Distributed Sequence Memory of Multidimensional Inputs in Recurrent Networks

Adam S. Charles   ADAMSC@PRINCETON.EDU
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08544, USA

Dong Yin   DONGYIN@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Christopher J. Rozell   CROZELL@GATECH.EDU
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA

Abstract

Recurrent neural networks (RNNs) have drawn interest from machine learning researchers because of their effectiveness at preserving past inputs for time-varying data processing tasks. To understand the success and limitations of RNNs, it is critical that we advance our analysis of their fundamental memory properties. We focus on echo state networks (ESNs), which are RNNs with simple memoryless nodes and random connectivity. In most existing analyses, the short-term memory (STM) capacity results conclude that the ESN network size must scale linearly with the input size for unstructured inputs. The main contribution of this paper is to provide general results characterizing the STM capacity for linear ESNs with multidimensional input streams when the inputs have common low-dimensional structure: sparsity in a basis or significant statistical dependence between inputs. In both cases, we show that the number of nodes in the network must scale linearly with the information rate and poly-logarithmically with the input dimension. The analysis relies on advanced applications of random matrix theory and results in explicit non-asymptotic bounds on the recovery error. Taken together, this analysis provides a significant step forward in our understanding of the STM properties in RNNs.

Keywords: short-term memory, recurrent neural networks, sparse signal recovery, low-rank recovery, restricted isometry property

1. Introduction

Recurrent neural networks (RNNs) have drawn interest from researchers because of their effectiveness at processing sequences of data (Jaeger, 2001; Lukoševičius, 2012; Hinaut et al., 2014). While deep networks have shown remarkable performance improvements at task such as image classification, RNNs have recently been successfully employed as layers in conventional deep neural networks to expand these tools into tasks with time-varying data (Sukhbaatar et al.; Gregor et al., 2015; Graves et al., 2013; Bashivan et al., 2016). This inclusion is becoming increasingly important as neural networks are being applied to a growing variety of inherently temporal high-dimensional data, such as video (Donahue et al., 2015), audio (Graves et al., 2013), EEG data (Bashivan et al., 2016), two-photon calcium imaging (Apthorpe et al., 2016). Despite the growing use of both deep and recurrent networks, theory characterizing the properties of such networks remain relatively unexplored. For deep neural networks, much of the computational power is often attributed to flexibility in learned representations (Mallat, 2016; Vardan et al., 2016; Patel et al., 2015). The power of RNNs, however, is tied to the ability of the recurrent network dynamics to act as a distributed memory substrate, preserving information about past inputs to leverage temporal dependencies for data processing tasks such as classification and prediction. To understand the success and limitations of RNNs, it is critical that we advance our analysis of the fundamental memory properties of these network structures.

There are many types of recurrent network structures that have been employed in machine learning applications, each with varying complexity in the network elements and the training procedures. In this paper we will focus on RNN structures known as echo state networks (ESNs). These networks have discrete time continuous-valued nodes $x[n] \in \mathbb{R}^M$ that evolve at time n in response to the inputs $s[n] \in \mathbb{R}^L$ according to the dynamics:

$$x[n+1] = f(Wx[n] + Zs[n] + \epsilon[n]), \tag{1}$$

where $W \in \mathbb{R}^{M \times M}$ is the connectivity matrix defining the recurrent dynamics, $Z \in \mathbb{R}^{M \times L}$ is the weight vector describing how the input drives the network, $f(\cdot) : \mathbb{R}^M \to \mathbb{R}^M$ is an element-wise nonlinearity evaluated at each node and $\epsilon[n] \in \mathbb{R}^M$ represents the error due to potential system imperfections (Jaeger, 2001; Wilson and Cowan, 1972; Amari, 1972; Sompolinsky et al., 1988; Maass et al., 2002). In an ESN, the connectivity matrix W is random and untrained, while the simple individual nodes have a single state variable with no memory. This is in contrast to approaches such as long short-term memory units (Sak et al., 2014; Lipton et al., 2016; Kalchbrenner et al., 2016) which have individual nodes with complex memory properties. As with many other recent papers, we will also focus on linear networks where $f(\cdot)$ is the identity function (Jaeger, 2001; Jaeger and Haas, 2004; White et al., 2004; Ganguli et al., 2008; Ganguli and Sompolinsky, 2010; Charles et al., 2014; Wallace et al., 2013).

The memory capacity of these networks has been studied in both the machine learning and computational neuroscience literature. In the approach of interest, the short-term memory (STM) of a network is characterized by quantifying the relationship between the transient network activity and the recent history of the exogenous input stream driving the network (Jaeger and Haas, 2004; Maass et al., 2002; Ganguli and Sompolinsky, 2010; Wallace et al., 2013; Verstraeten et al., 2007; White et al., 2004; Lukoševičius and Jaeger, 2009; Buonomano and Maass, 2009; Charles et al., 2014). Note that this is in contrast to alternative approaches that characterize long-term memory in RNNs through quantifying the number of distinct network attractors that can be used to stably remember

input patterns with the asymptotic network state. In the vast majority of the existing theoretical analysis of STM, the results conclude that networks with M nodes can only recover inputs of length $N \leq M$ (White et al., 2004; Wallace et al., 2013) when the inputs are unstructured.

However, in any machine-learning problem of interest, the input statistical structure is precisely what we intend to exploit to accomplish meaningful tasks. For one example, many signals are well-known to admit a sparse representation in a transform dictionary (Elad et al., 2010; Davies and Daudet, 2006). In fact, some classes of deep neural networks have been designed to induce sparsity at higher layers that may serve as inputs into the recurrent layers (LeCun et al., 2010; Kavukcuoglu et al., 2010). For another example, a collection of time-varying input streams (e.g., pixels or image features in a video stream) are often heavily correlated. In the specific case of single input streams ($L = 1$) with inputs that are K-sparse in a basis, recent work (Charles et al., 2014) has shown that the STM capacity can scale as favorably as $M = \mathcal{O}\left(K \log^{\gamma}(N)\right) \leq N$, where $\gamma \geq 1$ is a constant. In other words, the memory capacity can scale linearly with the information rate in the signal and only logarithmically with the signal dimension, resulting in the potential for recovery of inputs of length $N \gg M$. Unfortunately, existing analyses (Jaeger, 2001; White et al., 2004; Ganguli and Sompolinsky, 2010; Charles et al., 2014) are generally specific to the restricted case of single time-series inputs ($L = 1$) or unstructured inputs (Verstraeten et al., 2010).

Conventional wisdom is that structured inputs should lead to much higher STM capacity, though this has never been addressed with strong analysis in the general case of ESNs with multidimensional input streams. The main contribution of this paper is to provide general results characterizing the STM capacity for linear randomly connected ESNs with multidimensional input streams when the inputs are either sparse in a basis or have significant statistical dependence (with no sparsity assumption). In both cases, we show that the number of nodes in the network must scale linearly with the information rate and poly-logarithmically with the total input dimension. The analysis relies on advanced applications of random matrix theory, and results in non-asymptotic analysis of explicit bounds on the recovery error. Taken together, this analysis provides a significant step forward in our understanding of the STM properties in RNNs. While this paper is primarily focused on network structures in the context of RNNs in machine learning, these results also provide foundation for the theoretical understanding of recurrent network structures in biological neural networks, as well as the memory properties in other network structures with similar dynamics (e.g., opinion dynamics in social networks).

2. Background and Related Work

2.1 Short Term Memory in Recurrent Networks

Many approaches have been used to analyze the STM of randomly connected networks, including nonlinear networks (Sompolinsky et al., 1988; Massar and Massar, 2013; Faugeras et al., 2009; Rajan et al., 2010; Galtier and Wainrib, 2016; Wainrib, 2015) and linear networks (Jaeger, 2001; Jaeger and Haas, 2004; White et al., 2004; Ganguli et al., 2008; Ganguli and Sompolinsky, 2010; Charles et al., 2014; Wallace et al., 2013) with both discrete-time and continuous-time dynamics. These methods can be broadly be classified as either correlation-based methods (White et al., 2004; Ganguli et al., 2008) or uniqueness methods (Jaeger, 2001; Maass et al., 2002; Jaeger and Haas, 2004; Charles et al., 2014; Legenstein and Maass, 2007; Büsing et al., 2010). Correlation methods focus on quantifying the correlation between the network state and recent network inputs. In these studies, the STM is defined as the time of the oldest input where the correlation between the network

state and that input remains above a given threshold (White et al., 2004; Ganguli et al., 2008). These methods have mostly been applied to discrete-time systems, and have resulted in bounds on the STM that scale linearly with the number of nodes (i.e. $M > N$).

In contrast, uniqueness methods instead aim to show that different network states correspond to unique input sequences (i.e. the network dynamics are bijective).[1] For uniqueness methods, the STM is defined as the longest input length where this input-network state bijection still holds. These methods have been used under the term *separability property* for continuous-time liquid state machines (Maass et al., 2002; Vapnik and Chervonenkis, 1971; Legenstein and Maass, 2007; Wallace et al., 2013; Büsing et al., 2010) and under the term *echo-state property* for discrete-time ESNs (Jaeger, 2001; Yildiz et al., 2012; Buehner and Young, 2006; Manjunath and Jaeger, 2013). The echo-state property is the method most related to the approach we take here, and essentially derives the maximum length of the input signal such that the resulting network states remain unique. While this property guarantees a bijection between inputs and network states, it does not take into account input signal structure, does not capture the robustness of the mapping, and does not provide guarantees for stably recovering the input from the network activity.

2.2 Compressed Sensing

The compressed sensing literature and its recent extensions include many tools for studying the effects of random matrices applied to low-dimensional signals. Specifically, in the basic compressed sensing problem we desire to recover the signal $s \in \mathbb{R}^N$ from M measurements[2] generated from a random linear measurement operator,

$$x = \mathcal{A}(s) + \epsilon, \tag{2}$$

where $\epsilon \in \mathbb{R}^M$ represents the potential measurement errors. Typically, s is assumed to have low-dimensional structure and recovery is performed via a convex optimization program. The most common example is a sparsity model where s can be represented as

$$s = \Psi a,$$

where $\Psi \in \mathbb{R}^{N \times N}$ is a transform matrix and $a \in \mathbb{R}^N$ is the sparse coefficient representation of s with $K \ll N$ of its entries non-zero. Under this sparsity assumption, the coefficient representation is recoverable if the linear operator $\mathcal{A}$ satisfies the restricted isometry property (RIP) that guarantees uniqueness of the compressed measurements. Specifically, we say that $\mathcal{A}$ satisfies the RIP($2K$,δ) if for every $2K$-sparse signal s, the following condition is satisfied:

$$C(1 - \delta) \leq ||\mathcal{A}s||_2^2 / ||s||_2^2 \leq C(1 + \delta),$$

where $0 < \delta < 1$ and $C > 0$ is a positive constant. When $\mathcal{A}$ satisfies the RIP($2K$,δ) the sparse coefficients a can be recovered by solving an ℓ_1-norm based optimization function

$$a = \arg\min_a ||a||_1 \quad \text{such that} \quad ||x - \mathcal{A}(\Psi a)||_2 \leq ||\epsilon||_2, \tag{3}$$

---

1. We note that uniqueness-based methods imply recovery-based methods, modulo a recovery algorithm, as often recovery guarantees are based on some semblance of a bijection.

2. In the case of RNNs, the network node values act as the measurements of our system, prompting the use of M as the number of measurements in this section

up to a reconstruction error given by

$$||\boldsymbol{s} - \widehat{\boldsymbol{s}}||_2 \le \alpha \, ||\boldsymbol{\epsilon}||_2 + \beta \frac{\left|\left|\boldsymbol{\Psi}^T \left(\boldsymbol{s} - \boldsymbol{s}_K\right)\right|\right|_1}{\sqrt{K}}, \tag{4}$$

where α and β are constants (Candès et al., 2006). The first term of this recovery error bound depends on the norm of the measurement error $\boldsymbol{\epsilon}$, while the second term depends on the ℓ_1 difference between the true signal and the best K-sparse approximation of the true vector ($\boldsymbol{s}_K$). This term essentially measures how closely the signal matches the sparsity model. The ℓ_1 optimization program in (3) required for recovery can be solved by many efficient algorithms, including neurally plausible architectures (Rozell et al., 2010; Balavoine et al., 2012; Shapero et al., 2014; Charles et al., 2012).

When the data of interest is a matrix $\boldsymbol{S} \in \mathbb{R}^{L \times N}$, other low-dimensional models have also been explored. For example, as an alternative to a sparsity assumption, the successful low-rank model assumes that there are correlations between rows and columns such that $\boldsymbol{S}$ has rank $R < \min\{L, N\}$. We can then write the decomposition of the matrix as

$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{V}^*,$$

where $\boldsymbol{Q} \in \mathbb{R}^{L \times R}$ and $\boldsymbol{V}^* \in \mathbb{R}^{R \times N}$. There is a rich and growing literature dedicated to establishing guarantees for recovering low-rank matrices from incomplete measurements. Due to the difficulty of establishing a general matrix-RIP property for observations of a matrix (Recht et al., 2010), the guarantees in this literature more commonly use the optimality conditions for specific optimization procedures to show that the resulting solution has bounded error with high probability. The most common optimization program used for low-rank matrix recovery is the nuclear norm minimization,

$$\boldsymbol{S} = \arg \min_{\boldsymbol{S}} ||\boldsymbol{S}||_* \quad \text{such that} \quad ||\boldsymbol{x} - \mathcal{A}\left(\boldsymbol{S}\right)||_2 \le ||\boldsymbol{\epsilon}||_2, \tag{5}$$

where the nuclear norm $||\boldsymbol{S}||_*$ is defined as the sum of the singular values of $\boldsymbol{S}$ (Candès and Tao, 2010; Candès and Plan, 2010; Recht et al., 2010; Chen and Suter, 2004; Fazel, 2002; Singer and Cucuringu, 2010; Toh and Yun, 2010; Liu and Vandenberghe, 2009; Jaggi et al., 2010). This optimization procedure is similar to the ℓ_1-regularized optimization of Equation (3), however the nuclear-norm induces sparsity in the singular values rather than the matrix entries directly.

The solution to Equation (5) can be shown to satisfy performance guarantees via the dual-certificate approach (Candès and Plan, 2010; Ahmed and Romberg, 2015). This technique is a proof by construction and shows that a dual certificate (i.e., a vector whose projections into and out of the space spanned by the singular vectors of $\boldsymbol{S}$ are bounded) exists. Showing that such a certificate exists demonstrates that Equation (5) converges to a valid solution and is key to deriving accuracy bounds (Candès and Plan, 2010; Ahmed and Romberg, 2015). Specifically, if the dual certificate exists, then the solution to Equation (5) satisfies the recovery bound

$$\left|\left|\widehat{\boldsymbol{S}} - \boldsymbol{S}\right|\right|_F \le \left(4\sqrt{\min(N,L)\frac{2NL+M}{M}} + 2\right)\epsilon, \tag{6}$$

where the Forbenius norm $||\cdot||_F^2$ is defined as the sum of the squares of all the matrix entries. This bound demonstrates that perfect recovery is achievable in the case where there is no error ($\epsilon = 0$). We note that alternate optimization programs with similar guarantees have been proposed in the literature for inferring low-rank matrices (i.e. Ahmed and Romberg, 2015), but we will focus on nuclear norm optimization approaches due to the extensive literature on nuclear-norm solvers and the connections to sparse vector inference.

2.3 STM Capacity via the RIP

The ideas and tools from the compressed sensing literature have recently been used to show that *a-priori* knowledge of the input sparsity can lead to improvements recovery-based STM capacity results for ESNs. For a single input stream under a sparsity assumption, Ganguli and Sompolinsky (2010) analyzed an annealed version of the network dynamics to show that the network memory capacity can be larger than the network size. Building on this observation, Charles et al. (2014) provided an analysis of the exact network dynamics in an ESN (for the single input case of $L = 1$), yielding precise bounds on a network's STM capacity captured in the following theorem:

Theorem 1 (Theorem 4.1.1, Charles et al., 2014) *Suppose $N \geq M$, $N \geq K$, $N \geq O(1)$,[3] and $L = 1$. Let U be any unitary matrix of eigenvectors (containing complex conjugate pairs) of the connectivity matrix W and for M an even integer, denote the eigenvalues of W by $\{e^{jw_m}\}_{m=1}^{M}$. Let the first $M/2$ eigenvalues ($\{e^{jw_m}\}_{m=1}^{M/2}$) be chosen uniformly at random on the complex unit circle (i.e., $\{w_m\}_{m=1}^{M/2}$ is uniformly distributed over $[0, 2\pi)$) and the other $M/2$ eigenvalues as the complex conjugates of these values. Furthermore, let the entries of the input weights z be i.i.d. zero-mean Gaussian random variables with variance $\frac{1}{M}$. Given RIP conditioning δ and failure probability $N^{-\log^4 N} \leq \eta \leq \frac{1}{e}$, if*

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

then for a universal constant C, with probability $1 - \eta$ the mapping of length-N input sequences into M network state variables satisfies the RIP($2K$, δ).

This theorem proves a rigorous and non-asymptotic bound on the length of the input that can be robustly extracted from the network nodes. By showing the RIP property on the network dynamics, the recovery bound given in Equation (4) establishes the recovery performance for any N-length, K-sparse signal from the resulting network state at time N. In short, the number of required nodes scales linearly with the information rate of the signal (i.e., the sparsity level) and poly-logarithmically with the length of the input. The coherence factor $\mu^2(\Psi)$, defined as

$$\mu (\Psi) = \max_{n=1,...,N} \sup_{t \in [0,2\pi]} \left| \sum_{m=0}^{N-1} \Psi_{m,n} e^{-jtm} \right|,$$

expresses the *types* of sparsity that are efficiently stored in the network. Essentially this coherence factor is large (on the order of $\sqrt{N}$) for inputs sparse in the Fourier basis, and is very low (essentially a small constant) for inputs that are sparse in bases different from the Fourier bases (e.g. wavelet transforms). For the extreme case of Fourier-sparse inputs, the number of nodes must again exceed the number of inputs. When this coherence is low and $K \ll M$, this bound is a clear improvement over existing results as it allows for $N > M$. However, this result is restricted to single input streams with one type of low-dimensional structure. The current paper addresses the much more general problem of multidimensional inputs and other types of low-dimensional structure.

---

3. We use $O(1)$ notation to indicate that a variable is a finite constant.
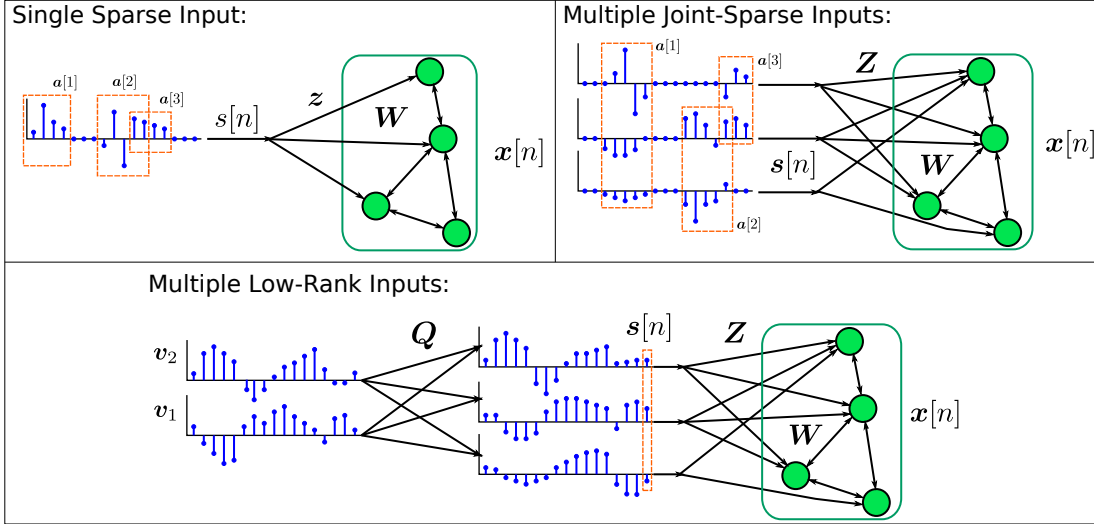
Figure 1: Echo-state networks can efficiently store inputs with a variety of low-dimensional structures. Top-Left: With a single input stream, the coefficients $\boldsymbol{a}$ represent chunks of activity in the input stream $s[n]$ (shown in red boxes). The raw input stream values then drive the network via Equation (1), resulting in a transient network state $\boldsymbol{x}[N]$ that encodes the input stream. Top-Right: In the case of multiple inputs that are jointly sparse, each coefficient can now represent a chunk of activity both across time and across input streams (as depicted by the red boxes). Bottom: When the multiple input streams are instead low-rank, each input stream is instead described by a linear combination of prototypical vectors $\boldsymbol{v}_k$. The matrix $\boldsymbol{Q}$ represents how the prototypical vectors are combined in order to obtain the input sequences fed into the network.

3. STM for Multi-Input Networks

In this work we will use the tools of random matrix theory to establish STM capacity results for recurrent networks under the general conditions of multiple simultaneous input streams and a variety of low-dimensional models. The temporal evolution of the linear network with multiple inputs is similar to the previous ESN definition, with the main difference being that the input at each time-step $\boldsymbol{s}[n] \in \mathbb{R}^L$ is a length L vector that drives the network through a feed-forward matrix $\boldsymbol{Z} \in \mathbb{R}^{M \times L}$ rather than a feed-forward vector,

$$
\begin{aligned}
\boldsymbol{x}[n] &= \boldsymbol{W}\boldsymbol{x}[n-1] + \sum_{l=1}^{L} \boldsymbol{z}_l s_l[n] + \widetilde{\boldsymbol{\epsilon}}[n] \\
&= \boldsymbol{W}\boldsymbol{x}[n-1] + \boldsymbol{Z}\boldsymbol{s}[n] + \widetilde{\boldsymbol{\epsilon}}[n].
\end{aligned}
\tag{7}
$$

We denote the columns of $\boldsymbol{Z}$ as $\boldsymbol{z}_l$ to separately notate the vectors mapping each input stream.

We can write the current network state as a linear function of the inputs by iterating Equation (7),

$$
\boldsymbol{x}[N] = \sum_{k=1}^{N} \boldsymbol{W}^{N-k} \boldsymbol{Z}\boldsymbol{s}[k] + \boldsymbol{\epsilon},
$$

where the error term $\epsilon = \sum_{k=1}^{N} \boldsymbol{W}^{N-k}\widetilde{\epsilon}[k]$ is the accumulated error, and then rewriting sum as a matrix-vector multiply,

$$\boldsymbol{x}[N] = \left[\boldsymbol{Z}, \boldsymbol{W}\boldsymbol{Z}, \cdots, \boldsymbol{W}^{N-1}\boldsymbol{Z}\right]\left[\boldsymbol{s}^T[N], \boldsymbol{s}^T[N-1], \cdots, \boldsymbol{s}^T[1]\right]^T + \epsilon.$$

Depending on the signal statistics in question, we will find it convenient in some cases to express the network dynamics in terms of a linear operator applied to an input matrix, i.e.

$$\boldsymbol{x}[N] = \mathcal{A}\left(\boldsymbol{S}\right) + \epsilon,$$

where $\boldsymbol{S} = \left[\boldsymbol{s}^T[N], \boldsymbol{s}^T[N-1], \cdots, \boldsymbol{s}^T[1]\right]^T$. In other cases, we find it more convenient to reorganize the columns into an effective measurement matrix applied to a vector of inputs. By defining the eigen-decomposition of $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$, we can re-write the dynamics process as

$$\boldsymbol{x}[N] = \boldsymbol{U}\left[\boldsymbol{D}^0\boldsymbol{U}^{-1}\boldsymbol{Z}, \boldsymbol{D}\boldsymbol{U}^{-1}\boldsymbol{Z}, \cdots, \boldsymbol{D}^{N-1}\boldsymbol{U}^{-1}\boldsymbol{Z}\right]\left[\boldsymbol{s}^T[N], \boldsymbol{s}^T[N-1], \cdots, \boldsymbol{s}^T[1]\right]^T + \epsilon.$$

To simplify this expression, we can reorganize the columns of the linear operator (and the rows of the vector of inputs) such that all the inputs corresponding to the lth input vector $\boldsymbol{z}_l$ create a single block. The kth row of the lth block of out matrix is now represented by $\boldsymbol{D}^{k-1}\boldsymbol{U}^{-1}\boldsymbol{z}_l$, which can be written as $\widetilde{\boldsymbol{Z}}_l \boldsymbol{d}_{k-1}$, where $\widetilde{\boldsymbol{Z}}_l = \boldsymbol{U}^{-1}\boldsymbol{z}_l$ and $\boldsymbol{d}_{k-1}$ is the vector of the diagonal elements of $\boldsymbol{D}$ raised to the $(k-1)$ power. We can more concisely by defining the matrix $\boldsymbol{F}$ consisting of the eigenvalues of $\boldsymbol{W}$ raised to different powers (i.e. $\boldsymbol{F}_{i,j} = d_i^{j-1}$), resulting in the expression

$$\boldsymbol{x}[N] = \boldsymbol{U}\left[\widetilde{\boldsymbol{Z}}_1\boldsymbol{F}, \widetilde{\boldsymbol{Z}}_2\boldsymbol{F}, \cdots, \widetilde{\boldsymbol{Z}}_L\boldsymbol{F}\right]\left[\boldsymbol{s}_1^T, \boldsymbol{s}_2^T, \cdots, \boldsymbol{s}_L^T\right]^T + \epsilon = \boldsymbol{A}\widetilde{\boldsymbol{s}} + \epsilon. \tag{8}$$

Since the eigenvalues of $\boldsymbol{W}$ here are restricted to reside on the unit circle, we note that $\boldsymbol{F}$ is a Vandermonde matrix whose rows are Fourier basis vectors. From Equation (8) we see that the current state is simply the sum of L compressed input streams, where the compression for each block essentially performs the same compression as for a single stream, but modulated by the different feed-forward vectors $\boldsymbol{z}_l$.

3.1 Sparse Multiple Inputs

To begin, we consider the direct extension of previous results based on sparsity models to the multi-input setting. In this setting we consider the model where the composite of all input signals is sparse in a basis $\boldsymbol{\Psi} \in \mathbb{R}^{NL \times NL}$ so that $\widetilde{\boldsymbol{s}} = \boldsymbol{\Psi}\widetilde{\boldsymbol{a}}$. This means that each signal stream can be written as $\boldsymbol{s}_l = \sum_{k=1}^{L} \boldsymbol{\Psi}^{l,k}\boldsymbol{a}_k$ where $\boldsymbol{\Psi}^{l,k}$ is the $\{l,k\}^{th}$ $N \times N$ block of $\widetilde{\boldsymbol{\Psi}}$. This signal model captures dependencies between input streams because a given coefficient can influence multiple channels. While in many application the basis $\boldsymbol{\Psi}$ is pre-specified (i.e. wavelet decomposition in image processing; Christopoulos et al., 2000), these bases can also be learned from exemplar data via dictionary learning algorithms (Olshausen and Field, 1996; Aharon et al.). This sparsity model can be a useful model for signals of interest, such as video signals, where similar sparse decompositions have been used for action recognition (Guha and Ward, 2012) and video categorization (Chiang et al., 2013). With this model, we will use a generalized notion of the coherence parameter used in (Charles et al., 2014):

$$\mu_S\left(\boldsymbol{\Psi}\right) = \max_{l,k=1,\ldots,L} \max_{n=1,\ldots,N} \sup_{t \in [0,2\pi]} \frac{\left|\sum_{m=0}^{N-1} \boldsymbol{\Psi}_{m,n}^{l,k} e^{-jtm}\right|}{\|\boldsymbol{\Psi}_m^{l,k}\|_2}. \tag{9}$$

In this case, each $N \times N$ block must be different from the Fourier basis to achieve high STM capacity. This restriction is reasonable, since if a single sub-block of $\mathbf{\Psi}$ was coherent with the Fourier basis, then at least one input stream could be sparse in a Fourier-like basis and hence would be unrecoverable. Using this network and signal model, we obtain the following theorem on the stability of the network representation:

Theorem 2 *Suppose $NL \geq M$, $N \geq K$ and $N \geq O(1)$. Let $\mathbf{U}$ be any unitary matrix of eigenvectors (containing complex conjugate pairs) and the entries of $\mathbf{Z}$ be i.i.d. zero-mean Gaussian random variables with variance $\frac{1}{M}$. For M an even integer, denote the eigenvalues of $\mathbf{W}$ by $\{e^{jw_m}\}_{m=1}^{M}$. Let the first $M/2$ eigenvalues ($\{e^{jw_m}\}_{m=1}^{M/2}$) be chosen uniformly at random on the complex unit circle (i.e., we chose $\{w_m\}_{m=1}^{M/2}$ uniformly at random from $[0, 2\pi)$) and the other $M/2$ eigenvalues as the complex conjugates of these values. For a given RIP conditioning δ, failure probability $N^{-\log^4 N} \leq \eta \leq \frac{1}{e}$, and coherence $\mu_S(\mathbf{\Psi})$ as defined as in Equation (9), if*

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

then $\mathbf{A}$ satisfies RIP-$(2K, \delta)$ with probability exceeding $1 - \eta$ for a universal constant C.

The proof of Theorem 2 is provided in Appendix A.1. Note that when $L = 1$, Theorem 2 reduces to Theorem 1. In this result we see that that the number of nodes relies only linearly on the underlying dimensionality (K) and poly-logarithmically on the total size of the input (NL). This means that under favorable coherence and sparsity conditions on the input, the network can again have STM capacities that are higher than the number of nodes in the network. Specifically, showing that $\mathbf{A}$ satisfies the RIP property, Theorem 2 ensures that standard recovery guarantees from the sparse inference literature hold. In particular, any K-sparse input is recoverable from the network state at time N up to the error bound of Equation (4) by solving the ℓ_1-regularized least-squares optimization of Equation (3).

3.2 Low Rank Multiple Inputs

Next we consider the case of a very different type of low-dimensional structure where the input signals are correlated but not necessarily sparse. Specifically, in this setting we assume that the inputs arise from a process where R prototypical signals combine linearly to form the various input streams. Such a signal structure could arise, for instance, due to correlations between input streams at spatially neighboring locations. A number of interesting applications display such correlations, including important measurement modalities in neuroscience (e.g. two-photon calcium imaging; Denk et al., 1990; Maruyama et al., 2014 and neural electrophysiological recordings; Berényi et al., 2014; Ahmed and Romberg, 2015), and remote sensing applications (e.g. hyperspectral imagery; Zhang et al., 2014; Veganzones et al., 2016). The applicability of RNNs and machine learning methods to data well described by this low-rank model is also increasingly relevant as there is increasing interest in applying neural network techniques to such data, either for detection (Apthorpe et al., 2016), classification (Chen et al., 2016, 2014), or as samplers via variational auto-encoders (Gao et al., 2016). In this case, we can write out the input matrix in the reduced form $\mathbf{S} = \mathbf{Q} \mathbf{V}^*$, where $\mathbf{V}^* \in \mathbb{R}^{R \times N}$ is the matrix whose rows may represent environmental causes generating the data and $\mathbf{Q} \in \mathbb{R}^{L \times R}$ represents the mixing matrix that defines the input stream. We

will assume both $L \geq R$ and $N \geq R$, meaning that $\boldsymbol{S}$ is low-rank. With this model we use a definition of coherence given by:

$$\mu_L^2 = R^{-1} \sup_{\omega \in [0,2\pi]} ||\boldsymbol{V}^* \boldsymbol{f}_\omega||_2^2. \tag{10}$$

where $\boldsymbol{f}_\omega = [1, e^{-j\omega}, \cdots e^{-j(N-1)\omega}]^T$ is the Fourier vector with frequency ω. This coherence parameter mirrors the coherence used for the sparse-input case. As μ_S measured the similarity between the measurement vectors and the sparsity basis $\boldsymbol{\Psi}$, μ_L measures the similarity between the measurements and the left singular vectors of the measured matrix. The intuition here is that measurements that align with the left singular vectors are unlikely to measure significant information about the $\boldsymbol{S}$.

To analyze the STM of the network dynamics with respect to low-rank signal statistics, we leverage the dual certificate approach (Candès and Plan, 2010, 2011a; Ahmed and Romberg, 2015) to derive the following theorem,

Theorem 3 *Suppose $NL \geq M$, $N \geq R$, $N \geq O(1)$ and $L \geq O(1)$. Let $\boldsymbol{z}$ be i.i.d. zero-mean Gaussian random variables with variance $\frac{1}{M}$. For M an even integer, denote the eigenvalues of $\boldsymbol{W}$ by $\{e^{jw_m}\}_{m=1}^M$. Let the first $M/2$ eigenvalues ($\{e^{jw_m}\}_{m=1}^{M/2}$) be chosen uniformly at random on the complex unit circle (i.e., we chose $\{w_m\}_{m=1}^{M/2}$ uniformly at random from $[0, 2\pi)$) and the other $M/2$ eigenvalues as the complex conjugates of these values. For a given coherence μ_L as defined as in Equation (10), if*

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

then, with probability at least $1 - O((LN)^{1-\beta}$, the minimization in Equation (5) recovers the rank-R input matrix $\boldsymbol{S}$ up to the error bound in Equation (6).

The proof of Theorem 3 is in Appendix A.2 and follows a golfing scheme to find an inexact dual certificate. In fact, we note that since our architecture is extremely similar mathematically to the architecture in (Ahmed and Romberg, 2015), our proof is also very similar. The main difference is that due to the unbounded nature of our distributions (i.e. the feed-forward vectors $\boldsymbol{Z}$ are Gaussian random variables) and the fact that our Fourier vectors are on the unit circle (rather than gridded), we can consider our proof as a generalization of the proof in (Ahmed and Romberg, 2015).

Theorem 3 is qualitatively similar to Theorem 2 in the way the STM capacity scales. In this case, the bound still scales linearly with the information rate as captured by the number of elements in the left and right matrices that compose $\boldsymbol{S}$: $RN + RL$. Interestingly, due to the left singular vectors interacting with the measurement operator first, the coherence term only affects the portion of the bound related to the number of elements in $\boldsymbol{Q}$. Additionally, as before, the number of total inputs LN only impacts the bound poly-logarithmically.

4. Simulation

To empirically verify that these theoretical STM scaling laws are representative of the empirical behavior, we generated a number of random networks and evaluated the recovery of (sparse or low-rank) input sequences in the presence of noise. For each simulation we generate a $M \times M$ random

(a) RMSE: No basis   (b) RMSE: Haar wavelet basis   (c) RMSE: DCT basis
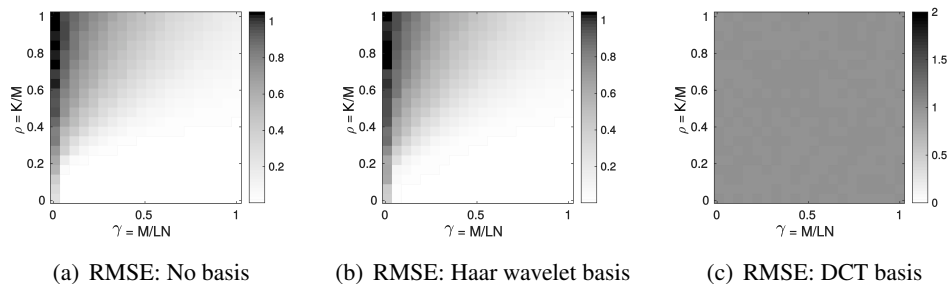
Figure 2: ESNs can have high STM capacity for multidimensional sparse inputs. Relative mean-squared error (rMSE) of the recovery for canonical sparse inputs (a) and Haar wavelet-sparse inputs (b) is very low for a range of sparsity and network sizes satisfying $M < LN$. The rMSE for DCT-sparse inputs (c), as predicted by our theoretical results, remains high (approximately 100% error).

orthogonal connectivity matrix $\boldsymbol{W}$[4] and a $M \times L$ random Gaussian feed-forward matrix $\boldsymbol{Z}$. In both cases we fixed the number of inputs to $L = 40$ and the number of time-steps to $N = 100$ while varying the network size M and underlying dimensionality of the input (i.e., the sparsity level or the input matrix rank). For the sparse input simulations, inputs were chosen with a uniformly random support pattern with random Gaussian values on the support. For low-rank simulations, the right singular vectors were chosen to be Gaussian random vectors, and the left singular values were chosen at random from a number of different basis sets.

In Figure 2 we show the relative mean-squared error of the input recovery as a function of the sparsity-to-network size ratio $\rho = K/M$ and the network size-to-input ratio $\gamma = M/NL$. Each pixel value represents the average recovery relative mean-squared error (rMSE), as calculated by

$$\text{RMSE} = \frac{\|\widehat{\boldsymbol{s}} - \boldsymbol{s}\|_2^2}{\|\boldsymbol{s}\|_2^2},$$

over 20 randomly generated trials with a noise level of $\|\boldsymbol{\epsilon}\|_2 \approx 0.01$. We show results for recovery of three different types of sparse signals: signals sparse in the canonical basis, signals sparse in a Haar wavelet basis, and signals sparse in a discrete cosine transform (DCT) basis. As our theory predicts, for canonical- and Haar wavelet-sparse signals the network has very favorable STM capacity results. The fact that the capacity achieves $M < NL$ is demonstrated by the area left of the $M = NL$ point ($\gamma = 1$) where the signal is recovered with high accuracy. Likewise, for the DCT-sparse signals we find that the inputs are never recovered well for any $M < NL$. This behavior is also predicted by our theory because of the unfavorable coherence properties of the DCT basis.

For the low-rank trials we see that recovery of low-rank inputs for a range of $M < LN$ is possible as predicted by the theoretical results. As with the sparse input case we consider three types of low-rank inputs. Instead of changing the sparsity basis, however, we change the right singular vectors $\boldsymbol{V}$ of the low-rank input matrix $\boldsymbol{S}$. We explore the cases where the elements of $\boldsymbol{V}$

---

4. Orthogonal connectivity matrices were obtained by running an orthogonalization procedure on a random Gaussian matrix.

(a) RMSE: No basis   (b) RMSE: Haar wavelet basis   (c) RMSE: DCT basis
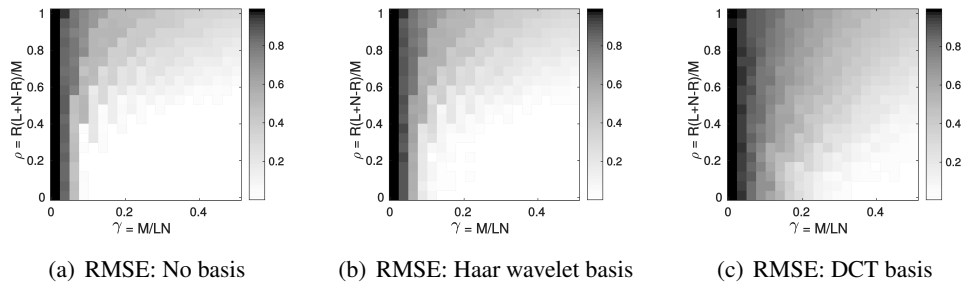
Figure 3: ESNs can have high STM capacity for multidimensional inputs with low-rank structure. The rMSE of the input recovery for inputs with canonical right singular vectors (a) and Haar wavelets for right singular vectors (b) is very low for a range of input rank and network sizes satisfying $M < LN$. The rMSE for inputs with DCT right singular vectors (c) remains high, a behavior predicted by our theoretical results.

are chosen from the canonical basis, the haar basis and the DCT basis. These results are shown in Figure 3 with plots similar to those in Figure 2, but with only showing the range $\gamma < 0.5$ to reduce computational time. As our theory predicts, the recovery of inputs with canonical- and Haar right singular vectors is more accurate for a larger range of ρ, γ pairs than the inputs with DCT right singular vectors.

5. Conclusions

Determining the fundamental limits of memory in recurrent networks is critical to understanding their behavior in machine learning tasks. In this work we show that randomly connected echo-state networks can exploit the low-dimensional structure in multidimensional input streams to achieve very high short-term memory capacity. Specifically, we show non-asymptotic bounds on recovery error for input sequences that have underlying low-dimensional statistics described by either joint sparsity or low-rank properties (with no sparsity). For multiple sparse inputs, we find that the network size must be linearly proportional to the input sparsity and only logarithmically dependent on the total input size (a combination of the input length and number of inputs). For inputs with low-rank structure, we find a similar dependency where the network size depends linearly on the underlying dimension of the inputs (the product of the input rank with the input length and the number of inputs) and logarithmically on the total input size. Both results continue to demonstrate that ESNs can have STM capacities much larger than the network size.

These results are a significant (conceptual and technical) generalization over previous work that provided theoretical guarantees in the case of a single sparse input (Charles et al., 2014). While the linear ESN structure is a simplified model, rigorous analysis of these networks has remained elusive due to the recurrence itself. These results isolate the properties of the transient dynamics due to the recurrent connections, and may provide one foundation for which to explore the analysis of other complex network attributes such as nonlinearities and spiking properties. We also note that knowledge of how well neural networks compresses structured signals could indicate methods to pick the size of recurrent network layers. Specifically, if a task is thought to require a certain time-

frame of an input signal, the overall sparsity (or rank) of the signal in that time-frame can be used in conjunction with the length of that time-frame to give a lower bound for the required number of nodes in the recurrent layer of the network.

While the current paper is restricted to orthogonal connectivity matrices, previous work (Charles et al., 2014) has shown that a number of network structures can satisfy these criteria, including some types of small-world network topologies. Additionally, we explored here low-rank and sparse inputs separately. The methods we have used to prove Theorem 3, however, have also been used to analyze recovery signals with other related structures (e.g. matrices that can be decomposed into the sum of a sparse and low-rank matrix) from compressive measurements (Candès and Plan, 2011b). Our bounds presented here therefore could open up avenues for similar analysis of other low-dimensional signal classes.

While the context of this paper is focused on the role of recurrent networks as a tool in machine learning tasks, these results may also lead to a better understanding of the STM properties in other networked systems (e.g., social networks, biological networks, distributed computing, etc.). With respect to the literature relating recurrent ESN and liquid-state machines to working memory in biological systems, the notion of sparsity has much of the same flavor as the concept of *chunking* (Gobet et al., 2001). Chunking is the concept of humans learning to remember items in highly correlated groups rather than remembering items individually as a way of artificially increasing their working memory. Similarly, the use of sparsity bases allow RNNs to 'chunk' items according to the basis elements. Thus, each basis counts only as one item (the true underlying sparsity) and the network needs only store these elements rather than storing every input separately.

Acknowledgments

A. Appendix

A.1 RIP for Multiple Gaussian Feed-forward Vectors

In this appendix we prove Theorem 2, showing that the matrix representing the network evolution with L inputs and *i.i.d.* Gaussian feed-forward vectors satisfies the RIP. Recall that we have $\boldsymbol{x}[N] = \boldsymbol{A}\widetilde{\boldsymbol{s}}$, where $\boldsymbol{A} \in \mathbb{R}^{M \times NL}$ is derived in Section 3 and that $\widetilde{\boldsymbol{s}}$ (the vectorization of $\boldsymbol{S}$) is sparse with respect to the basis Ψ, meaning that there is a K-sparse signal $\boldsymbol{a}$ such that $\widetilde{\boldsymbol{s}} = \Psi\boldsymbol{a}$. Similar to (Charles et al., 2014), this proof is based on showing conditions on M such that $\boldsymbol{A}$ satisfies the RIP with respect to $\boldsymbol{\Psi}$, i.e.

$$(1 - \delta) \left\|\boldsymbol{a}\right\|_2^2 \leq \left\|\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{a}\right\|_2^2 \leq (1 + \delta) \left\|\boldsymbol{a}\right\|_2^2,$$

holds with high probability for all K-sparse $\boldsymbol{a}$. This is equivalent to bounding the following probability of the event

$$\left\|(\boldsymbol{A}\boldsymbol{\Psi})^H \boldsymbol{A}\boldsymbol{\Psi} - I\right\|_K \leq \delta. \tag{11}$$

where the norm $||\boldsymbol{A}||_K$ is defined as

$$||\boldsymbol{A}||_K := \sup_{\boldsymbol{y} \ is \ K-sparse} \frac{\boldsymbol{y}^H \boldsymbol{A} \boldsymbol{y}}{||\boldsymbol{y}||_2^2}.$$

First we bound the expectation of $\left|\left|(\boldsymbol{A}\boldsymbol{\Psi})^H \boldsymbol{A}\boldsymbol{\Psi} - I\right|\right|_K$, and use the result to bound the tail probability of the event (11).

Proof

A.1.1 EXPECTATION

First we let

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \widetilde{\boldsymbol{Z}}_1 \boldsymbol{F} & \widetilde{\boldsymbol{Z}}_2 \boldsymbol{F} & \cdots & \widetilde{\boldsymbol{Z}}_L \boldsymbol{F} \end{bmatrix} \boldsymbol{\Psi},$$

Since $(\boldsymbol{A}\boldsymbol{\Psi})^H \boldsymbol{A}\boldsymbol{\Psi} = \widehat{\boldsymbol{A}}^H \widehat{\boldsymbol{A}}$, then for any $\boldsymbol{a} \in \mathbb{R}^{NL}$, $||\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{a}||_2 = \left|\left|\widehat{\boldsymbol{A}}\boldsymbol{a}\right|\right|_2$. Therefore we only need to prove that

$$\left|\left|\widehat{\boldsymbol{A}}^H \widehat{\boldsymbol{A}} - \boldsymbol{I}\right|\right|_K \le \delta,$$

holds with high probability when M is large enough. Let V_i^{H} denote the ith row of $\widehat{\boldsymbol{A}}$, i.e.,

$$V_i^{\mathrm{H}} = \sum_{l=1}^{L} \tilde{\boldsymbol{z}}_{i,l} \boldsymbol{F}_i^H \boldsymbol{\Psi}_i.$$

Let $\boldsymbol{B}_1 = \sum_{i=1}^{M/2} V_i V_i^H - \boldsymbol{I}/2$ and $\boldsymbol{B}_2 = \sum_{i=M/2+1}^{M} V_i V_i^H - \boldsymbol{I}/2$; it is easy to check that $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are complex conjugates, and that $\widehat{\boldsymbol{A}}^H \widehat{\boldsymbol{A}} - \boldsymbol{I} = \boldsymbol{B}_1 + \boldsymbol{B}_2$. Therefore $\left|\left|\widehat{\boldsymbol{A}}^H \widehat{\boldsymbol{A}} - \boldsymbol{I}\right|\right|_K \le 2 \left|\left|\boldsymbol{B}_1\right|\right|_K$, and we only need to show that $||\boldsymbol{B}_1||_K \le \delta$ with high probability when M is large enough. We can easily see that when $i \ne j$, $1 \le i, j \le M/2$, V_i^H and V_j^H are independent.

First we show that $\mathbb{E}\left[||\boldsymbol{B}_1||_K\right]$ is small with high probability when M is large enough. We then show that $||\boldsymbol{B}_1||_K$ is concentrated around its mean with high probability when M is large enough. By Lemma 6.7 in Rauhut (2010), for a Rademacher sequence ϵ_i, $i = 1, ..., M/2$, we have

$$\mathbb{E}\left[||\boldsymbol{B}_1||_K\right] = \mathbb{E}\left[\left|\left|\sum_{i=1}^{M/2}(V_i V_i^H - \frac{1}{M}\boldsymbol{I})\right|\right|_K\right] \le 2\mathbb{E}\left[\left|\left|\sum_{i=1}^{M/2} \epsilon_i V_i V_i^H\right|\right|_K\right].$$

We can now apply Lemma 8.2 from Rauhut (2010), giving us

$$
\begin{aligned}
\mathbb{E}\left[||\boldsymbol{B}_1||_K\right] &\le 2\mathbb{E}\left[\left|\left|\sum_{i=1}^{M/2} \epsilon_i V_i V_i^H\right|\right|_K\right] \\
&\le 2\mathbb{E}\left[\mathbb{E}\left[C_0 V_{\max}\sqrt{K}\log(100K)\sqrt{\log(4NL)\ln(5M)}\right.\right. \\
&\qquad\qquad \left.\left. \cdot \sqrt{\left|\left|\sum_{i=1}^{M/2} V_i V_i^H\right|\right|_K}\ |V_i, i = 1, ..., M/2\right]\right] \\
&\le \sqrt{C_1 K \log^4(NL)}\sqrt{\mathbb{E}[V_{\max}^2]\mathbb{E}\left[||\boldsymbol{B}_1||_K + \frac{1}{2}\right]}, \qquad (12)
\end{aligned}
$$

14

where the last inequality results due to $25K \leq NL$, $5M \leq 4NL$, the Cauthy-Schwarz inequality and the triangle inequality, and $V_{\max} = \max_{1 \leq l \leq M/2} ||V_l||_\infty$. Note that the pth element of V_l^H can be written

$$V_l^H(p) = \sum_{i=1}^{L} \tilde{z}_{l,i} F_l^H \Psi_i(p).$$

and since we know that

$$\sum_{i=1}^{L} |F_l^H \Psi_i(p)|^2 \leq \sum_{i=1}^{L} ||\Psi_i(p)||_2^2 \mu^2(\Psi) = \mu^2(\Psi),$$

and

$$V_{\max} = \max_{\substack{1 \leq l \leq M/2 \\ 1 \leq p \leq NL}} |V_l(p)|$$

we can now use Corollary 5. Setting $Q = MNL/2$, $\mu_0 = \mu(\Psi)$ in Corollary 5 yields

$$\mathcal{P}\left[V_{\max}^2 > \frac{\mu^2(\Psi)}{M} \log \frac{MNL}{2\eta}\right] \leq \eta, \tag{13}$$

and

$$\mathbb{E}\left[V_{\max}^2\right] \leq \frac{\mu^2(\Psi)}{M}(\log \frac{MNL}{2} + 1) \leq C_2 \frac{\mu^2(\Psi)}{M} \log(NL). \tag{14}$$

Returning to the inequality in Equation (12), considering (14), we have

$$\mathbb{E}\left[||B_1||_K\right] \leq a\sqrt{\mathbb{E}\left[||B_1||_K\right] + 1},$$

where $a = \sqrt{C_1 C_2 K \log^5(NL)\mu^2(\Psi)/M}$. Then $\mathbb{E}\left[||B_1||_K\right] \leq \frac{a^2}{2} + a\sqrt{\frac{1}{2} + \frac{a^2}{4}}$. When $a \leq 1/2$, we get $\mathbb{E}\left[||B_1||_K\right] \leq a$. Let $0 < A' \leq a \leq 1/2$, and we can conclude that when

$$M \geq \frac{C_3 K \mu^2(\Psi) \log^5(NL)}{\delta'},$$

then

$$\mathbb{E}\left[||B_1||_K\right] \leq \delta'.$$

A.1.2 TAIL BOUND

Now we study the tail bound of $||B_1||_K$. First we construct a second set of random variables V_l', which are independent of V_l and are identically distributed as V_l. Additionally, we let

$$\widetilde{B}_1 = \sum_{i=1}^{M/2} \left(V_i V_i^H - V_i' V_i'^H\right),$$

and then according to Charles et al. (2014), there is

$$\mathbb{E}\left[\left|\left|\widetilde{B}_1\right|\right|_K\right] \leq 2\mathbb{E}\left[||B_1||_K\right], \tag{15}$$

$$\mathcal{P}\left[||\boldsymbol{B}_1||_K > 2\mathbb{E}\left[||\boldsymbol{B}_1||_K\right] + u\right] \leq 2\mathcal{P}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K > u\right]. \tag{16}$$

Now since we have

$$\left|\left|V_i V_i^H - V_i' V_i'^H\right|\right|_K \leq 2\max\{\left|\left|V_i V_i^H\right|\right|_K, \left|\left|V_i' V_i'^H\right|\right|_K\},$$

and

$$\left|\left|V_i V_i^H\right|\right|_K \leq \sup_{y \text{ is } K-sparse} ||V_i||_\infty^2 \frac{||y||_1^2}{||y||_2^2} \leq K V_{\max}^2,$$

then we know that $\max_{1 \leq i \leq M/2} \left|\left|V_i V_i^H\right|\right|_K \leq K V_{\max}^2$ and $\max_{1 \leq i \leq M/2} \left|\left|V_i' V_i'^H\right|\right|_K \leq K V_{\max}'^2$, where $V_{\max}' = \max_{1 \leq l \leq M/2} ||V_l'||_\infty$ then by Equation (13), we obtain

$$\mathcal{P}\left[\max_{1 \leq i \leq M/2} \left|\left|V_i V_i^H\right|\right|_K > \frac{K\mu^2(\boldsymbol{\Psi})}{M}\log\frac{MNL}{2\eta}\right] \leq \mathcal{P}\left[V_{\max}^2 > \frac{\mu^2(\boldsymbol{\Psi})}{M}\log\frac{MNL}{2\eta}\right] \leq \eta.$$

Since the probability theorems depend on bounded random variables, we define **F** to denote the following event

$$\mathbf{F} = \left\{\max\left\{\max_{1 \leq i \leq M/2} \left|\left|V_i V_i^H\right|\right|_K, \max_{1 \leq i \leq M/2} \left|\left|V_i' V_i'^H\right|\right|_K\right\} \leq \frac{K\mu^2(\boldsymbol{\Psi})}{M}\log\frac{MNL}{2\eta}\right\},$$

such that $\mathcal{P}\left[\mathbf{F}^C\right] \leq 2\eta$. Furthermore, we define $\mathbf{I}_F$ as the indicator function of **F**, and let $\widehat{\boldsymbol{B}}_1 = \sum_{i=1}^{M/2} \xi_i\left(V_i V_i^H - V_i' V_i'^H\right)\mathbf{I}_F$, where $\xi = \{\xi_i\}, i = 1, 2, ..., M/2$ is a Rademacher sequence and independent of V_i. The truncated variable $Y_i = \xi_i\left(V_i V_i^H - V_i' V_i'^H\right)\mathbf{I}_F$ has a symmetric distribution and $||Y_i||_K$ is bounded by $B_{\max} := \frac{2K\mu^2(\boldsymbol{\Psi})}{M}\ln\frac{MNL}{2\eta}$. By Proposition 19 in Tropp et al. (2009), we have

$$\mathcal{P}\left[\left|\left|\widehat{\boldsymbol{B}}_1\right|\right|_K > C_4(u\mathbb{E}\left[\left|\left|\widehat{\boldsymbol{B}}_1\right|\right|_K\right] + tB_{\max})\right] \leq e^{-u^2} + e^{-t}, \tag{17}$$

for all $u, t \geq 1$. Following Tropp et al. (2009), we find that

$$\mathcal{P}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K > v\right] \leq \mathcal{P}\left[\left|\left|\widehat{\boldsymbol{B}}_1\right|\right|_K > v\right] + \mathcal{P}\left[\mathbf{F}^C\right], \tag{18}$$

and

$$\mathbb{E}\left[\left|\left|\widehat{\boldsymbol{B}}_1\right|\right|_K\right] \leq \mathbb{E}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K\right]. \tag{19}$$

By combining Equations (17), (18), and (19), we get

$$\mathcal{P}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K > C_4(u\mathbb{E}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K\right] + tB_{\max})\right] \leq e^{-u^2} + e^{-t} + 2\eta.$$

In Equation (20), let $\eta < 1/e$, $u = \sqrt{\log\eta^{-1}}$ and $t = \log\eta^{-1}$. These values yield

$$\mathcal{P}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K > C_4(\sqrt{\log\eta^{-1}}\mathbb{E}\left[\left|\left|\widetilde{\boldsymbol{B}}_1\right|\right|_K\right] + (\log\eta^{-1})B_{\max})\right] \leq 4\eta. \tag{20}$$

16

Now we can combine Equations (15), (16), and (20) to get

$$\mathcal{P}\left[||\boldsymbol{B}_1||_K > 2\mathbb{E}\left[||\boldsymbol{B}_1||_K\right] + 2C_4\sqrt{\log \eta^{-1}}\mathbb{E}\left[||\boldsymbol{B}_1||_K\right] + C_4(\log \eta^{-1})B_{\max}\right] \leq 8\eta.$$

By choosing

$$M \geq \frac{C_3 K \mu^2(\boldsymbol{\Psi}) \log^5(NL)}{\delta'^2}, \tag{21}$$

where $\delta' < 1/2$, we obtain

$$\mathcal{P}\left[||\boldsymbol{B}_1||_K > 2\delta' + 2C_4\delta'\sqrt{\log \eta^{-1}} + C_5 \log \eta^{-1} \frac{\delta'^2 \log\left(\frac{1}{2}MNL\eta^{-1}\right)}{\log^5(NL)}\right] \leq 8\eta.$$

We observe now that

$$\frac{\log(\frac{1}{2}MNL\eta^{-1})}{\log^5(NL)} \leq \frac{2\log(NL) + \log \eta^{-1}}{\log^5(NL)},$$

indicating that when $\eta \geq (NL)^{-\log^4(NL)}$, i.e., $\log \eta^{-1} \leq \log^5(NL)$, then

$$\frac{\log(\frac{1}{2}MNL\eta^{-1})}{\log^5(NL)} \leq C_6,$$

, for a constant C_6. This inequality reduces our probability statement to

$$\mathcal{P}\left[||\boldsymbol{B}_1||_K > 2\delta' + 2C_4\delta'\sqrt{\log \eta^{-1}} + C_5 C_6 \log \eta^{-1}\delta'^2\right] \leq 8\eta.$$

For an arbitrary δ with $0 < \delta < 1/2$, we let

$$\delta' = \frac{\delta}{2C_4 C_7 \sqrt{\log \eta^{-1}}}, \tag{22}$$

resulting in

$$2C_4\delta'\sqrt{\log \eta^{-1}} = \frac{\delta}{C_7},$$

$$2\delta' = \frac{\delta}{C_4 C_7 \sqrt{\log \eta^{-1}}} \leq \frac{\delta}{C_4 C_7},$$

$$C_5 C_6 \log \eta^{-1}\delta'^2 = C_5 C_6 \frac{\delta^2}{4C_4^2 C_7^2} \leq \frac{C_5 C_6}{8C_4^2 C_7^2}\delta.$$

Then we can see that if

$$C_7 \geq \max\left\{3, \frac{3}{C_4}, \sqrt{\frac{3C_5 C_6}{8C_4^2}}\right\},$$

then

$$\mathcal{P}\left[||\boldsymbol{B}_1||_K > \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3}\right] < 8\eta.$$

17

Now by plugging (22) into (21), we know that when $(NL)^{-\ln^4 (NL)} \leq \eta < 1/e$, there exists a constant C such that when

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

there is

$$\mathrm{P}(\|\boldsymbol{B}_1\|_K > \delta) \leq 8\eta,$$

which completes the proof. ∎

A.1.3 LEMMAS FOR THEOREM 2

Lemma 4 *Suppose we have n complex Gaussian random variables, z_1, z_2, $\cdots z_n$, $z_i = x_i + \mathrm{j}y_i$, where x_i and y_i denote the real and imaginary parts of z_i. Let x_i and y_i i.i.d Gaussian r.v.'s with mean 0 and variance $1/2M$. ϕ_1, ϕ_2, $\cdots \phi_n$ r.v.'s which are independent of z_i for all i and satisfy*

$$\sum_{i=1}^{n} |\phi_i|^2 \leq \mu_0^2,$$

then for $w = \sum_{i=1}^{n} z_i \phi_i$,

$$\mathcal{P}\left[|w|^2 > u\right] \leq e^{-\frac{Mu}{\mu_0^2}}.$$

Proof We use x and y to denote the real and imaginary parts of w, and a_i and b_i to denote the real and imaginary parts of ϕ_i. We have

$$w = \sum_{i=1}^{n}(a_i x_i - b_i y_i) + \mathrm{j}\sum_{i=1}^{n}(a_i y_i + b_i x_i)x + \mathrm{j}y.$$

Then conditioned on a_i and b_i, x and y have distribution $\mathcal{N}(0, \frac{1}{2M}\sum_{i=1}^{n}(a_i^2 + b_i^2))$.

The next step is to prove the conditional independence of x and y. Since

$$\mathrm{Cov}(a_i x_i - b_i y_i, a_i y_i + b_i x_i | a_i, b_i) = \frac{a_i b_i}{2M} - \frac{a_i b_i}{2M} = 0,$$

when $i \neq j$, x_i, y_i are independent of x_j, y_j, and x and y are conditionally independent. Thus, conditioned on a_i and b_i,

$$\frac{2M}{\sum_{i=1}^{n}(a_i^2 + b_i^2)}|w|^2,$$

is χ^2 distributed. We use χ_2 to denote a two-degree χ^2 distributed random variable. According to the results on χ^2 distributions, we have

$$\mathcal{P}\left[|w|^2 > u|a_i, b_i\right] = \mathcal{P}\left[\chi_2 > \frac{2Mu}{\sum_{i=1}^{n}(a_i^2 + b_i^2)}|a_i, b_i\right] = e^{-\frac{Mu}{\sum_{i=1}^{n}(a_i^2+b_i^2)}} \leq e^{-\frac{Mu}{\mu_0^2}}. \quad (23)$$

Noticing that Equation (23) holds for all possible values of a_i and b_i completes the proof. ∎

Corollary 5 *For Q r.v.'s, w_1, w_2, $\cdots$, w_Q, let $w_i = \sum_{l=1}^{n} z_{il}\phi_{il}$ and $z_{il} = x_{il} + \mathrm{j}y_{il}$, where x_{il}, y_{il}, $1 \leq i \leq Q$, $1 \leq l \leq n$ are i.i.d. Gaussian distributed with mean 0 and variance $1/2M$. Suppose for any i, there is*

$$\sum_{l=1}^{n} |\phi_{il}|^2 \leq \mu_0^2.$$

And let $w_{\max} = \max_{1 \leq i \leq Q} |w_i|$, then for $\eta > 0$, we have

$$\mathcal{P}\left[w_{\max}^2 > \frac{\mu_0^2}{M}\log\frac{Q}{\eta} \right] \leq \eta,$$

and

$$\mathbb{E}\left[w_{\max}^2 \right] \leq \frac{\mu_0^2}{M}(\ln Q + 1).$$

Proof According to Lemma 4 and by using union bound, we have

$$\mathcal{P}\left[w_{\max}^2 > u \right] \leq Qe^{-\frac{Mu}{\mu_0^2}}.$$

Let $\eta = Qe^{-\frac{Mu}{\mu_0^2}}$ and there is

$$\mathcal{P}\left[w_{\max}^2 > \frac{\mu_0^2}{M}\log\frac{Q}{\eta} \right] \leq \eta.$$

Then we have

$$
\begin{aligned}
\mathbb{E}\left[w_{\max}^2 \right] &= \int_0^\infty \mathcal{P}\left[w_{\max}^2 > u \right] du \\
&\leq \int_0^{\frac{\mu_0^2}{M}\ln Q} 1\, du + \int_{\frac{\mu_0^2}{M}\log Q}^\infty Qe^{-\frac{Mu}{\mu_0^2}}\, du \\
&= \frac{\mu_0^2}{M}(\log Q + 1).
\end{aligned}
$$

∎

A.2 Proof of Low-rank Recovery

In this appendix we prove Theorem 3 where a low-rank input matrix S can be recovered from the network state $x[N]$ via nuclear norm optimization (Candès and Tao, 2010; Candès and Plan, 2010; Recht et al., 2010). To prove this theorem we use the dual certificate approach used to prove similar results in (Ahmed and Romberg, 2015; Candès and Plan, 2011a). In this methodology we seek a certificate Y whose projections into and out of the space spanned by the singular vectors of S are bounded appropriately. Specifically if we consider the singular value decomposition of S as

$$S = Q\Sigma V^*$$

and we consider the projection $\mathcal{P}_T$ which projects a matrix into the space T spanned by the left and right singular vectors,

$$\mathcal{P}_T\left(\boldsymbol{W}\right) = \boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{W} + \boldsymbol{W}\boldsymbol{V}\boldsymbol{V}^* - \boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{W}\boldsymbol{V}\boldsymbol{V}^* \tag{24}$$

the conditions for the dual certificate are that $\mathcal{A}$ is injective on T and there exists a matrix $\boldsymbol{Y}$ which satisfies

$$\left\|\mathcal{P}_T\left(\boldsymbol{Y}\right) - \boldsymbol{Q}\boldsymbol{V}^H\right\|_F \leq \frac{1}{2\sqrt{2}\gamma} \tag{25}$$

$$\left\|\mathcal{P}_{T^\perp}\left(\boldsymbol{Y}\right)\right\| \leq \frac{1}{2} \tag{26}$$

where the projection $\mathcal{P}_{T^\perp}$ is the projection onto the perpendicular space to T,

$$\mathcal{P}_{T^\perp}\left(\boldsymbol{W}\right) = \left(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^*\right)\boldsymbol{W}\left(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*\right)$$

The remainder of this proof will be devoted to demonstrating that there does exist a certificate $\boldsymbol{Y}$ by iteratively devising $\boldsymbol{Y}$ via a golfing scheme (Gross, 2011; Candès and Plan, 2011a; Ahmed et al., 2014). The golfing scheme essentially generates an iterative method which defined a series of certificate vectors $\boldsymbol{Y}_k$ for $k \in [1, \cdots, \kappa]$ which converge to a certificate $\boldsymbol{Y}_\kappa$ which satisfies the necessary conditions. As in (Ahmed and Romberg, 2015), we can initialize the 0th iterate to zero, and define the kth iterate in terms of the $\boldsymbol{Y}_{k-1}$ as

$$\boldsymbol{Y}_k = \boldsymbol{Y}_{k-1} + \kappa\mathcal{A}_k^*\mathcal{A}_k(\boldsymbol{Q}\boldsymbol{V}^* - \mathcal{P}_T\left(\boldsymbol{Y}_{k-1}\right)),$$

where

$$\mathcal{A}\left(\boldsymbol{W}\right) = \mathrm{vec}\left(\langle\boldsymbol{A}_n, \boldsymbol{W}\rangle\right). \tag{27}$$

We can see that since every iterate has $\mathcal{A}_k^*$ applied to it, every iteration is projected in to the range of $\mathcal{A}^*$, indicating that the final iteration $\boldsymbol{Y}$ will also be in the range of $\mathcal{A}^*$. In (Ahmed and Romberg, 2015), Asif and Romberg define a simpler iteration

$$\widetilde{\boldsymbol{Y}}_k = (\mathcal{P}_T - \kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T)\widetilde{\boldsymbol{Y}}_{k-1},$$

which is expressed in terms of the modified certificate

$$\widetilde{\boldsymbol{Y}}_k = \mathcal{P}_T\left(\boldsymbol{Y}_k\right) - \boldsymbol{Q}\boldsymbol{V}^*.$$

What remains now is to demonstrate that this iterative procedure converges, with high probability, to a certificate which satisfies the desired dual certificate conditions. We start by using Lemma 9 and observing that the Forbenious norm of the kth iterate is well bounded with probability $1 - O((LN)^{-\beta})$ by

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{Y}}_k\right\|_F &\leq \max_k \|\mathcal{P}_T - \kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T\| \left\|\widetilde{\boldsymbol{Y}}_{k-1}\right\|_F \\
&\leq 2^{-k}\left\|\widetilde{\boldsymbol{Y}}_0\right\|_F \\
&\leq 2^{-k}\left\|\boldsymbol{Q}\boldsymbol{V}^*\right\|_F \\
&\leq 2^{-k}\sqrt{R},
\end{aligned}$$

so long that $M \leq c\beta\kappa R(N + \mu_0^2 L)\log^2(LN)$. As in (Ahmed and Romberg, 2015) we observe that when we choose $\kappa \geq 0.5\log_2(8\gamma^2 R)$, the bound for the Frobenious norm of $\widetilde{Y}_\kappa$ is bounded by $\left|\left|\widetilde{Y}_\kappa\right|\right|_F \leq (2\sqrt{2}\gamma)^{-1}$.

To show that the second condition on the certificate is also satisfies, we apply Lemma 10. We begin with writing the quantity we wish to bound in terms of the past golfing scheme iterate

$$
\begin{aligned}
||\mathcal{P}_{T^\perp}(Y_\kappa)|| &\leq \sum_{k=1}^{\kappa}\left|\left|\mathcal{P}_{T^\perp}\left(\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1}\right)\right|\right| \\
&= \sum_{k=1}^{\kappa}\left|\left|\mathcal{P}_{T^\perp}\left(\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1} - \widetilde{Y}_{k-1}\right)\right|\right| \\
&\leq \sum_{k=1}^{\kappa}\left|\left|\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1} - \widetilde{Y}_{k-1}\right|\right| \\
&\leq \sum_{k=1}^{\kappa}\left|\left|\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1} - \widetilde{Y}_{k-1}\right|\right|_F \\
&\leq \sum_{k=1}^{\kappa}\max_{k\in[1,\dots\kappa]}\left|\left|\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1} - \widetilde{Y}_{k-1}\right|\right|_F \\
&\leq \sum_{k=1}^{\kappa}\frac{1}{2}2^{-k} \\
&\leq \frac{1}{2}
\end{aligned}
$$

We use Lemma 10 to bound the maximum spectral norm of $\kappa\mathcal{A}_k^*\mathcal{A}_k\widetilde{Y}_{k-1} - \widetilde{Y}_{k-1}$ with probability $1 - O((LN)^{1-\beta})$. Taking $\kappa \geq \log(LN)$ shows that the final certificate Y_κ satisfies all the desired properties, completing the proof.

A.2.1 MATRIX BERNSTEIN INEQUALITY AND OLICZ NORM

The lemmas required in our main result depend heavily on the matrix Bernstein inequality (Tropp, 2012). This inequality uses the variance measure and Oricz norm of a matrix to bound the largest singular value of the matrix. The matrix Bernstein inequality is summarized as

Theorem 6 (Matrix Bernstein's Inequality) *Let $X_i \in \mathbb{R}^{L,N}$, $i \in [1,\dots,M]$ be M random matrices such that $\mathbb{E}[X_i] = 0$ and $||X_i||_{\psi_\alpha} < U_\alpha < \infty$ for some $\alpha \geq 1$. Then with probability $1 - e^{-t}$, the spectral norm of the sum is bounded by*

$$
\left|\left|\sum_{i=1}^{M}X_i\right|\right| \leq C\max\left\{\sigma_X\sqrt{t + \log(L + N)}, U_\alpha\log^{1/\alpha}\left(\frac{MU_\alpha^2}{\sigma_X^2}\right)(t + \log(L + N))\right\},
$$

for some constant C and the variance parameter defined by

$$
\sigma_X = \max\left\{\left|\left|\sum_{i=1}^{M}\mathbb{E}[X_iX_i^*]\right|\right|^{1/2}, \left|\left|\sum_{i=1}^{M}\mathbb{E}[X_i^*X_i]\right|\right|^{1/2}\right\}.
$$

where Orlicz-α norm $||X||_{\psi_\alpha}$ is defined as

$$||X||_{\psi_\alpha} = \inf \left\{ y > 0 | \mathbb{E}\left[e^{||X||^\alpha/y^\alpha} \right] \leq 2 \right\}. \tag{28}$$

In particular we will use the matrix Bernstein inequality with the Orlicz-1 and Orlicz-2 norms, since subgaussian and subexponential random variables have bounded Orlicz-2 and -1 norms, respectively. To calculate these norms, we find the following lemmas from (Tropp, 2012; Ahmed and Romberg, 2015) useful:

Lemma 7 (Lemma 5.14, Tropp, 2012) *A random variable X is subgaussian iff X^2 is subexponential. Furthermore,*

$$||X||_{\psi_2}^2 \leq \left|\left| X^2 \right|\right|_{\psi_1} \leq 2 ||X||_{\psi_2}^2.$$

Lemma 8 (Lemma 7, Ahmed and Romberg, 2015) *Let X_1 and X_2 be two subgaussian ranfom variables. Then the product $X_1 X_2$ is a subexponential random variable with*

$$||X_1 X_2||_{\psi_1} \leq c ||X_1||_{\psi_2} ||X_2||_{\psi_2}.$$

Lemma 7 relates the Orlicz-1 and -2 norms for a random variable and it's square. Lemma 8 allows us to factor an Orlicz-1 norm of a sub-exponential random variable as the product of two subgaussian random variables. Finally we find useful the following calculation for the Orlicz-1 norm of the norm of a random Gaussian vector z_n with *i.i.d.* zero-mean and variance σ^2 entries:

$$\begin{aligned}
\left|\left| ||z_n||_2^2 \right|\right|_{\psi_1} &= \inf \left\{ y : \mathbb{E}\left[e^{||z_n||_2^2/y} \right] \leq 2 \right\} \\
&= \inf \left\{ y : \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} e^{-z_n^2(1/2\sigma^2 - 1/y)} dz_n \leq 2^{\frac{1}{M}} \right\} \\
&= \frac{2\sigma^2}{1 - 4^{-\frac{1}{M}}}. \tag{29}
\end{aligned}$$

A.2.2 BOUND ON $||\kappa \mathcal{P}_T \mathcal{A}_k^* \mathcal{A}_k \mathcal{P}_T - \mathcal{P}_T||$

Lemma 9 *Let $\mathcal{P}_T$ be defined as in Equation (24) and $\mathcal{A}_k$ be the restricted measurement operator as defined in Equation (27). Then if the number of nodes scale as*

$$M \geq c\beta\kappa R \left(N + \mu_0^2 L \right) \log^2(LN),$$

for a constant $\beta > 1$, then with probability greater then $1 - O(\kappa(LN)^{-\beta})$, we have

$$\max_{k \in [1,...,\kappa]} ||\kappa \mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - \mathcal{P}_T|| \leq \frac{1}{2}.$$

Proof

Lemma 9 bounds the operator norm

$$||\kappa \mathcal{P}_T \mathcal{A}_k^* \mathcal{A}_k \mathcal{P}_T - \mathcal{P}_T||.$$

Since $\mathbb{E}\left[\mathcal{A}_k^*\mathcal{A}_k\right] = \frac{1}{\kappa}\mathcal{I}$, this norm is equivalent to

$$
\begin{aligned}
\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathcal{P}_T &= \kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathbb{E}\left[\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T\right] \\
&= \kappa\sum_{n\in\Gamma_k}\left(\mathcal{P}_T(\boldsymbol{A}_n)\otimes\mathcal{P}_T(\boldsymbol{A}_n) - \mathbb{E}\left[\mathcal{P}_T(\boldsymbol{A}_n)\otimes\mathcal{P}_T(\boldsymbol{A}_n)\right]\right).
\end{aligned}
$$

We can also define here $\mathcal{L}_n(\boldsymbol{C}) = \langle\mathcal{P}_T(\boldsymbol{A}_n), \boldsymbol{C}\rangle\mathcal{P}_T(\boldsymbol{A}_n)$ which has $||\mathcal{L}_n|| = ||\mathcal{P}_T(\boldsymbol{A}_n)||_F^2$ which gives us

$$
\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathbb{E}\left[\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T\right] = \kappa\sum_{n\in\Gamma_k}\left(\mathcal{L}_n - \mathbb{E}\left[\mathcal{L}_n\right]\right).
$$

To calculate the variance, we can use the symmetry of $\mathcal{L}_n$ to only calculate

$$
\kappa^2\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\mathcal{L}_n^2\right] - \mathbb{E}\left[\mathcal{L}_n\right]^2\right\| \leq \kappa^2\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\mathcal{L}_n^2\right]\right\| = \kappa^2\left\|\mathbb{E}\left[\sum_{n\in\Gamma_k}\|\mathcal{P}_T(\boldsymbol{A}_n)\|_F^2\mathcal{L}_n\right]\right\|.
$$

We now need to bound $\|\mathcal{P}_T(\boldsymbol{A}_n)\|_F^2$, which can be done by the following:

$$
\begin{aligned}
\|\mathcal{P}_T(\boldsymbol{A}_n)\|_F^2 &= \langle\mathcal{P}_T(\boldsymbol{A}_n), \boldsymbol{A}_n\rangle \\
&= \langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*, \boldsymbol{z}_n\boldsymbol{f}_n^*\rangle + \langle\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{z}_n\boldsymbol{f}_n^*\rangle - \langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{z}_n\boldsymbol{f}_n^*\rangle \\
&= \|\boldsymbol{f}_n\|_2^2\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2 + \|\boldsymbol{z}_n\|_2^2\|\boldsymbol{V}^*\boldsymbol{f}_n\|_2^2 - \|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2\|\boldsymbol{V}^*\boldsymbol{f}_n\|_2^2 \\
&\leq N\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2 + \|\boldsymbol{z}_n\|_2^2\|\boldsymbol{V}^*\boldsymbol{f}_n\|_2^2.
\end{aligned}
$$

Using this calculation we can write

$$
\begin{aligned}
\left\|\mathbb{E}\left[\sum_{n\in\Gamma_k}\|\mathcal{P}_T(\boldsymbol{A}_n)\|_F^2\mathcal{L}_n\right]\right\| &\leq \left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\left(N\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2 + \|\boldsymbol{z}_n\|_2^2\|\boldsymbol{V}^*\boldsymbol{f}_n\|_2^2\right)\mathcal{L}_n\right]\right\| \\
&\leq N\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2\mathcal{L}_n\right]\right\| \\
&\qquad + \sup\|\boldsymbol{V}^*\boldsymbol{f}_n\|_\infty\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\|\boldsymbol{z}_n\|_2^2\mathcal{L}_n\right]\right\| \\
&\leq N\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2\mathcal{L}_n\right]\right\| + R\mu_0^2\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\|\boldsymbol{z}_n\|_2^2\mathcal{L}_n\right]\right\|.
\end{aligned}
$$

We now need to bound these two quantities. First we look to bound the first quantity

$$
\begin{aligned}
\left\|\sum_{n\in\Gamma_k}\mathbb{E}\left[\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2(\mathcal{P}_T(\boldsymbol{A}_n)\otimes\mathcal{P}_T(\boldsymbol{A}_n))\right]\right\| &\leq \|\mathcal{P}_T\|\left\|\mathbb{E}\left[\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2(\boldsymbol{A}_n\otimes\boldsymbol{A}_n)\right]\right\|\|\mathcal{P}_T\| \\
&\leq \left\|\mathbb{E}\left[\|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2\{z_n[\alpha]z_n^*[\beta]\boldsymbol{f}_n\boldsymbol{f}_n^*\}_{\alpha,\beta}\right]\right\|.
\end{aligned}
$$

Expanding, we have:

$$
\begin{aligned}
* &= \left\| \mathbb{E}\left[\left(\sum_{l=1}^{L} \|\boldsymbol{q}_l\|_2^2 |z_n[l]|^2 + 2\sum_{l\neq m} \mathrm{Re}(\langle \boldsymbol{q}_l, \boldsymbol{q}_m\rangle z_n[l] z_n^*[m]) \right) z_n[\alpha] z_n^*[\beta] \boldsymbol{I}_N \right]_{\alpha,\beta} \right\| \\
&= \left\| \left\{ \frac{1}{M^2}\sum_{l=1}^{L} \|\boldsymbol{q}_l\|_2^2 \boldsymbol{I}_N \delta_{\alpha=\beta} + \frac{2}{M^2}\langle \boldsymbol{q}_\alpha, \boldsymbol{q}_\beta\rangle \boldsymbol{I}_N \delta_{\alpha\neq\beta} \right\}_{\alpha,\beta} \right\| \\
&= \frac{1}{M^2}\left\| \left\{ \|\boldsymbol{Q}\|_F^2 \boldsymbol{I}_N \delta_{\alpha=\beta} + 2\langle \boldsymbol{q}_\alpha, \boldsymbol{q}_\beta\rangle \boldsymbol{I}_N \delta_{\alpha\neq\beta} \right\}_{\alpha,\beta} \right\|,
\end{aligned}
$$

giving us

$$
\left\| \sum_{n\in\Gamma_k} \mathbb{E}\left[\|\boldsymbol{Q}^* \boldsymbol{z}_n\|_2^2 (\mathcal{P}_T(\boldsymbol{A}_n) \otimes \mathcal{P}_T(\boldsymbol{A}_n)) \right] \right\| \leq \frac{1}{M\kappa}\|\boldsymbol{Q}\|_F^2 + \|\boldsymbol{Q}\boldsymbol{Q}^*\| \leq \frac{R+1}{M\kappa}.
$$

Similarly, for the second term we can take $\boldsymbol{Q} = \boldsymbol{I}_L$ to get

$$
\left\| \sum_{n\in\Gamma_k} \mathbb{E}\left[\|\boldsymbol{z}_n\|_2^2 (\mathcal{P}_T(\boldsymbol{A}_n) \otimes \mathcal{P}_T(\boldsymbol{A}_n)) \right] \right\| \leq \frac{1}{M\kappa}\|\boldsymbol{I}\|_F^2 = \frac{L}{M\kappa}.
$$

Putting the pieces together, we get

$$
\sigma_X^2 = \kappa R \frac{N + \mu_0^2 L}{M}.
$$

To use the matrix Bernstein inequality, it now remains to bound the following Orlicz-1 norm: $\kappa \left\| \mathcal{L}_n - \mathbb{E}\left[\mathcal{L}_n\right] \right\|_{\psi_1}$. By the PSD quality of $\mathcal{L}_n$ and its expectation,

$$
\left\| \mathcal{L}_n - \mathbb{E}\left[\mathcal{L}_n\right] \right\|_{\psi_1} \leq \max\left\{ \|\mathcal{L}_n\|_{\psi_1} - \|\mathbb{E}\left[\mathcal{L}_n\right]\|_{\psi_1} \right\}.
$$

The norm of $\|\mathbb{E}\left[\mathcal{L}_n\right]\|$ can be calculated via

$$
\|\mathbb{E}\left[\mathcal{L}_n\right]\| = \|\mathbb{E}\left[\mathcal{P}_T(\boldsymbol{A}_n)\right]\|_F^2 = \mathbb{E}\left[\sum_m |z_n[m]|^2 |f_n[m]|^2 \right] = \frac{1}{M}\|\boldsymbol{f}_n\|_2^2 = \frac{N}{M},
$$

indicating that the second term is simply $\|\mathbb{E}\left[\mathcal{L}_n\right]\|_{\psi_1} = N/(M\log(2))$. To calculate $\|\mathbb{E}\left[\mathcal{L}_n\right]\|_{\psi_1}$, we use the definition of the Orlitcz-1 norm in Equation (28) to see that

$$
\|\mathcal{L}_n\|_{\psi_1} = \left\| \|\mathcal{P}_T(\boldsymbol{A}_n)\|_2^2 \right\|_{\psi_1} \leq N\left\| \|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2 \right\|_{\psi_1} + RN\mu_0^2 \left\| \|\boldsymbol{z}_n\|_2^2 \right\|_{\psi_1},
$$

where the inequality stems form the fact that $\|\boldsymbol{V}^*\boldsymbol{f}_n\|_2^2 \leq RN\mu_0^2$ and $\|\boldsymbol{f}_n\|_2^2 \leq N$. Using the result in Equation (29) with $\sigma^2 = 1/M$ in the first term and in $\sigma^2 = R/M$ in the second term yields

$$
\begin{aligned}
\|\mathbb{E}\left[\mathcal{L}_n\right]\|_{\psi_1} &\leq N\left\| \|\boldsymbol{Q}^*\boldsymbol{z}_n\|_2^2 \right\|_{\psi_1} + RN\mu_0^2 \left\| \|\boldsymbol{z}_n\|_2^2 \right\|_{\psi_1} \\
&\leq \frac{1}{M}\left(\frac{N}{M(1 - 4^{-\frac{1}{R}})} + \frac{R\mu_0^2}{1 - 4^{-\frac{1}{L}}} \right) \\
&\leq \frac{2R\left(N + L\mu_0^2\right)}{\log(2)M}.
\end{aligned}
$$

We now have appropriate bounds on both the variance and Orliscz norm, permitting a bound on the largest singular value via the Matrix Bernstein inequality. Specifically, we can see that the first term in Theorem 6 with $t = \beta \log(LN) > \log(N + L)$ is bounded as

$$\sigma_X \sqrt{t + \log(L + N)} \leq \sqrt{2\kappa R\beta \frac{N + \mu_0^2 L}{M} \log(LN)}.$$

Likewise we can bound the second term:

$$
\begin{aligned}
U_1 \log\left(\frac{MU_1^2}{\sigma_X^2}\right)(t + \log(L + N)) &\leq 2\beta U_1 \log\left(\frac{4\Delta\kappa R(N + \mu_0^2 L)}{\log^2(2)M}\right)\log(LN) \\
&\leq c\frac{\beta\kappa R(N + \mu_0^2 L)}{M}\log^2(LN).
\end{aligned}
$$

Thus to appropriately bound

$$\|\kappa\mathcal{P}_T\mathcal{A}_k^*\mathcal{A}_k\mathcal{P}_T - \mathcal{P}_T\| \leq c\max\left\{\sqrt{\frac{\kappa R\beta(N + \mu_0^2 L)}{M}\log(LN)}, \frac{\beta\kappa R(N + \mu_0^2 L)}{M}\log^2(LN)\right\},$$

we can see that we would need

$$M \geq C\beta\kappa R(N + \mu_0^2 L)\log^2(LN).$$

Taking the union bound over the κ partitions completes the proof of the lemma. ∎

A.2.3 BOUND ON $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{G})\|$

Lemma 10 *Let $\mathcal{A}_k$ be defined as in Equation (27), $\kappa < M$ be the number of steps in the golfing scheme and assume that $M \leq LN$. Then as long as*

$$M \geq c\beta\kappa \max\left(N + \mu_0^2 L\right)\log^2(NL),$$

where μ_k^2 is the coherence term defined by

$$\mu_k^2 = R^{-1}\sup_{\omega\in[0,2\pi]}\left\|\widetilde{\boldsymbol{Y}}_k^*\boldsymbol{f}_\omega\right\|_2^2, \tag{30}$$

then with probability at least $1 - O(M(LN)^{-\beta})$, we have

$$\max_k\left\|\kappa\mathcal{A}_k^*\mathcal{A}_k(\widetilde{\boldsymbol{Y}}_{k-1}) - \widetilde{\boldsymbol{Y}}_{k-1}\right\| \leq 2^{-(k+1)}.$$

Proof

Lemma 10 essentially bounds the operator norm of $\kappa\mathcal{A}^*\mathcal{A} - \mathcal{I}$. In particular, to prove Theorem 2, the reduced version with $\kappa = 1$ is needed. Lemma 2 essentially uses the matrix Bernstein inequality to accomplish this task, taking

$$X_n = \kappa(\langle\boldsymbol{G}, \boldsymbol{A}_n\rangle\boldsymbol{A}_n - \mathbb{E}\left[\langle\boldsymbol{G}, \boldsymbol{A}_n\rangle\boldsymbol{A}_n\right]),$$

and we just need to control $||\sum \mathbb{E}[X_n X_n^*]||$ and $||\sum \mathbb{E}[X_n^* X_n]||$. To bound the second of these, we can calculate

$$
\begin{aligned}
\left\| \sum_{n \in \Gamma_k} \mathbb{E}[X_n^* X_n] \right\| &\leq \kappa^2 \left\| \sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle|^2 \boldsymbol{A}_n \boldsymbol{A}_n^*\right] \right\| \\
&= \kappa^2 \left\| \sum_{n \in \Gamma_k} \mathbb{E}\left[\|\boldsymbol{f}_n\|_2^2 |\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] \right\| \\
&\leq \frac{3N\kappa}{M} \|\boldsymbol{G}\|_F^2 ,
\end{aligned}
$$

where the second inequality is due to Lemma 11 and $\|\boldsymbol{f}_n\|_2^2 \leq N$. For the other expectation

$$
\begin{aligned}
\left\| \sum_{n \in \Gamma_k} \mathbb{E}[X_n X_n^*] \right\| &\leq \kappa^2 \left\| \sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle|^2 \boldsymbol{A}_n^* \boldsymbol{A}_n\right] \right\| \\
&= \frac{L\kappa^2}{M^2} \left\| \sum_{n \in \Gamma_k} \mathbb{E}\left[\|\boldsymbol{G}\boldsymbol{f}_n\|_2^2 \boldsymbol{f}_n \boldsymbol{f}_n^*\right] \right\| \\
&\leq \frac{L\kappa^2}{M^2} \sup_\omega(\|\boldsymbol{G}\boldsymbol{f}_\omega\|_2^2) \left\| \sum_{n \in \Gamma_k} \mathbf{1}_N \right\| \\
&\leq \frac{L\kappa}{M} \sup_\omega(\|\boldsymbol{G}\boldsymbol{f}_\omega\|_2^2) \\
&= \frac{L\kappa}{M} \mu^2 \|\boldsymbol{G}\|_F^2.
\end{aligned}
$$

Using these bounds, we can write

$$
\sigma_X^2 \leq \frac{\kappa}{M} \|\boldsymbol{G}\|_F^2 \max\{\mu_0 L, 3N\} ,
$$

and to use Proposition 1 we just need to bound $\|X\|_{\psi_2}$. To start, we can see that

$$
\begin{aligned}
U_1 &= \|X\|_{\psi_1} \leq 2\kappa \|\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle \boldsymbol{A}_n\|_{\psi_1} \\
&\leq c\kappa \|\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle\|_{\psi_2} \|\|\boldsymbol{A}_n\|_F\|_{\psi_2} \\
&\leq c\kappa \|\langle \boldsymbol{G}, \boldsymbol{A}_n \rangle\|_{\psi_2} \sqrt{\left\|\|\boldsymbol{f}_n\|_2^2 \|\boldsymbol{z}_n\|_2^2\right\|_{\psi_2}} \\
&= c\kappa \sqrt{\frac{N}{M(1 - 4^{-1/L})}} \|\|\mathrm{trace}(\boldsymbol{f}_n \boldsymbol{z}_n^* \boldsymbol{G})\|_{\psi_2} \\
&\leq c\kappa \sqrt{\frac{N}{M^2(1 - 4^{-1/L})}} \sum_{l=1}^{L} \|\langle \boldsymbol{g}_l, \boldsymbol{f}_n \rangle\|_{\psi_2} \\
&\leq c\kappa \sqrt{\frac{N^2 L \mu_0^2 \|\boldsymbol{G}\|_F^2}{M^2}}.
\end{aligned}
$$

We can now apply the matrix Bernstein theorem with the calculated values of U_1 and σ_X. Again using $t = \beta \log(LN)$, the first portion of the bound is

$$\sigma_X \sqrt{t + \log(L + N)} \leq c \, \|\boldsymbol{G}\|_F \sqrt{\frac{\kappa\beta}{M} \max\{\mu_k^2 L, N\} \log(LN)},$$

and the second portion of the bound is

$$U_1 \log\left(\frac{\Delta U_1^2}{\sigma_X}\right)(t + \log(L + N))$$

$$\leq c \, \|\boldsymbol{G}\|_F \, \kappa \sqrt{\frac{LN\mu_k^2}{M}} \log\left(\frac{LN\mu_k^2}{M^2} \frac{Mc \, \|\boldsymbol{G}\|_F^2 \, \kappa^2}{\kappa \, \|\boldsymbol{G}\|_F^2 \max\{\mu_k^2 L, N\}}\right) \beta \log(LN)$$

$$\leq c \, \|\boldsymbol{G}\|_F \, \kappa \sqrt{\frac{LN\mu_k^2}{M}} \log\left(\frac{c\Delta\kappa LN\mu_k^2}{M \max\{\mu_k^2 L, N\}}\right) \beta \log(LN)$$

$$\leq c\beta \, \|\boldsymbol{G}\|_F \, \kappa \sqrt{\frac{LN\mu_k^2}{M}} \log\left(\min\{\mu_k^2 L, N\}\right) \log(LN).$$

We can now use Lemma 12 to bound $\mu_k^2 \leq \mu_0^2$ with probability $1 - O(M(LN)^{-\beta})$ and Lemma 9 to bound $\|\boldsymbol{G}_k\|_F \leq 2^{-k}\sqrt{R}$, which gives us a bound of

$$\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\boldsymbol{G}\| \leq c2^{-k/2} \max\left\{\sqrt{\frac{\kappa\beta R \log(LN)}{M}} \log(\max\{\mu_0^2 L, N\}),\right.$$

$$\left.\sqrt{\mu_0^2 LN \frac{\beta\kappa}{M}} \log(LN) \log(\min\{\mu_0^2 L, N\})\right\}.$$

Simplifying the bound using $R \leq \min\{L, N\}$,

$$\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\boldsymbol{G}\| \leq c2^{-k/2} \max\left\{\sqrt{\frac{\kappa\beta R \log(LN)}{M}} \log(LN), \sqrt{\mu_0^2 LN \frac{\beta\kappa}{M}} \log^2(LN))\right\}$$

Taking

$$M \geq c\beta\kappa R \max\{N, L\mu_0^2\} \log^2(LN),$$

proves the lemma. To simplify the bound on the probability, we note that Lemma 12 holds with probability $1 - O(M(LN)^{-\beta})$ and this lemma holds with probability $1 - O(\kappa(LN)^{-\beta})$. Since $\kappa < M$ and assuming that $M \leq LN$, we can write that the result holds with probability $1 - O((LN)^{1-\beta})$. Additionally, since Lemma 12 holds when

$$M \geq c\beta\kappa R \left(N + L\mu_0^2\right) \log^2(LN) \geq c\beta\kappa R \max\{N, L\mu_0^2\} \log^2(LN),$$

Then both lemmas hold under the same condition. ∎

27

A.2.4 BOUND ON $\mathbb{E}\left[|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right]$

Lemma 11 bounds the spectrum of the expected matrix

$$\mathbb{E}\left[|\langle \boldsymbol{G}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] :$$

Lemma 11 *Suppose* $\boldsymbol{A}_n = \boldsymbol{z}_n \boldsymbol{f}_n^*$ *be defined as the outer product of an* i.i.d. *random Gaussian vector* $\boldsymbol{z}_n$ *with zero mean and variance* $1/M$ *and a random Fourier vector* $\boldsymbol{f}_n$. *Then the operator* $|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*$ *satisfies*

$$\mathbb{E}_z\left[|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] \preceq \frac{3}{M^2}\|\boldsymbol{C}^*\boldsymbol{f}_n\|_2^2 \boldsymbol{I}_M,$$

and

$$\mathbb{E}_{z,f}\left[|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] \preceq \frac{3}{M^2}\|\boldsymbol{C}\|_F^2 \boldsymbol{I}_M.$$

Proof

To begin the proof, we look at the expectation of each element of the matrix. We first calculate the expectation with respect to $\boldsymbol{z}_n$,

$$
\begin{aligned}
* &= \mathbb{E}_z\left[\left|\sum_{l=1}^{L} z_n[l]\boldsymbol{c}_l^*\boldsymbol{f}_n\right|^2 z_n[\alpha]z_n^*[\beta]\right] \\
&= \mathbb{E}_z\left[\left(\sum_{l=1}^{L} z_n[l]\boldsymbol{c}_l^*\boldsymbol{f}_n\right)^* \left(\sum_{l=1}^{L} z_n[l]\boldsymbol{c}_l^*\boldsymbol{f}_n\right) z_n[\alpha]z_n^*[\beta]\right] \\
&= \mathbb{E}_z\left[\sum_{l=1}^{L} |z_n[l]|^2|\boldsymbol{c}_l^*\boldsymbol{f}_n|^2 z_n[\alpha]z_n^*[\beta] + 2\sum_{k\neq l}\mathrm{Re}\left(z_n^*[l]z_n[k]\langle \boldsymbol{c}_l^*\boldsymbol{f}_n, \boldsymbol{c}_k^*\boldsymbol{f}_n\rangle\right) z_n[\alpha]z_n^*[\beta]\right] \\
&= \left(\frac{3}{2M^2}|\boldsymbol{c}_\alpha^*\boldsymbol{f}_n|^2 + \frac{1}{M^2}\|\boldsymbol{C}\boldsymbol{f}_n\|_2^2\right)\delta_{\alpha=\beta} + \frac{2}{M^2}\langle \boldsymbol{c}_\alpha^*\boldsymbol{f}_n, \boldsymbol{c}_\beta^*\boldsymbol{f}_n\rangle\delta_{\alpha\neq\beta}.
\end{aligned}
$$

We can then use the matrix formulation

$$
\begin{aligned}
\mathbb{E}_z\left[|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] &= \frac{3}{2M^2}\mathrm{diag}(\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^*) + \frac{1}{M^2}\|\boldsymbol{C}\boldsymbol{f}_n\|_2^2\boldsymbol{I}_M \\
&\qquad + \frac{2}{M^2}\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^* + \frac{2}{M^2}\mathrm{diag}(\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^*) \\
&= \frac{1}{M^2}\|\boldsymbol{C}\boldsymbol{f}_n\|_2^2\boldsymbol{I}_M + 2\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^* - \frac{1}{2}\mathrm{diag}(\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^*) \\
&\preceq \frac{3}{M^2}\|\boldsymbol{C}\boldsymbol{f}_n\|_2^2\boldsymbol{I}_M,
\end{aligned}
$$

where to obtain the result we first use the linearity of the expectation along with the with the positive-semidefinite property of $\mathrm{diag}(\boldsymbol{C}\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{C}^*)$, proving the fist portion of the Lemma. To prove the second portion we simply take an expectation with respect to $\boldsymbol{f}_n$:

$$\mathbb{E}_{z,f}\left[|\langle \boldsymbol{C}, \boldsymbol{A}_n\rangle|^2 \boldsymbol{z}_n \boldsymbol{z}_n^*\right] \preceq \mathbb{E}_f\left[\frac{3}{M^2}\|\boldsymbol{C}\boldsymbol{f}_n\|_2^2\boldsymbol{I}_M\right] \preceq \frac{3}{M^2}\|\boldsymbol{C}\|_F^2\boldsymbol{I}_{,M},$$

completing the proof. ∎

A.2.5 CONTRACTIVE PROPERTY OF μ_k^2

Lemma 12 *Let μ_k^2 be the coherence factor as defined in Equation* (30), *and additionally assume that $L > 1$ and that $LN > R\mu_0^4$. If*

$$M \geq c\beta\kappa R \left(N + L\mu_0^2\right) \log^2(LN),$$

then with probability at least $1 - O(\kappa(LN)^{-\beta})$,

$$\mu_k^2 \leq 2^{-1}\mu_{k-1}^2,$$

for all $k \in [1, \cdots, \kappa]$.

Proof

   In Lemma 12 we show that the coherence term reduces at each golfing iteration. Observe that

$$\mu_k^2 = \frac{1}{R}\sup_\omega \sum_{l=1}^{L}\langle \widetilde{\boldsymbol{Y}}_k, \boldsymbol{e}_l\boldsymbol{f}^*\rangle^2$$

$$= \frac{1}{R}\sup_\omega \sum_{l=1}^{L}\left(\sum_{n\in\Gamma_k}\kappa\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle - \langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{e}_l\boldsymbol{f}^*\rangle\right)^2$$

$$= \frac{1}{R}\sup_\omega \sum_{l=1}^{L}\left(\sum_{n\in\Gamma_k}\kappa\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle - \mathbb{E}\left[\kappa\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle\right]\right)^2.$$

To bound this quantity we use the scalar Bernstein inequality on each of the inner quantities

$$\sum_{n\in\Gamma_k} X_n = \sum_{n\in\Gamma_k}\kappa\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle - \mathbb{E}\left[\kappa\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle\right].$$

As in the matrix Bernstein formulation, we require both the variance and Orlicz norm. First we find the variance,

$$\sum_{n\in\Gamma_k}\mathbb{E}\left[X_nX_n^*\right] = \kappa^2 \sum_{n\in\Gamma_k}\mathbb{E}\left[|\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle|^2|\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right]$$

$$-|\mathbb{E}\left[\langle\mathcal{P}_T\left(\boldsymbol{A}_n\right), \boldsymbol{e}_l\boldsymbol{f}^*\rangle\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle\right]|^2$$

$$\leq \kappa^2 \sum_{n\in\Gamma_k}\mathbb{E}\left[|\langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*, \boldsymbol{e}_l\boldsymbol{f}^*\rangle + \langle\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}, \boldsymbol{e}_l\boldsymbol{f}^*\rangle\right.$$

$$\left.+\langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{e}_l\boldsymbol{f}^*\rangle|^2|\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right]$$

$$\leq \kappa^2 \sum_{n\in\Gamma_k}\mathbb{E}\left[\left(|\langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*, \boldsymbol{e}_l\boldsymbol{f}^*\rangle|^2 + |\langle\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{f}^*\rangle z_n[l]|^2\right.\right.$$

$$\left.\left.+|\langle\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{e}_l\boldsymbol{f}^*\rangle|^2\right)|\langle\widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right].$$

29

This sum consists of three terms. The first of which can be bounded using Lemma 11,

$$
\sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{Q}\boldsymbol{Q}^* \boldsymbol{z}_n \boldsymbol{f}_n^*, \boldsymbol{e}_l \boldsymbol{f}^* \rangle|^2 |\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n \rangle|^2\right]
$$

$$
= \sum_{n \in \Gamma_k} \mathbb{E}\left[|\boldsymbol{f}_n^* \boldsymbol{f} \langle \boldsymbol{q}_l, \boldsymbol{Q}^* \boldsymbol{z}_n \rangle|^2 |\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n \rangle|^2\right]
$$

$$
\leq 3 \sum_{n \in \Gamma_k} \mathbb{E}\left[\boldsymbol{f}^* \boldsymbol{f}_n \boldsymbol{f}_n^* \boldsymbol{f} \boldsymbol{q}_l^* \boldsymbol{Q}^* \left\|\widetilde{\boldsymbol{Y}}_{k-1} \boldsymbol{f}_n\right\|_2^2 \boldsymbol{I}_L \boldsymbol{Q} \boldsymbol{q}_l\right]
$$

$$
\leq \frac{3R\mu_{k-1}^2}{M^3} \|\boldsymbol{q}_l\|_2^2 \sum_{n \in \Gamma_k} \boldsymbol{f}^* \mathbb{E}\left[\boldsymbol{f}_n \boldsymbol{f}_n^*\right] \boldsymbol{f}
$$

$$
= \frac{3NR\mu_{k-1}^2}{\kappa M^2} \|\boldsymbol{q}_l\|_2^2 \, .
$$

For the second term we have

$$
\sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{f}_n^* \boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{f}^* \rangle|^2 |z_n[l]|^2 |\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n \rangle|^2\right]
$$

$$
= \sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{V}^* \boldsymbol{f}_n, \boldsymbol{V}^* \boldsymbol{f} \rangle|^2 |z_n[l]|^2 |\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n \rangle|^2\right]
$$

$$
= \sum_{n \in \Gamma_k} \mathbb{E}\left[\boldsymbol{f}^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f}_n \boldsymbol{f}_n^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f} |\langle \widetilde{\boldsymbol{Y}}_{k-1}, z_n[l] \boldsymbol{z}_n \boldsymbol{f}_n^* \rangle|^2\right] \, .
$$

Using the fact that $|z_n[l]|^2 = \boldsymbol{e}_l^* \boldsymbol{z}_n \boldsymbol{z}_n^* \boldsymbol{e}_l$ and Lemma 11, we obtain

$$
\sum_{n \in \Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{f}_n^* \boldsymbol{V}\boldsymbol{V}^*, \boldsymbol{f}^* \rangle|^2 |z_n[l]|^2 |\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n \rangle|^2\right]
$$

$$
\leq \frac{3}{M^2} \sum_{n \in \Gamma_k} \mathbb{E}\left[\boldsymbol{f}^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f}_n \boldsymbol{f}_n^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f} \boldsymbol{e}_l^* \left\|\widetilde{\boldsymbol{Y}}_{k-1} \boldsymbol{f}_n\right\|_2^2 \boldsymbol{I}_L \boldsymbol{e}_l\right]
$$

$$
\leq \frac{3R\mu_{k-1}^2}{M^2} \sum_{n \in \Gamma_k} \mathbb{E}\left[\boldsymbol{f}^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f}_n \boldsymbol{f}_n^* \boldsymbol{V}\boldsymbol{V}^* \boldsymbol{f}\right]
$$

$$
\leq \frac{3R\mu_{k-1}^2}{M^2} |\Gamma_k| \|\boldsymbol{V}^* \boldsymbol{f}\|_2^2
$$

$$
\leq \frac{3R^2 \mu_{k-1}^2 \mu_0^2}{\kappa M^1} \, .
$$

Finally, for the third term, we have

$$\sum_{n\in\Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{QQ}^*\boldsymbol{z}_n\boldsymbol{f}_n^*\boldsymbol{VV}^*, \boldsymbol{e}_l\boldsymbol{f}^*\rangle|^2|\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right]$$

$$= \sum_{n\in\Gamma_k} \mathbb{E}\left[|\langle \boldsymbol{V}^*\boldsymbol{f}_n, \boldsymbol{V}^*\boldsymbol{f}\rangle|^2|\langle \boldsymbol{q}_l, \boldsymbol{Q}^*\boldsymbol{z}_n\rangle|^2|\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right]$$

$$= \sum_{n\in\Gamma_k} \mathbb{E}\left[\boldsymbol{f}^*\boldsymbol{VV}^*\boldsymbol{f}_n\boldsymbol{f}_n^*\boldsymbol{VV}^*\boldsymbol{f}|\langle \boldsymbol{q}_l, \boldsymbol{Q}^*\boldsymbol{z}_n\rangle|^2|\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\right]$$

$$\leq \sum_{n\in\Gamma_k} \mathbb{E}\left[||\boldsymbol{V}^*\boldsymbol{f}||_2^2\, \boldsymbol{q}_l^*\boldsymbol{Q}^*\boldsymbol{z}_n\boldsymbol{z}_n^*|\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle|^2\boldsymbol{Qq}_l\right]$$

$$\leq \frac{3}{M^2}\,||\boldsymbol{q}_l||_2^2\,||\boldsymbol{V}^*\boldsymbol{f}||_2^2\sum_{n\in\Gamma_k}\mathbb{E}\left[\left|\left|\widetilde{\boldsymbol{Y}}_{k-1}\boldsymbol{f}_n\right|\right|_2^2\right]$$

$$\leq \frac{3R^2\mu_0^2\mu_{k-1}^2}{\kappa M}\,||\boldsymbol{q}_l||_2^2\,.$$

Summing the three bounds and using $||\boldsymbol{q}_l|| \leq 1$ yields

$$\sigma_X^2 \leq 9\kappa\left(\frac{NR\mu_{k-1}^2}{M^2}\,||\boldsymbol{q}_l||_2^2 + 2\frac{R^2\mu_0^2\mu_{k-1}^2}{M}\right).$$

To use the Bernstein inequality it remains to find the Orlicz-1 norm of X_n. From Lemma 10 we have

$$\left|\left|\langle \widetilde{\boldsymbol{Y}}_{k-1}, \boldsymbol{A}_n\rangle\right|\right|_{\psi_2}^2 = \left|\left|\boldsymbol{z}_n^*\widetilde{\boldsymbol{Y}}_{k-1}\boldsymbol{f}_n\right|\right|_{\psi_2}^2 \leq \left|\left|||\boldsymbol{Q}^*\boldsymbol{z}_n||_2\,||\boldsymbol{\Lambda}_{k-1}\boldsymbol{V}^*\boldsymbol{f}_n||_2\right|\right|_{\psi_2}^2 \leq c\frac{R\mu_{k-1}^2}{M}.$$

For the first term we have

$$||\boldsymbol{f}_n^*\boldsymbol{f}\langle \boldsymbol{q}_l, \boldsymbol{Q}^*\boldsymbol{z}_n\rangle||_{\psi_2}^2 \leq |||\boldsymbol{f}_n^*||_2\,||\boldsymbol{f}||_2\,|\langle \boldsymbol{q}_l, \boldsymbol{Q}^*\boldsymbol{z}_n\rangle||_{\psi_2}^2$$

$$\leq N^2\,||\boldsymbol{q}_l^*\boldsymbol{Q}^*\boldsymbol{z}_n||_{\psi_2}^2$$

$$\leq c\frac{N^2\,||\boldsymbol{q}_l||_2^2}{M^2}.$$

For the second term we have

$$||\langle \boldsymbol{V}^*\boldsymbol{f}_n, \boldsymbol{V}^*\boldsymbol{f}\rangle z_n[l]||_{\psi_2}^2 \leq |||\boldsymbol{V}^*\boldsymbol{f}_n||_2\,||\boldsymbol{V}^*\boldsymbol{f}||_2\,z_n[l]||_{\psi_2}^2 \leq c\frac{R^2}{M}\mu_0^4.$$

Similarly, for the final term we have

$$||\langle \boldsymbol{V}^*\boldsymbol{f}_n, \boldsymbol{V}^*\boldsymbol{f}\rangle\langle \boldsymbol{q}_l, \boldsymbol{Q}^*\boldsymbol{z}_n\rangle||_{\psi_2}^2 \leq |||\boldsymbol{V}^*\boldsymbol{f}_n||_2\,||\boldsymbol{V}^*\boldsymbol{f}||_2\,\boldsymbol{q}_l^*\boldsymbol{Q}^*\boldsymbol{z}_n||_{\psi_2}^2 \leq c\frac{R^2}{M}\mu_0^4\,||\boldsymbol{q}_l||_2^2\,.$$

Now we can calculate the total Orlicz norm as

$$||X_n||_{\psi_1}^2 \leq c\kappa^2\frac{R\mu_{k-1}^2}{M}\left(\frac{N^2\,||\boldsymbol{q}_l||_2^2}{M^2} + \frac{R^2}{M}\mu_0^4 + \frac{R^2}{M}\mu_0^4\,||\boldsymbol{q}_l||_2^2\right)$$

$$\leq c\frac{\kappa^2RN^2}{M^3}\,||\boldsymbol{q}_l||_2^2\,\mu_{k-1}^2 + c\frac{2\kappa^2R^3}{M^2}\mu_0^4\mu_{k-1}^2.$$

Since we wish to bound the square of the sum of terms, we calculate the square values of the two terms in the Bernstein inequality. The first term is bounded by

$$t\sigma_X^2 \le c\beta\kappa \frac{R}{M}\mu_{k-1}^2 \left(\frac{N}{M}\|\boldsymbol{q}_l\|_2^2 + 2R\mu_0^2\right)\log(LN),$$

and the second term is bounded by

$$t^2 U_\alpha^2 \log^2\left(\frac{|\Gamma_k|U_\alpha^2}{\sigma_X^2}\right) \le t^2 U_\alpha^2 \log^2\left(c\frac{|\Gamma_k|M\kappa^2 R\mu_{k-1}^2\left(\frac{N^2}{M}\|\boldsymbol{q}_l\|_2^2 + 2R^2\mu_0^4\right)}{M^2\kappa R\mu_{k-1}^2\left(\frac{N}{M}\|\boldsymbol{q}_l\|_2^2 + 2R\mu_0^2\right)}\right)$$

$$\le t^2 c\kappa^2 \frac{R}{M^2}\mu_{k-1}^2 \left(\frac{N^2}{M}\|\boldsymbol{q}_l\|_2^2 + 2R^2\mu_0^4\right)\log^2\left(c\frac{N^2\|\boldsymbol{q}_l\|_2^2 + 2MR^2\mu_0^4}{N\|\boldsymbol{q}_l\|_2^2 + 2RM\mu_0^2}\right)$$

$$\le c\beta^2\kappa^2 \frac{R}{M^2}\mu_{k-1}^2 \left(\frac{N^2}{M}\|\boldsymbol{q}_l\|_2^2 + 2R^2\mu_0^4\right)\log^4\left(LN\right),$$

where the last step assumes $L > 1$ and $LN > R\mu_0^4$. Each summand is then bounded by the maximum of these two quantities with probability $1 - O(|\Gamma_k|(LN)^{-\beta})$, the $|\Gamma_k|$ term coming from the union bound over all terms in each inner sum.

Using this bound on each summand, we obtain the total bound by taking a union bound, summing over $l \in [1, \cdots, L]$, and dividing by R, yielding a bound of the maximum of

$$t\sigma_X^2 \le c\beta\kappa \frac{R}{M}\mu_{k-1}^2 \left(\frac{N}{M} + 2L\mu_0^4\right)\log(LN),$$

and

$$t^2 U_\alpha^2 \log^2\left(\frac{|\Gamma_k|U_\alpha^2}{\sigma_X^2}\right) \le \beta^2 c\kappa^2 \frac{R}{M^2}\mu_{k-1}^2 \left(\frac{N}{M} + 2RL\mu_0^2\right)\log^4\left(LN\right),$$

with probability $1 - O(M(NL)^{-\beta})$. To complete the proof, we note that if we have

$$M \ge c\beta\kappa R\left(N + L\mu_0^2\right)\log^2(LN),$$

then both terms in this bound are less than μ_{k-1}^2. ∎

References

M. Aharon, M. Elad, A. Bruckstein, and Y. Katz. K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, 2006.

A. Ahmed and J. Romberg. Compressive multiplexing of correlated signals. *IEEE Transactions on Information Theory*, 61(1):479–498, 2015.

A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

S.-I. Amari. Characteristics of random nets of analog neuron-like elements. *IEEE Transactions on Systems, Man and Cybernetics*, (5):643–657, 1972.

N. J. Apthorpe, A. J. Riordan, R. E. Aguilar, J. Homann, Y. Gu, D. W. Tank, and S. H. Seung. Automatic neuron detection in calcium imaging data using convolutional networks. *Advances in neural information processing systems (NIPS)*, pages 3270–3278, Dec. 2016.

A. Balavoine, J. Romberg, and C. J. Rozell. Convergence and rate analysis of neural networks for sparse approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23: 1377–1389, Sept. 2012.

P. Bashivan, I. Rish, M. Yeasin, and N. Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *International Conference on Learning Representations*, 2016.

A. Berényi, Z. Somogyvári, A. J. Nagy, L. Roux, J. D. Long, S. Fujisawa, E. Stark, A. Leonardo, T. D. Harris, and G. Buzsáki. Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. *Journal of neurophysiology*, 111(5):1132–1149, 2014.

M. Buehner and P. Young. A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824, 2006.

D. V Buonomano and W. Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113–125, 2009.

L. Büsing, B. Schrauwen, and R. Legenstein. Connectivity, dynamics, and memory in reservoir computing with binary and analog neurons. *Neural computation*, 22(5):1272–1311, 2010.

E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

E. J. Candès and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011a.

E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011b.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

E. J. Candès, J. Romberg, and Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52 (2):489–509, 2006.

A. S. Charles, P. Garrigues, and C. J. Rozell. A common network architecture efficiently implements a variety of sparsity-based inference problems. *Neural Computation*, 24(12):3317–3339, 2012.

A. S. Charles, H. L. Yap, and C. J. Rozell. Short-term memory capacity in networks via the restricted isometry property. *Neural computation*, 26(6):1198–1235, 2014.

P. Chen and D. Suter. Recovering the missing components in a large noisy low-rank matrix: Application to sfm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1051–1063, 2004.

Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6): 2094–2107, 2014.

Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

C.-K. Chiang, C.-H. Liu, C.-H. Duan, and S.-H. Lai. Learning component-level sparse representation for image and video categorization. *IEEE Transactions on Image Processing*, 22(12): 4775–4787, 2013.

C. Christopoulos, A. Skodras, and T. Ebrahimi. The jpeg2000 still image coding system: an overview. *IEEE transactions on consumer electronics*, 46(4):1103–1127, 2000.

M. E. Davies and L. Daudet. Sparse audio representations using the mclt. *Signal processing*, 86(3): 457–470, 2006.

W. Denk, J. H. Strickler, and W. W. Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 248(4951):73–6, 1990.

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

M. Elad, M. A. T. Figueiredo, and Y. Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6)972–982, 2010.

O. Faugeras, J. Touboul, and B. Cessac. A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Frontiers in computational neuroscience*, 3(1):1–26, Feb. 2009.

M. Fazel. Matrix rank minimization with applications. *PhD Thesis, Stanford University*, 2002.

M. Galtier and G. Wainrib. A local echo state property through the largest Lyapunov exponent. *Neural Networks*, 76:39–45, 2016.

S. Ganguli and H. Sompolinsky. Short-term memory in neuronal networks through dynamical compressed sensing. *Advances in Neural Information Processing Systems (NIPS)*, pages 667–675, 2010.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970, 2008.

Y. Gao, E. Archer, L. Paninski, and J. P. Cunningham. Linear dynamical neural population models through nonlinear embeddings. *Advances in Neural Information Processing Systems (NIPS)*, pages 163–171, Dec. 2016.

F. Gobet, P. C. R. Lane, S. Croker, P. C. H. Cheng, G. Jones, I. Oliver, and J. M. Pine. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6):236–243, 2001.

A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, 2013.

K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

T. Guha and R. K. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1576–1588, 2012.

X. Hinaut, M. Petit, G. Pointeau, and P. F. Dominey. Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Frontiers in Neurorobotics*, 8(16):1–17, 2014.

H. Jaeger. Short term memory in echo state networks. *GMD Report 152 German National Research Center for Information Technology*, 5:1–60, 2001.

H. Jaeger and H. Haas. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

M. Jaggi, M. Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. *International Conference on Machine Learning (ICML)*, pages 471–478, 2010.

N. Kalchbrenner, A. Graves, and I. Danihelka. Grid long short-term memory. *International Conference on Learning Representations*, 2016.

K. Kavukcuoglu, P. Sermanet, Y. L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. *Advances in Neural Information Processing Systems (NIPS)*, pages 1090–1098, 2010.

Y. LeCun, K. Kavukcuoglu, C. Farabet, et al. Convolutional networks and applications in vision. IEEE International Symposium on Circuits and Systems *ISCAS*, pages 253–256, 2010.

R. Legenstein and W. Maass. Edge of chaos and prediction of computational performance for neural circuit models. *Neural Networks*, 20(3):323–334, 2007.

Z. Lipton, D. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. *International Conference on Learning Representations*, 2016.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

M. Lukoševičius. A practical guide to applying echo state networks. *Neural Networks: Tricks of the Trade*, pages 659–686. 2012.

M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A*, 374(2065):20150203, 2016.

G. Manjunath and H. Jaeger. Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696, 2013.

R. Maruyama, K. Maeda, H. Moroda, I. Kato, M. Inoue, H. Miyakawa, and T. Aonishi. Detecting cells using non-negative matrix factorization on calcium imaging data. *Neural Networks*, 55:11–19, 2014.

M. Massar and S. Massar. Mean-field theory of echo state networks. *Physical Review E*, 87(4):042809, 2013.

B. A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609, Jun 1996.

A. B. Patel, T. Nguyen, and R. G. Baraniuk. A probabilistic theory of deep learning. *arXiv preprint arXiv:1504.00641*, 2015.

K. Rajan, L. F. Abbott, and H. Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 82(1):011903, 2010.

H. Rauhut. Compressive sensing and structured random matrices. *Theoretical Found. and Numerical Methods for Sparse Recovery*, 9:1–92, 2010.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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*, 20(10):2526–2563, Oct 2010.

H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Conference of International Speech Communication Association (INTERSPEECH)*, 2014.

S. Shapero, M. Zhu, P. Hasler, and C. J. Rozell. Optimal sparse approximation with integrate and fire neurons. *International Journal of Neural Systems*, 24(05):1440001, Aaug 2014.

A. Singer and M. Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259, 1988.

S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 2431–2439.

K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

J.A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, and R.G. Baraniuk. Beyond Nyquist: efficient sampling of sparse bandlimited signals. *IEEE Transactions on Information Theory*, 56(1), Jan. 2009. ISSN 0018-9448.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

P. Vardan, Y. Romano, and M. Elad. Convolutional neural networks analyzed via convolutional sparse coding. *arXiv preprint arXiv:1607.08194*, 2016.

M. A. Veganzones, M. Simoes, G. Licciardi, N. Yokoya, J.M. Bioucas-Dias, and J. Chanussot. Hyperspectral super-resolution of locally low rank images from complementary multisource data. *IEEE Transactions on Image Processing*, 25(1):274–288, 2016.

D. Verstraeten, B. Schrauwen, M. dHaene, and D. Stroobandt. An experimental unification of reservoir computing methods. *Neural Networks*, 20(3):391–403, 2007.

D. Verstraeten, J. Dambre, X. Dutoit, and B. Schrauwen. Memory versus non-linearity in reservoirs. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2010.

G. Wainrib. Context dependent representation in recurrent neural networks. *arXiv preprint http://arxiv.org/pdf/1506.06602.pdf*, 2015.

E. Wallace, R. M. Hamid, and P. E. Latham. Randomly connected networks have short temporal memory. *Neural Computation*, 25:1408–1439, 2013.

O. L. White, D. D. Lee, and H. Sompolinsky. Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14):148102, 2004.

H. R. Wilson and J. D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1, 1972.

I. B. Yildiz, H. Jaeger, and S. J. Kiebel. Re-visiting the echo state property. *Neural networks*, 35: 1–9, 2012.

H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4729–4743, 2014.