Toeplitz Compressed Sensing Matrices with Applications to Sparse Channel Estimation

Jarvis Haupt, Waheed U. Bajwa, Gil Raz, and Robert Nowak

Abstract

Compressed sensing (CS) has recently emerged as a powerful signal acquisition paradigm. In essence, CS enables the recovery of high-dimensional but sparse (or nearly sparse) vectors from relatively few linear observations in the form of projections of the signal onto a collection of test vectors. Existing results show, for example, that if the entries of the test vectors are independent realizations of random variables with certain distributions, such as zero-mean Gaussian, then with high probability the resulting observations sufficiently encode the information in the unknown signal and recovery can be accomplished by solving a tractable convex optimization. This work provides a significant extension of current CS theory. A novel technique is proposed that allows theoretical treatment of CS in settings where the entries of the test vectors exhibit structured statistical dependencies, from which it follows that CS can be effectively utilized in linear, time-invariant (LTI) system identification problems. An immediate application is in the area of sparse channel estimation, where the main results of this work can be applied to the recovery of sparse (or nearly sparse) wireless multipath channels. Specifically, it is shown in the paper that time-domain probing of a wireless channel with a (pseudo-)random binary sequence, along with the utilization of CS reconstruction techniques, provides significant improvements in estimation accuracy when compared with traditional least-squares based linear channel estimation strategies. Abstract extensions of the main results, utilizing the theory of equitable graph coloring to tolerate more general statistical dependencies across the test vectors, are also discussed.

I. Introduction

Compressed sensing (CS) describes a new signal acquisition paradigm in which sparse, high-dimensional vectors $\boldsymbol{\beta} \in \mathbb{R}^n$ can be accurately recovered from a small number of linear observations of the form $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} \in \mathbb{R}^k$. The conditions under which CS succeeds depend on the structure of the measurement matrix $\boldsymbol{X} \in \mathbb{R}^{k \times n}$. Specifically, recent theoretical results establish that the observations $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ can be used to efficiently recover any "sparse enough" signal provided that the matrix $\boldsymbol{X}$ satisfies the so-called *restricted isometry property* (RIP) [1]–[3]. Essentially, the RIP condition guarantees that all column submatrices of $\boldsymbol{X}$ (the $k \times m$ submatrices formed by

This research was supported in part by the DARPA Analog-to-Information Program. This paper was presented in part at the 14th IEEE/SP Workshop on Statistical Signal Processing, Madison, WI, August 2007 and at the 42nd Annual Conference on Information Sciences and Systems, Princeton, NJ, March 2008.

JH, WUB and RN are with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706 USA. GR is with GMR Research and Technology, Concord, MA 01742 USA. (E-mails: jdhaupt@wisc.edu, bajwa@cae.wisc.edu, raz@gmrtech.com, nowak@engr.wisc.edu).

all subsets of m columns, where m is proportional to the level of sparsity in $\boldsymbol{\beta}$) are well-conditioned. The use of such measurement matrices to recover sparse vectors leads to remarkable results. First, the number of observations required for accurate recovery, k, is on the order of the number of nonzero entries in the signal, which can be far fewer than the ambient signal dimension n. In addition, signal recovery can be accomplished by solving a tractable convex optimization. And both of these results remain true even when $\boldsymbol{y}$ is corrupted by some (deterministic or stochastic) additive noise $\boldsymbol{\eta} \in \mathbb{R}^k$, where "recovery" in these settings means that the signal estimate is close to the true signal in terms of a suitable error metric (such as the mean squared error).

This paper explores whether the sparsity-exploiting power of CS can be used for the identification of discrete, linear, time-invariant (LTI) systems. Specifically, we examine the effectiveness of random probing of LTI systems having sparse impulse responses, coupled with the utilization of CS reconstruction methods. The practical importance of this problem is evidenced by many wireless communication applications in which the underlying multipath channel can be modeled as an LTI system with a sparse impulse response [4]. Compared to the conventional channel estimation methods that do not explicitly account for the underlying multipath sparsity, reliable estimation of the channel impulse response in these settings can lead to significant reductions in transmission energy and improvements in spectral efficiency.

Existing results in CS show that if the entries of the observation (measurement) matrix are independent realizations of random variables with certain distributions, such as zero-mean Gaussian, then with high probability the resulting matrix satisfies the RIP [5]. Toeplitz matrices are matrices having constant diagonals and they arise naturally in the context of estimation of wireless multipath channels. The major contribution of this work is a significant extension of CS theory to observation matrices containing statistical dependencies across rows, from which it follows that random Toeplitz matrices satisfy the RIP with high probability. As a result, recent advances from the theory of CS can be leveraged to devise quantitative error bounds for convex/linear programming based sparse channel estimation schemes. Our proofs rely on a novel technique that facilitates analysis of the (structured) statistical dependencies arising from the Toeplitz structure and, in general, our techniques can be applied in certain settings to obtain deviation bounds for both the ℓ_2-norms of random vectors having dependent entries, and inner products between certain dependent random vectors. The proofs and techniques outlined here generalize and build upon our own previous works, which were the first to provide theoretical performance guarantees for CS using random Toeplitz-structured matrices [4], [6].

The remainder of this paper is organized as follows. In the rest of this section, we provide the requisite background on existing results in CS. In Section II, we describe how estimation of sparse wireless multipath channels fits into the CS framework. In particular, we show that time-domain probing of a wireless channel with a (pseudo-)random binary sequence, along with the utilization of CS reconstruction techniques, provides significant improvements in estimation accuracy when compared with traditional least-squares based linear channel estimation strategies. The major contributions of the paper appear in Section III, where we establish the RIP for random Toeplitz matrices comprised of either Gaussian or bounded random variables. Finally, in Section IV, we present extensions of the main results of the paper to accommodate more general statistical dependencies, and we discuss connections with

previous works.

A. Compressed Sensing and the Restricted Isometry Property

Consider the problem of recovering an unknown vector $\boldsymbol{\beta} \in \mathbb{R}^n$ from a collection of linear observations,

$$y_j = \sum_{i=1}^{n} x(j)_i \ \beta_i = \boldsymbol{x}(j)'\boldsymbol{\beta}, \quad j = 1, \ldots, k \tag{1}$$

where the vectors $\boldsymbol{x}(j) \in \mathbb{R}^n$ are specified "test vectors." The observation process can be written more compactly in the matrix formulation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{y} \in \mathbb{R}^k$ and $\boldsymbol{X}$ is the $k \times n$ observation matrix whose rows are the test vectors. This very general signal model encompasses a wide variety of applications, including magnetic resonance imaging, digital imaging, radio frequency surveillance. When the number of observations, k, equals or exceeds the dimension of the unknown signal, n, (the so-called overdetermined setting) then results from classical linear algebra show that any unknown signal can be recovered exactly using a suitable set of test vectors. The complete set of basis vectors from any orthonormal transform suffices, for example.

Compressed sensing (CS) primarily addresses the question of what is possible when the number of observations is *fewer* than the dimension of the unknown signal, the so-called underdetermined setting. The seminal works in CS established that signals can theoretically be recovered exactly from such incomplete observations, provided the signals are sparse. Further, CS is a viable practical technique because recovery is tractable—it can be accomplished by convex optimization [7]–[9].

Proofs of these remarkable initial results all rely on the same properties of the observation matrix, namely that the submatrix of the observation matrix corresponding to the true signal subspace should behave almost like an orthonormal matrix. One concise way to quantify this is by the restricted isometry property (RIP), first introduced in [1]. The RIP, defined below, can be leveraged to establish a series of fundamental results in CS.

Definition 1 (Restricted Isometry Property): The observation matrix $\boldsymbol{X}$ is said to satisfy the restricted isometry property of order S with parameter $\delta_S \in (0, 1)$, if

$$(1 - \delta_S)\|\boldsymbol{z}\|_{\ell_2}^2 \leq \|\boldsymbol{X}\boldsymbol{z}\|_{\ell_2}^2 \leq (1 + \delta_S)\|\boldsymbol{z}\|_{\ell_2}^2 \tag{2}$$

holds for all S-sparse vectors $\boldsymbol{z} \in \mathbb{R}^n$.[1]

Remark 1: We will sometimes make use of shorthand notation, where instead of saying that a matrix $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S, we will say $\boldsymbol{X}$ satisfies $RIP(S, \delta_S)$.

While there is currently no known way to test in polynomial time whether a given matrix satisfies RIP, certain probabilistic constructions of matrices can be shown to satisfy RIP with high probability [1], [7]–[9]. The following result is representative [5].

Lemma 1: Let $\boldsymbol{X}$ be a $k \times n$ matrix whose entries are independent and identically distributed (i.i.d.), drawn from one of the following zero-mean distributions, each having variance $1/k$:

[1]A vector $\boldsymbol{z} \in \mathbb{R}^n$ is said to be S-sparse if $\|\boldsymbol{z}\|_{\ell_0} \leq S$, where $\|\boldsymbol{z}\|_{\ell_0}$ counts the number of nonzero entries in $\boldsymbol{z}$.

- $x_{i,j} \sim \mathcal{N}(0, 1/k)$,
- $x_{i,j} \sim \begin{cases} 1/\sqrt{k} & \text{with prob. } 1/2 \\ -1/\sqrt{k} & \text{w.p. } 1/2 \end{cases}$,
- $x_{i,j} \sim \begin{cases} \sqrt{3/k} & \text{w.p. } 1/6 \\ 0 & \text{w.p. } 2/3 \\ -\sqrt{3/k} & \text{w.p. } 1/6 \end{cases}$.

For any $\delta_S \in (0,1)$ and any $c_1 < \delta_S^2(3-\delta_S)/48$, set

$$c_2 = \frac{192 \log(12/\delta_S)}{3\delta_S^2 - \delta_S^3 - 48c_1}. \tag{3}$$

Then whenever $k \geq c_2 S \log n$, $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability at least $1 - \exp(-c_1 k)$.

The initial contributions to the theory of CS established, essentially, that any sparse signal can be recovered exactly from a collection of linear observations if the corresponding observation matrix satisfies RIP. The following result is a generalization that also describes the recovery of signals that are not exactly sparse.

Lemma 2 (Noiseless Recovery [3]): Let $\boldsymbol{X}$ be an observation matrix satisfying RIP of order $2S$ with parameter $\delta_{2S} < \sqrt{2} - 1$, and let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ be a vector of observations of any sparse unknown signal $\boldsymbol{\beta}$ having no more than S nonzero entries. The estimate $\widehat{\boldsymbol{\beta}}$ obtained as the solution of

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{z} \in \mathbb{R}^n}{\arg\min} \|\boldsymbol{z}\|_{\ell_1} \text{ subject to } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{z} \tag{4}$$

satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0 \frac{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}\|_{\ell_1}^2}{S}, \tag{5}$$

where $\boldsymbol{\beta}_S$ is the vector formed by setting all but the S largest entries (in magnitude) of $\boldsymbol{\beta}$ to zero, and

$$c_0 = 4\left(\frac{1 - \delta_{2S} + \sqrt{2}\delta_{2S}}{1 - \delta_{2S} - \sqrt{2}\delta_{2S}}\right)^2. \tag{6}$$

Note that in the case where the signal $\boldsymbol{\beta}$ has no more than S nonzero entries, this result guarantees that signal recovery is exact. Further, as shown in Lemma 1, there exist matrices for which this recovery is possible (with high probability) using only $k = O(S \log n)$ observations. Finally, the optimization program (4), which goes by the name of *basis pursuit*, is computationally tractable because it can be recast as a linear program. In this sense the optimality of noiseless CS is evident.

Suppose now that the observations are corrupted by some additive noise. That is, the observations are given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^k$ is either deterministic, or it is a vector whose entries are i.i.d. realizations of some zero-mean random variable. In either case, it turns out that CS can be used to obtain a stable recovery of the unknown signal.

The first work to establish theoretical results in the stochastic noise setting was [10], which used a reconstruction procedure that required a combinatorial search. The result presented here gives similar reconstruction error bounds, but is based on the RIP condition and utilizes a tractable convex optimization that goes by the name of *Dantzig*

selector. The original specification of the result in [2] assumed a specific signal class, but the proof actually provides a more general oracle result.

*Lemma 3 (**Dantzig Selector**—Recovery in Stochastic Noise [2]):* Let $\boldsymbol{X}$ be an observation matrix satisfying RIP of order $2S$ such that $\delta_{2S} < \sqrt{2}-1$ for some integer $S \geq 1$, and let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ be a vector of noisy observations of $\boldsymbol{\beta} \in \mathbb{R}^n$, where the entries of $\boldsymbol{\eta}$ are i.i.d. zero-mean Gaussian variables with variance σ^2. Choose $\lambda_n = \sqrt{2(1+a)(1+\delta_1)\log n}$ for any $a \geq 0$. The estimator

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\arg\min} \|\boldsymbol{z}\|_{\ell_1} \text{ subject to } \|\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{z})\|_{\ell_\infty} \leq \sigma\lambda_n \tag{7}$$

satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0' \min_{1\leq m\leq S} \left(\sigma\lambda_n\sqrt{m} + \frac{\|\boldsymbol{\beta}_m - \boldsymbol{\beta}\|_{\ell_1}}{\sqrt{m}}\right)^2 \tag{8}$$

with probability at least $1 - \left(\sqrt{\pi(1+a)\log n}\cdot n^a\right)^{-1}$. The constant $c_0' = 16/\left(1-\delta_{2S}-\sqrt{2}\delta_{2S}\right)^2$, and as in the previous lemma, $\boldsymbol{\beta}_m$ is the vector formed by setting all but the m largest entries (in magnitude) of $\boldsymbol{\beta}$ to zero.

Remark 2: The sufficient condition stated in the original result in [2] was $\delta_{2S} + \theta_{S,2S} < 1$, where $\theta_{S,2S}$ is called the $S, 2S$*-restricted orthogonality parameter* of the matrix $\boldsymbol{X}$. In general, the S, S'-restricted orthogonality parameter is such that

$$\left|(\boldsymbol{X}\boldsymbol{\alpha})'(\boldsymbol{X}\boldsymbol{\beta})\right| \leq \theta_{S,S'}\|\boldsymbol{\alpha}\|_{\ell_2}\|\boldsymbol{\beta}\|_{\ell_2} \tag{9}$$

holds for all vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ having no more than S and S' nonzero entries, respectively, and such that the nonzero entries of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ occur at disjoint subsets of indices. However, it can be shown that $\theta_{S,2S} \leq \sqrt{2}\delta_{2S}$, from which the RIP condition stated in the lemma follows.

Notice that the reconstruction error in (8) is essentially comprised of two factors. One factor is due to the "estimation error" that arises from determining m unknown quantities from noisy data, while the other is due to the "approximation error" or bias arising from estimating the unknown vector using only m components. For a given signal class, the best rate of error decay is obtained by balancing the two terms. That is, the best value of m is the value m^* such that

$$\frac{\|\boldsymbol{\beta}_{m^*} - \boldsymbol{\beta}\|_{\ell_1}}{m^*} = \sigma\lambda_n. \tag{10}$$

Thus, to make the optimal rates achievable, the observation matrix should be chosen to recover signals with sparsity S that is at least as large as the "effective sparsity" m^*.

Finally, when the perturbation (noise) vector is deterministic, a relaxation of the optimization program in (4) guarantees stable signal recovery. The result given below is from [3], which refines the initial result of [11].

Lemma 4 (Recovery in Deterministic Noise [3]): Let $\boldsymbol{X}$ be an observation matrix satisfying RIP of order $2S$ such that $\delta_{2S} < \sqrt{2}-1$ for some integer $S \geq 1$, and let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ be a vector of noisy observations of $\boldsymbol{\beta} \in \mathbb{R}^n$, where $\|\boldsymbol{\eta}\|_{\ell_2} \leq \epsilon$. The solution of

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{z}\in\mathbb{R}^n}{\arg\min} \|\boldsymbol{z}\|_{\ell_1} \text{ subject to } \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{z}\|_{\ell_2} \leq \epsilon \tag{11}$$

satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0'' \left(\epsilon + \frac{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}\|_{\ell_1}}{\sqrt{S}} \right)^2, \tag{12}$$

where $\boldsymbol{\beta}_S$ is as defined earlier and

$$c_0'' = \frac{16\,(1+\delta_{2S})}{\left(1-\delta_{2S}-\sqrt{2}\delta_{2S}\right)^2}. \tag{13}$$

Note that this result differs significantly from the stochastic noise result. Indeed, applying this result directly to the stochastic noise setting (in which case $\|\boldsymbol{\eta}\|_{\ell_2} \sim \sqrt{k}\sigma$) would yield an approximation error that grows like the *number of observations* times the noise power. On the other hand, the Dantzig selector results in a much better reconstruction error bound, with the estimation error scaling like the *sparsity level* times the noise power. In other words, the estimation error bound of the Dantzig selector is adaptive to the sparsity level, while the error bound of the relaxed ℓ_1 minimization in (11) is not. The difference in the two reconstruction error bounds could be significant, especially when the number of observations is far greater than the sparsity (or effective sparsity).

II. Random Toeplitz Matrices and Sparse Channel Estimation

We now describe a different problem domain which yields an observation model similar to the canonical CS setting. Consider point-to-point communication between two single-antenna transceivers over a wideband wireless multipath channel. Such single-antenna communication channels can be characterized as discrete, linear, time-invariant systems—see, e.g., [4] for further details. Optimal demodulation and decoding in wireless communication systems often requires accurate knowledge of the channel impulse response. Typically, this is accomplished by probing the channel with a known training sequence and linearly processing the channel output. Many real-world channels of practical interest, such as underwater acoustic channels [12], digital television channels [13] and residential ultrawideband channels [14], however, tend to have sparse or approximately sparse impulse responses. On the other hand, conventional linear channel estimation schemes, such as the least-squares method, fail to capitalize on the anticipated sparsity of the aforementioned channels. In contrast, it is established in this section that a channel estimate obtained as a solution to the Dantzig selector significantly outperforms a least-squares based channel estimate in terms of the mean squared error (MSE) when it comes to learning sparse (or approximately sparse) channels.

To begin with, let $\{x_i\}_{i=1}^p$, $p \in \mathbb{N}$, denote the training sequence, and consider using this sequence as the input to a wireless channel characterized by a finite (discrete) impulse response $\boldsymbol{\beta} \in \mathbb{R}^n$. The resulting observations $\boldsymbol{y} \in \mathbb{R}^{n+p-1}$ are described by the discrete-time convolution between the training signal $\boldsymbol{x}$ and the impulse response $\boldsymbol{\beta}$, and corruption by an additive noise vector $\boldsymbol{\eta}$: $\boldsymbol{y} = \boldsymbol{x} * \boldsymbol{\beta} + \boldsymbol{\eta}$. More specifically, if we use the notational convention that $x_i = 0$ for $i \notin \{1, 2, \ldots, p\}$, then each observation can be written as a sum,

$$y_j = \sum_{i=1}^{p} \beta_i \; x_{j+1-i} + \eta_j, \tag{14}$$

where the observations are corrupted by independent additive white Gaussian noise $\{\eta_j\}$ of variance σ^2 and contain a channel impulse response component only for $j = 1, \ldots, n+p-1$. The resulting input-output relation can be

expressed as a matrix-vector product

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n+p-2} \\ y_{n+p-1} \end{bmatrix} = \begin{bmatrix} x_1 & & 0 \\ x_2 & \ddots & \\ \vdots & \ddots & x_1 \\ x_p & & x_2 \\ & \ddots & \vdots \\ 0 & & x_p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix} \tag{15}$$

and the goal is to obtain an estimate of the channel impulse response $\boldsymbol{\beta}$ from knowledge of the observations $\boldsymbol{y}$ and training signal $\boldsymbol{x}$. Note that, by symmetry, either the training signal or the impulse response could be rewritten as a convolution matrix in the above formulation. Writing it in this way casts the channel estimation problem into the canonical CS framework.

The purpose of channel estimation in communication systems is to aid them in achieving their primary objective, which is reliable communication of data (information) from one point to another. Further, because of the dynamic nature of the wireless medium, the impulse response of a channel is bound to change over time [15]. As such, the input data sequence at the transmitter is periodically interspersed with the training sequence so as to maintain an up-to-date estimate of the channel impulse response at the receiver. We treat two facets of the sparse channel estimation problem in this section. The first one, corresponding to the lack of a "guard interval" of length $n-1$ between the data and training sequence, most closely resembles the canonical CS observation model, where the number of observations is far fewer than the length of the unknown signal. Specifically, consider a setting where the length of the training sequence $p = n + k - 1$ for some $k \geq 1$ and the training sequence is immediately preceded and succeeded by the data sequence. In this case, the first and last $n-1$ observations in (15) also contain contributions from the *unknown* data, rendering them useless for estimation purposes (the 0's in the convolution matrix in (15) would be replaced by the data sequence). Therefore, the channel estimation problem in this case reduces to reconstructing the unknown impulse response $\boldsymbol{\beta}$ from $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$, where the observation matrix $\boldsymbol{X}$ is a "partial" Toeplitz matrix of the form

$$\boldsymbol{X} = \begin{bmatrix} x_n & x_{n-1} & \dots & x_2 & x_1 \\ x_{n+1} & x_n & \dots & x_3 & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n+k-1} & x_{n+k-2} & \dots & x_{k+1} & x_k \end{bmatrix}. \tag{16}$$

The second setting we consider corresponds to the case when the training sequence is immediately preceded and succeeded by $n-1$ zeros (i.e., a guard interval of length $n-1$ exists between the data and the training sequence). In this setting, the channel estimation problem corresponds to obtaining an estimate of the channel impulse response from the "full" set of observations described by (15). Notice that when $p \geq n$, the partial Toeplitz matrix described in (16) above is a submatrix of the observation matrix in this setting. In contrast, when $p < n$, every row of the

observation matrix in this setting has at least one zero entry, and in the limiting case when $p = 1$, the observation matrix is just a scaled version of the $n \times n$ identity matrix.

The question we address in this section for both of the aforementioned settings is whether random binary probing, along with the use of a nonlinear Dantzig selector based estimator, can be employed to efficiently estimate a sparse channel, quantified by the condition $\|\boldsymbol{\beta}\|_{\ell_0} = S \ll n$. Note that initial theoretical analysis of CS systems that utilized random observation matrices relied inherently upon statistical independence among observations. The problem considered here is significantly more challenging—the Toeplitz structure of the (partial and full) observation matrices introduces statistical dependencies among observations and hence, existing techniques can no longer be employed. Instead, we develop a novel technique in Section III that facilitates analysis in the presence of such (structured) dependencies.

A. MSE of Least-Squares Channel Estimates

Estimation of an unknown vector $\boldsymbol{\beta}$ from linear observation models of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ is a well-studied problem in the area of estimation theory—see, e.g., [16]. Traditionally, channel estimates are usually obtained from $\boldsymbol{y}$ by solving the least-squares (LS) problem (or a variant of it). Note that in the case that the observation matrix $\boldsymbol{X}$ is given by (16), LS solution requires that $k \geq n$ so as to obtain a meaningful channel estimate [16]. Under this assumption, the LS channel estimate is given by

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{17}$$

where the observation matrix $\boldsymbol{X}$ corresponds to (16) in the case of training without guard intervals and to the full Toeplitz matrix in the case of training with guard intervals. Below, we lower bound the MSE performance of an LS channel estimate corresponding to random binary probing (binary phase shift keying signaling).

Theorem 1: Let the training sequence $\{x_i\}_{i=1}^{p}$ be given by a sequence of i.i.d. binary random variables taking values $+1$ or -1 with probability $1/2$ each. Further, let $p = n + k - 1$ for some $k \geq n$ for the case of training without guard intervals. Then the MSE of the LS channel estimate $\widehat{\boldsymbol{\beta}}_{LS}$ is lower bounded by

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq \frac{n\sigma^2}{p} \tag{18}$$

and

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq \frac{n\sigma^2}{k} \tag{19}$$

for training with and without guard intervals, respectively. Further, the equality in the above two expressions hold if and only if the corresponding observation matrices have orthogonal columns.

Proof: To establish this theorem, note that for both the cases of training with or without guard intervals

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] = \text{trace}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}\sigma^2. \tag{20}$$

Further, let $\{\lambda_i\}_{i=1}^n$ denote the n eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. Then, from elementary linear algebra, we have

$$\begin{aligned}\text{trace}\left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\} &= \sum_{i=1}^{n}\frac{1}{\lambda_i} = n\left(\frac{\sum_{i=1}^{n}\frac{1}{\lambda_i}}{n}\right)\\ &\overset{(a)}{\geq} n\left(\frac{n}{\sum_{i=1}^{n}\lambda_i}\right) = \frac{n^2}{\text{trace}\left\{\boldsymbol{X}'\boldsymbol{X}\right\}}\end{aligned} \tag{21}$$

where (a) follows from the arithmetic-harmonic means inequality. Also, from the arithmetic-harmonic means inequality, the equality in (a) holds if and only if $\lambda_1 = \lambda_2 = \cdots = \lambda_n$, resulting in the condition that $\boldsymbol{X}$ must have orthogonal columns for the equality to hold in (a). Finally, note that $\text{trace}\left\{\boldsymbol{X}'\boldsymbol{X}\right\} = np$ for the case of training with guard intervals, while $\text{trace}\left\{\boldsymbol{X}'\boldsymbol{X}\right\} = nk$ for the case of training without guard intervals and this completes the proof of the theorem. ■

B. MSE of Dantzig Selector Channel Estimates

We have seen in the previous section that the MSE of an LS channel estimate is lower bounded by (ambient dimension of $\boldsymbol{\beta}$) $\cdot\ \sigma^2/$(# of *effective* observations). Conventional channel learning techniques based on the LS criterion, however, fail to take into account the anticipated sparsity of the channel impulse response. To get an idea of the potential MSE gains that are possible when incorporating the sparsity assumption into the channel estimation strategy, we compare the performance of an LS based channel estimator to that of a channel estimation strategy that has been equipped with an *oracle*. The oracle does not reveal the true $\boldsymbol{\beta}$, but does inform us of the indices of nonzero entries of $\boldsymbol{\beta}$. Clearly this represents an ideal estimation strategy and one cannot expect to attain its performance level. Nevertheless, it is the target that one should consider.

To begin with, let $T_* \subset \{1,\ldots,n\}$ be the set of indices of the S nonzero entries of $\boldsymbol{\beta}$ and suppose that an oracle provides us with the sparsity pattern T_*. Then an ideal channel estimate $\boldsymbol{\beta}^*$ can be obtained for both the cases of training with or without guard intervals by first forming a *restricted* LS estimator from $\boldsymbol{y}$

$$\boldsymbol{\beta}_{T_*} = (\boldsymbol{X}'_{T_*}\boldsymbol{X}_{T_*})^{-1}\boldsymbol{X}'_{T_*}\boldsymbol{y} \tag{22}$$

where $\boldsymbol{X}_{T_*}$ is a submatrix obtained by extracting the S columns of $\boldsymbol{X}$ corresponding to the indices in T_*, and then setting $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}_{T_*}$ on the indices in T_* and zero on the indices in T_*^c. Appealing to the proof of Theorem 1, the MSE of this oracle channel estimator can be lower bounded as

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_{\ell_2}^2\right] &= \text{trace}\left\{(\boldsymbol{X}'_{T_*}\boldsymbol{X}_{T_*})^{-1}\right\}\sigma^2\\ &\geq \frac{S^2\sigma^2}{\text{trace}\left\{\boldsymbol{X}'_{T_*}\boldsymbol{X}_{T_*}\right\}}\end{aligned} \tag{23}$$

which results in the lower bound of $S\sigma^2/k$ for training without the guard intervals and $S\sigma^2/p$ for training with the guard intervals. In other words, the MSE of an oracle based channel estimate is lower bounded by (# of nonzero entries of $\boldsymbol{\beta}$) $\cdot\ \sigma^2/$(# of *effective* observations). Comparison of this lower bound with that for the MSE of an LS channel estimate shows that linear channel estimates based on the LS criterion may be at a significant disadvantage when it comes to estimating sparse channels. Finally, notice that in the case of training without guard

intervals (corresponding to the observation matrix given by (16)), the oracle estimator only requires that $k \geq S$ as opposed to $k \geq n$ for an LS channel estimate.

While the ideal channel estimate $\boldsymbol{\beta}^*$ is impossible to construct in practice, we now show that it is possible to obtain a more reliable estimate of $\boldsymbol{\beta}$ as a solution to the Dantzig selector (DS). This is accomplished by readily adapting the results from Lemma 3 in Section I-A and Theorems 5 and 6 in Section III-A. Below, we summarize these adapted results in terms of two theorems.

Theorem 2 (Training Without Guard Intervals): Let the training sequence $\{x_i\}_{i=1}^p$ be given by a sequence of i.i.d. binary random variables taking values $+1$ or -1 with probability $1/2$ each. Further, let $p = n + k - 1$ for some $k \geq 4c_2 S^2 \log n$. Choose $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the DS channel estimate

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{z} \in \mathbb{R}^n}{\arg\min} \|\boldsymbol{z}\|_{\ell_1} \text{ subject to } \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z})\|_{\ell_\infty} \leq \sigma\lambda_n\sqrt{k} \tag{24}$$

satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq 2c_0'(1+a)\log n \cdot \left(\frac{S\sigma^2}{k}\right) \tag{25}$$

with probability at least $1 - 2\max\left\{\left(\sqrt{\pi(1+a)\log n}\cdot n^a\right)^{-1}, \exp(-c_1 k/4S^2)\right\}$. Here, the observation matrix $\boldsymbol{X}$ corresponds to the partial Toeplitz matrix given in (16), c_0' is as defined in Lemma 3, and c_1 and c_2 are positive constants that depend only on S and are given in Theorem 5.

Theorem 3 (Training With Guard Intervals): Let the training sequence $\{x_i\}_{i=1}^p$ be given by a sequence of i.i.d. binary random variables taking values $+1$ or -1 with probability $1/2$ each. Further, let $p \geq 4c_2 S^2 \log n$ and choose $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the DS channel estimate

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{z} \in \mathbb{R}^n}{\arg\min} \|\boldsymbol{z}\|_{\ell_1} \text{ subject to } \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z})\|_{\ell_\infty} \leq \sigma\lambda_n\sqrt{p} \tag{26}$$

satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq 2c_0'(1+a)\log n \cdot \left(\frac{S\sigma^2}{p}\right) \tag{27}$$

with probability at least $1 - 2\max\left\{\left(\sqrt{\pi(1+a)\log n}\cdot n^a\right)^{-1}, \exp(-c_1 p/4S^2)\right\}$. Here, the observation matrix $\boldsymbol{X}$ corresponds to the full Toeplitz matrix given in (15), c_0' is as defined in Lemma 3, and c_1 and c_2 are positive constants that depend only on S and are given in Theorem 6.

The proofs of these theorems are essentially a direct application of Lemma 3 and Theorems 5 and 6, and are therefore omitted for the sake of brevity. These two theorems show that the DS channel estimate achieves squared error (roughly) within a factor of $\log n$ of the oracle based MSE lower bound of (# of nonzero entries of $\boldsymbol{\beta}$) $\cdot$ $\sigma^2/$(# of *effective* observations).

The appeal of the DS channel estimator, however, goes beyond the estimation of truly sparse channels. Indeed, it is to be expected that physical channels in certain scattering environments happen to be only "approximately" sparse [14]. Specifically, rearrange (and reindex) the entries of the channel impulse response $\boldsymbol{\beta}$ by decreasing order of magnitude: $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(n)}|$. We term a wireless channel *approximately sparse* if the ordered entries $\{\beta_{(j)}\}$ of its impulse response decay with the index j. The following theorem, which focuses on the case

of training with guard intervals and basically follows from Lemma 3 and Theorem 6, quantifies the reconstruction performance of the DS channel estimator in this setting.

Theorem 4 (Estimating Approximately Sparse Channels): Let the training sequence $\{x_i\}_{i=1}^p$ be given by a sequence of i.i.d. binary random variables taking values $+1$ or -1 with probability $1/2$ each. Fix any $S \in \mathbb{N}$, and choose $p \geq 4c_2 S^2 \log n$ and $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the DS channel estimate (26) satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0' \min_{1 \leq m \leq S} \left(\frac{\sigma \lambda_n \sqrt{m}}{\sqrt{p}} + \frac{\sum_{j=m+1}^{n} |\beta_{(j)}|}{\sqrt{m}} \right)^2 \tag{28}$$

with probability at least $1 - 2\max\left\{ \left(\sqrt{\pi(1+a)\log n} \cdot n^a \right)^{-1}, \exp(-c_1 p / 4S^2) \right\}$. Here, the observation matrix $\boldsymbol{X}$ corresponds to the full Toeplitz matrix given in (15), and the constants c_0', c_1 and c_2 are as in Theorem 3.

While the immediate significance of this result is obscured by the minimization over m in (28), its implications can be better understood by focusing on a specific decay structure of the ordered entries of $\boldsymbol{\beta}$. One such decay structure, which is widely-studied in the literature [17], assumes that the j-th largest entry of $\boldsymbol{\beta}$ obeys

$$|\beta_{(j)}| \leq R \cdot j^{-\alpha - 1/2} \tag{29}$$

for some $R > 0$ and $\alpha > 1/2$. The parameter α here controls the rate of decay of the magnitudes of the ordered entries. Under this decay condition, the summation in (28) can be explicitly written as $\sum_{j=m+1}^{n} |\beta_{(j)}| \leq C_\alpha R\, m^{-\alpha+1/2}$, where $C_\alpha > 0$ is a constant that depends only on α. We then have the following corollary of Theorem 4.

Corollary 1: Suppose that the channel impulse response $\boldsymbol{\beta} \in \mathbb{R}^n$ obeys (29) and let $\{x_i = \pm 1\}_{i=1}^p$ be the random binary sequence used to probe the channel for the case of training with guard intervals. Choose $p \geq C_2 \left(\log n\right)^{\frac{2\alpha-3}{2\alpha-1}} \left(\sigma^2\right)^{-\frac{2}{2\alpha-1}}$ and $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the reconstruction error of the DS channel estimate (26) is upper bounded by

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq C_0 \left(\log n\right)^{\frac{2\alpha}{2\alpha+1}} \cdot \left(\frac{\sigma^2}{p}\right)^{\frac{2\alpha}{2\alpha+1}} \tag{30}$$

with probability at least $1 - 2\max\left\{ \left(\sqrt{\pi(1+a)\log n} \cdot n^a \right)^{-1}, \exp\left(-C_1 \left(\log n \cdot \sigma^2\right)^{\frac{2}{2\alpha+1}} p^{\frac{2\alpha-1}{2\alpha+1}} \right) \right\}$. Here, the absolute constants $C_0(a, \alpha, c_0', R), C_1(a, \alpha, c_1, R)$, and $C_2(a, \alpha, c_2, R)$ are strictly positive and depend only on the parameters $a, \alpha, c_0', c_1, c_2$, and R.

It is instructive at this point to compare the reconstruction error performance of the DS channel estimate (given in (30)) with that of the LS channel estimate. Notice that since the MSE lower bound of $O(n\sigma^2/p)$ (given in (18)) holds for the LS channel estimate for all $\beta \in \mathbb{R}^n$, it remains valid under the decay condition (29). On the other hand, ignoring the $\log n$ factor in (30), we see that the reconstruction error of the DS solution essentially behaves like $O\left((\sigma^2/p)^{\frac{2\alpha}{2\alpha+1}}\right)$. Thus, even in the case of an approximately sparse channel impulse response, the DS channel estimate shows an MSE improvement by a factor of (roughly) $O(n \cdot (\sigma^2/p)^{1/(2\alpha+1)})$ over the LS MSE of $n\sigma^2/p$. In fact, it can also be shown that $O\left((\sigma^2/p)^{\frac{2\alpha}{2\alpha+1}}\right)$ is the minimax MSE rate for the class of channels exhibiting the decay (29) and hence, the performance of the DS channel estimator comes within a $\log n$ factor of a minimax

estimator. Finally, note that performance guarantees similar to the ones provided in Theorem 4 and Corollary 1 can also be obtained from Lemma 3 and Theorem 5 for the case of training without guard intervals.

III. Random Toeplitz Matrices Satisfy RIP

Because of the ubiquity of binary phase shift keying signaling in wireless communications, the channel estimation results in the previous section were stated in terms of random binary (± 1) probe sequences. However, the results also hold in settings where the probe sequence consists of realizations of random variables drawn from any bounded zero-mean distribution. In fact, the same results hold if the probe sequence is drawn from certain unbounded zero-mean distributions, such as the Gaussian distribution.

More generally, Toeplitz CS matrices have some additional benefits compared to completely independent (i.i.d.) random CS matrices. First, Toeplitz matrices are more efficient to generate and store. A $k \times n$ (random) partial Toeplitz matrix only requires the generation and storage of $k+n$ independent realizations of a random variable, while a fully-random matrix of the same size requires the generation and storage of kn random quantities. In addition, the use of Toeplitz matrices in CS applications leads to a general reduction in computational complexity. Performing a matrix-vector multiplication between a fully-random $k \times n$ matrix and an $n \times 1$ vector requires kn operations. In contrast, multiplication by a Toeplitz matrix can be performed in the frequency domain, because of the convolutional nature of Toeplitz matrices. Using fast Fourier transforms, the complexity of the multiplication can be reduced to $O(n \log n)$ operations, resulting in a significant speedup of the mixed-norm optimizations that are essential to several commonly-utilized CS reconstruction procedures such as GPSR [18] and SpaRSA [19]. Depending on the computational resources available, this speedup can literally be the difference between *intractable* and *solvable* problems.

In this section we establish the main claim of this paper, that Toeplitz matrices with entries drawn from either zero-mean bounded distributions or the zero-mean Gaussian distribution satisfy the restricted isometry property (RIP). Recall the definition of the Restricted Isometry Property from (1) of Section I-A. The RIP statement is essentially a statement about singular values, and to establish RIP for a given matrix it suffices to bound the extremal eigenvalues of the Gram matrices of all column submatrices (having no more than S columns) in the range $(1-\delta_S, 1+\delta_S)$. We will use this interpretation in our proofs, and the main results will be obtained using *Geršgorin's Disc Theorem*, which is an elegant result in classical eigenanalysis. We state this result here as a lemma, without proof. There are many valid references—see, for example, [20].

Lemma 5 (Geršgorin): The eigenvalues of an $m \times m$ matrix $\boldsymbol{M}$ all lie in the union of m discs $d_i = d_i(c_i, r_i)$, $i = 1, 2, \ldots, m$, centered at $c_i = M_{i,i}$, and with radius

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} |M_{i,j}|. \tag{31}$$

To begin, we consider any subset of column indices $T \subset \{1, \ldots, n\}$ of size $|T| \leq S$, and let $\boldsymbol{X}_T$ be the submatrix formed by retaining the columns of $\boldsymbol{X}$ indexed by the entries of T. The singular values of $\boldsymbol{X}_T$ are the eigenvalues of its $|T| \times |T|$ Gram matrix $\boldsymbol{G}(\boldsymbol{X}, T) = \boldsymbol{X}_T' \boldsymbol{X}_T$. Suppose that, for some integer $S \geq 1$ and some

positive values δ_d and δ_o chosen such that $\delta_d + \delta_o = \delta_S \in (0,1)$, every diagonal element of $\boldsymbol{G}(\boldsymbol{X},T)$ satisfies $|G_{i,i}(\boldsymbol{X},T) - 1| < \delta_d$ and every off-diagonal element $G_{i,j}(\boldsymbol{X},T)$, $i \neq j$, satisfies $|G_{i,j}(\boldsymbol{X},T)| < \delta_o/S$. Then the center of each Geršgorin disc associated with the matrix $\boldsymbol{G}(\boldsymbol{X},T)$ deviates from 1 by no more than δ_d and the radius of each disc is no larger than $(S-1)\delta_o/S < \delta_o$. By Lemma 5, the eigenvalues of $G(\boldsymbol{X},T)$ are all in the range $(1-\delta_d-\delta_o, 1+\delta_d+\delta_o) = (1-\delta_S, 1+\delta_S)$.

Now, notice that every Gram matrix $\boldsymbol{G}(\boldsymbol{X},T)$ is a submatrix of the full Gram matrix $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{X},\{1,\ldots,n\})$. Thus, instead of considering each submatrix separately, we can instead establish the above conditions on the elements of the full Gram matrix $\boldsymbol{G}$, and that suffices to ensure that the eigenvalues of *all* submatrices (formed by any choice of T, $|T| \leq S$) are controlled simultaneously. In the proofs that follow, we will show that every diagonal element of $\boldsymbol{G}$ is close to one (with high probability), and every off-diagonal element is bounded in magnitude (again, with high probability), and the final result will follow from a simple union bound.

It is instructive to note that because of the convolutional structure imposed by the linear, time-invariant observation model we consider here, the sufficient conditions to establish on the diagonal and off-diagonal elements of the Gram matrix of the resulting observation matrix essentially amount to properties of the autocorrelation function of the probe sequence. For the full observation matrix shown in (15), for example, each diagonal element is identical and equal to the autocorrelation of the probe sequence at lag zero. Similarly, each off-diagonal element corresponds to the autocorrelation at different nonzero lags (as stated in Section II, the probe sequence is assumed to be zero outside of the specified range). For the partial observation matrix of (16), the diagonal and off-diagonal elements correspond to windowed versions of the autocorrelation function at different lags. In the following subsections we quantify these autocorrelations for certain random input sequences. However, we note that the proof technique described above can be used to establish RIP for *any* input sequence (including possibly deterministic sequences); one would only need to verify that the autocorrelation function of the sequence satisfies the required conditions.

A. Bounded Entries

First we establish RIP for random Toeplitz matrices, for both the full observation matrices as shown in (15) as well as the partial matrices like (16), when the probe sequence $\{x_i\}$ consists of i.i.d. realizations of any bounded zero-mean random variable. We scale the distributions on x_i appropriately so that columns of the observation matrices are unit-normed in expectation. Suitable distributions are

- $x_i \sim \text{unif}\left[-\sqrt{3/\xi}, \sqrt{3/\xi} \right]$,
- $x_i \sim \begin{cases} 1/\sqrt{\xi} & \text{with prob. } 1/2 \\ -1/\sqrt{\xi} & \text{w.p. } 1/2 \end{cases}$,
- For $q \in (0,1)$, $x_i \sim \begin{cases} 1/\sqrt{\xi q} & \text{w.p. } q/2 \\ 0 & \text{w.p. } 1-q \\ -1/\sqrt{\xi q} & \text{w.p. } q/2 \end{cases}$,

where $\xi = k$ for partial matrices and $\xi = p$ for full matrices.

Before we state the first main results of the paper, we provide two lemmas that will be useful in the proofs. First, we describe the concentration of a sum of squares of bounded random variables.

Lemma 6: Let x_i, $i = 1, \ldots, k$ be a sequence of i.i.d., zero-mean bounded random variables such that $|x_i| \leq a$, and with variance $\mathbb{E}\left[x_i^2\right] = \sigma^2$. Then,

$$\Pr\left(\left|\sum_{i=1}^{k} x_i^2 - k\sigma^2\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{ka^4}\right) \tag{32}$$

Proof: Recall Hoeffding's inequality, which states that a sequence of k independent bounded random variables z_i satisfying $a_i \leq z_i \leq b_i$ with probability one, satisfies

$$\Pr\left(|s_k - \mathbb{E}\left[s_k\right]| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{k}(b_i - a_i)^2}\right), \tag{33}$$

where $s_k = \sum_{i=1}^{k} z_i$. In our case, we let $z_i = x_i^2$, so $z_i \in [0, a^2]$ with probability one, and since $s_k = \sum_{i=1}^{k} x_i^2$, $\mathbb{E}\left[s_k\right] = k\sigma^2$. The result follows. ■

Next, we describe how the inner product between vectors whose entries are bounded random variables concentrates about its mean.

Lemma 7: Let x_i and y_i, $i = 1, \ldots, k$ be sequences of i.i.d., zero-mean, bounded random variables satisfying $|x_i| \leq a$ (and thus $|x_i y_i| \leq a^2$). Then,

$$\Pr\left(\left|\sum_{i=1}^{k} x_i y_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2ka^4}\right). \tag{34}$$

Proof: Again we apply Hoeffding's inequality to the sum $s_k = \sum_{i=1}^{k} z_i$, this time with $z_i = x_i y_i$. In this case we have $-a^2 \leq z_i \leq a^2$ and since the elements are independent and have zero mean, $\mathbb{E}\left[s_k\right] = 0$. The result follows. ■

We are now in a position to state and prove the first main result of the paper.

Theorem 5: Let $\{x_i\}_{i=1}^{n+k-1}$ be a sequence whose entries are i.i.d. realizations of bounded zero-mean random variables with variance $\mathbb{E}\left[x_i^2\right] = 1/k$, satisfying $|x_i| \leq \sqrt{c/k}$ for some $c \geq 1$ (several such distributions are given above). Let

$$\boldsymbol{X} = \begin{bmatrix} x_n & x_{n-1} & \ldots & x_2 & x_1 \\ x_{n+1} & x_n & \ldots & x_3 & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n+k-1} & x_{n+k-2} & \ldots & x_{k+1} & x_k \end{bmatrix}, \tag{35}$$

be the $k \times n$ Toeplitz matrix generated by the sequence, and assume $n > 2$. Then, for any $\delta_S \in (0, 1)$, there exist constants c_1 and c_2 depending only on δ_S and c, such that whenever $k \geq c_2 S^2 \log n$, $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability exceeding $1 - \exp(-c_1 k/S^2)$.

Proof: Following the discussion of Geršgorin's Theorem, we need to establish conditions on the diagonal and off-diagonal elements of the Gram matrix $\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{X}$. Applying Lemma 6 we see that each diagonal element $G_{i,i} = \sum_{j=1}^{k} x_j^2$ satisfies

$$\Pr\left(|G_{i,i} - 1| \geq \delta_d\right) \leq 2\exp\left(-\frac{2k\delta_d^2}{c^2}\right) \tag{36}$$

and by the union bound

$$\Pr\left(\bigcup_{i=1}^{n}\{|G_{i,i}-1|\geq\delta_d\}\right)\leq 2n\exp\left(-\frac{2k\delta_d^2}{c^2}\right). \tag{37}$$

This establishes the required condition on the diagonal elements of the Gram matrix.

Next we treat the off-diagonal elements. Notice that entry $G_{i,j}$ is simply the inner product between columns i and j of the matrix $\boldsymbol{X}$. For example, one such term for the matrix specified in Theorem 5 is given by

$$G_{n-1,n}=x_1x_2+x_2x_3+x_3x_4+x_4x_5+\cdots+x_kx_{k+1}. \tag{38}$$

One issue is immediately apparent—the entries of the sum are not independent, so standard concentration inequalities cannot be applied directly. In the example here, the first two terms are dependent (they both depend on x_2), as are the second and third (both depend on x_3), and the third and fourth (both depend on x_4). But notice that the first and third terms are independent, as are the second and fourth, etc. Overall the sum may be split into two sums of i.i.d. random variables, where each component sum is formed simply by grouping alternating terms. The number of terms in each sum is either the same (if k is even) or differs by one if k is odd.

In fact this decomposition into two sums over independent entries is possible for every $G_{i,j}$, and this observation is the key to tolerating the dependencies that arise from the structure in the sensing matrix. Note that the terms in any such sum are each dependent with *at most two* other terms in the sum. Each sum can be rearranged such that the dependent terms are "chained"—that is, the ℓ-th (rearranged) term is dependent with (at most) the $(\ell-1)$-st term and the $(\ell+1)$-st terms. This rearranged sum has the same structure as the example above, and can be split in a similar fashion simply by grouping alternating terms.

Rewrite the sum $G_{i,j}=\sum_{i=1}^{k}z_i$, where the z_i's are identically distributed zero-mean random variables that satisfy $-c/k\leq z_i\leq c/k$. When k is even, the sum can be decomposed as

$$G_{i,j}=\sum_{i=1}^{t_1=k/2}z_{\pi_1(i)}+\sum_{i=1}^{t_2=k/2}z_{\pi_2(i)} \tag{39}$$

where t_1 and t_2 denote the number of terms in each sum, and $z_{\pi_1(i)}$ and $z_{\pi_2(i)}$ denote the rearranged and reindexed terms. The permutation operators π_1 and π_2 need not be known explicitly—it is enough to simply know such operators exist. When k is odd, we can write

$$G_{i,j}=\sum_{i=1}^{t_1=(k-1)/2}z_{\pi_1(i)}+\sum_{i=1}^{t_2=(k+1)/2}z_{\pi_2(i)}. \tag{40}$$

Generically, we write $G_{i,j}=G_{i,j}^1+G_{i,j}^2$. Applying Lemma 7 with $a^2=c/k$ to the component sums having t_1 and t_2 terms gives

$$\begin{aligned}
&\Pr\left(|G_{i,j}|\geq\frac{\delta_o}{S}\right)\\
&\quad\leq\ \Pr\left(\left\{|G_{i,j}^1|>\frac{\delta_o}{2S}\right\}\ \text{or}\ \left\{|G_{i,j}^2|>\frac{\delta_o}{2S}\right\}\right)\\
&\quad\leq\ 2\max\left\{\Pr\left(|G_{i,j}^1|>\frac{\delta_o}{2S}\right),\Pr\left(|G_{i,j}^2|>\frac{\delta_o}{2S}\right)\right\}\\
&\quad\leq\ 2\max\left\{2\exp\left(-\frac{k^2\delta_o^2}{8t_1c^2S^2}\right),2\exp\left(-\frac{k^2\delta_o^2}{8t_2c^2S^2}\right)\right\}.
\end{aligned} \tag{41}$$

It is easy to see that larger values of t_1 and t_2 decrease the error exponent, resulting in bounds that decay more slowly. For our purposes, to obtain a uniform bound independent of the parity of k, we use the (loose) upper bound $t_1 \leq t_2 < k$ to obtain

$$\Pr\left(|G_{i,j}| \geq \frac{\delta_o}{S}\right) \leq 4\exp\left(-\frac{k\delta_o^2}{8c^2S^2}\right). \tag{42}$$

To establish the condition for every off-diagonal element, we first note that, by symmetry, $G_{i,j} = G_{j,i}$. Thus, the total number of *unique* off-diagonal elements $G_{i,j}$ is $(n^2-n)/2 < n^2/2$, and we can apply the union of events bound to obtain

$$\Pr\left(\bigcup_{i=1}^{n}\bigcup_{\substack{j=1\\ j\neq i}}^{n}\left\{|G_{i,j}| \geq \frac{\delta_o}{S}\right\}\right) \leq 2n^2\exp\left(-\frac{k\delta_o^2}{8c^2S^2}\right). \tag{43}$$

This establishes the required condition on the off-diagonal elements of the Gram matrix.

Now, recall that RIP of order S holds with a prescribed $\delta_S \in (0,1)$ where $\delta_S = \delta_d + \delta_o$, when every diagonal element deviates from 1 by no more than δ_d, and every off-diagonal element is less than δ_o/S in magnitude. To obtain the result claimed in Theorem 5, we assume $n \geq 3$, let $\delta_d = \delta_o = \delta_S/2$ and use the union bound to obtain

$$\begin{aligned}\Pr\left(\boldsymbol{X} \text{ does not satisfy } RIP(S,\delta_S)\right) &\leq 2n^2\exp\left(-\frac{k\delta_o^2}{8c^2S^2}\right) + 2n\exp\left(-\frac{2k\delta_d^2}{c^2}\right) &(44)\\ &\leq 3n^2\exp\left(-\frac{k\delta_S^2}{32c^2S^2}\right). &(45)\end{aligned}$$

For $c_1 < \delta_S^2/32c^2$, the upper bound

$$\Pr\left(\boldsymbol{X} \text{ does not satisfy } RIP(S,\delta_S)\right) \leq \exp\left(-\frac{c_1 k}{S^2}\right) \tag{46}$$

holds whenever

$$k \geq \left(\frac{96c^2}{\delta_S^2 - 32c_1c^2}\right) S^2\log n, \tag{47}$$

which proves the theorem. ■

The same technique can be applied to the full observation matrices as in (15). This leads to the second main result of the paper.

Theorem 6: Let $\{x_i\}_{i=1}^{p}$ be a sequence whose entries are i.i.d. realizations of bounded zero-mean random variables with variance $\mathbb{E}\left[x_i^2\right] = 1/p$, satisfying $|x_i| \leq \sqrt{c/p}$ for some $c \geq 1$ (the example distributions listed at the start of the section again suffice). Let

$$\boldsymbol{X} = \begin{bmatrix} x_1 & & 0 \\ x_2 & \ddots & \\ \vdots & \ddots & x_1 \\ x_p & & x_2 \\ & \ddots & \vdots \\ 0 & & x_p \end{bmatrix} \tag{48}$$

be the $(n+p-1)\times n$ full Toeplitz matrix generated by the sequence, and assume $n > 2$. Then, for any $\delta_S \in (0,1)$ there exist constants c_1 and c_2 depending only on δ_S and c, such that for any sparsity level $S \leq c_2\sqrt{p/\log n}$, $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability exceeding $1-\exp(-c_1 p/S^2)$.

Remark 3: Notice the difference in the statements of results in Theorems 5 and 6, which highlight an inherent difference in the respective observation models. In the setting of Theorem 5, the user is allowed the flexibility to obtain more measurements "on the fly," and the resulting (rescaled) matrices satisfy RIP with higher orders S (or smaller parameters δ_S). Contrast that with the setting of Theorem 6, where the number of observations is fixed a priori. This effectively imposes an upper limit on the order S (or a lower limit on the parameter δ_S) for which RIP is satisfied.

Proof: The proof proceeds in a similar fashion to the proof of Theorem 5. Each column of the "full" observation matrix now contains p entries of the probe sequence, and is identical modulo an integer shift. From Lemma 6, the diagonal elements of the Gram matrix satisfy

$$\Pr\left(\bigcup_{i=1}^{n}\{|G_{i,i}-1|\geq\delta_d\}\right)\leq 2\exp\left(-\frac{2p\delta_d^2}{c^2}\right). \tag{49}$$

The off-diagonal elements are still composed of sums of dependent random variables, however, in this case the number of nonzero terms comprising each sum varies. At most (when i and j differ by 1), $G_{i,j}$ will consist of a sum of $p-1$ terms. On the other extreme, if $p \leq |j-i|$, each term of the inner product is zero trivially. In any event, we can still apply the results of Lemma 7 and upper-bound the error for each term by the worst-case behavior. This gives

$$\begin{aligned}
&\Pr\left(|G_{i,j}|\geq\frac{\delta_o}{S}\right)\\
&\quad\leq\quad \Pr\left(\left\{|G_{i,j}^1|>\frac{\delta_o}{2S}\right\}\text{ or }\left\{|G_{i,j}^2|>\frac{\delta_o}{2S}\right\}\right)\\
&\quad\leq\quad 2\max\left\{\Pr\left(|G_{i,j}^1|>\frac{\delta_o}{2S}\right),\Pr\left(|G_{i,j}^2|>\frac{\delta_o}{2S}\right)\right\}\\
&\quad\leq\quad 2\max\left\{2\exp\left(-\frac{p^2\delta_o^2}{8t_1c^2S^2}\right),2\exp\left(-\frac{p^2\delta_o^2}{8t_2c^2S^2}\right)\right\}.
\end{aligned} \tag{50}$$

Notice that now, regardless of the parity of p, the number of terms in each partial sum (t_1 and t_2) is no greater than $p/2$. The bound

$$\Pr\left(\bigcup_{i=1}^{n}\bigcup_{\substack{j=1\\ j\neq i}}^{n}\left\{|G_{i,j}|\geq\frac{\delta_o}{S}\right\}\right)\leq 2n^2\exp\left(-\frac{p\delta_o^2}{4c^2S^2}\right). \tag{51}$$

follows. As before, we let $\delta_d=\delta_o=\delta_S/2$ and assume $n\geq 3$, to obtain

$$\Pr\left(\boldsymbol{X}\text{ does not satisfy }RIP(S,\delta_S)\right)\leq 3n^2\exp\left(-\frac{p\delta_S^2}{16c^2S^2}\right). \tag{52}$$

For any $c_1 < \delta_S^2/16c^2$ and

$$S\leq\sqrt{\frac{\delta_S^2-16c_1c^2}{48c^2}}\cdot\sqrt{\frac{p}{\log n}}, \tag{53}$$

the matrix $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability at least $1 - \exp(-c_1 p/S^2)$, proving the theorem. ■

B. Gaussian Entries

Results analogous to those of Theorems 5 and 6 can also be obtained if the entries of the probe sequence are drawn independently from certain unbounded distributions. For example, probe sequences consisting of i.i.d. Gaussian entries also generate Toeplitz matrices that satisfy RIP.

Following the proof techniques above, we first need to establish that the sum of squares of i.i.d. Gaussian random variables concentrates about its mean. For that, we utilize the following result from [21, Sec. 4, Lem. 1].

Lemma 8: Let $\{x_i\}_{i=1}^k$ be i.i.d. Gaussian variables with mean 0 and variance σ^2. The sum of squares of the x_i's satisfies

$$\Pr\left(\sum_{i=1}^{k} x_i^2 - k\sigma^2 \geq 2\sigma^2\sqrt{kt} + 2\sigma^2 t\right) \leq \exp(-t) \tag{54}$$

and

$$\Pr\left(\sum_{i=1}^{k} x_i^2 - k\sigma^2 \leq -2\sigma^2\sqrt{kt}\right) \leq \exp(-t). \tag{55}$$

For $0 \leq t \leq 1$, the symmetric bound

$$\Pr\left(\left|\sum_{i=1}^{k} x_i^2 - k\sigma^2\right| \geq 4\sigma^2\sqrt{kt}\right) \leq 2\exp(-t) \tag{56}$$

follows.

In addition, we can quantify the concentration of inner products between zero-mean Gaussian random vectors as follows.

Lemma 9: Let x_i and y_i, $i = 1, \ldots, k$ be sequences of i.i.d., zero-mean Gaussian random variables with variance σ^2. Then,

$$\Pr\left(\left|\sum_{i=1}^{k} x_i y_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{4\sigma^2(k\sigma^2 + t/2)}\right). \tag{57}$$

Proof: The proof basically follows the derivation of Bernstein's Inequality. Using the Chernoff bound, we obtain

$$\Pr\left(\sum_{i=1}^{k} x_i y_i \geq t\right) \leq \exp(-st) \prod_{i=1}^{k} \mathbb{E}\left[\exp\left(s x_i y_i\right)\right], \tag{58}$$

which holds for all $s \geq 0$ and all $t > 0$. Fix a term inside the product and expand the exponential in a Taylor Series, which gives

$$\begin{aligned} \mathbb{E}\left[\exp\left(s x_i y_i\right)\right] &= \mathbb{E}\left[1 + (s x_i y_i) + \frac{(s x_i y_i)^2}{2!} + \frac{(s x_i y_i)^3}{3!} + \frac{(s x_i y_i)^4}{4!} + \ldots\right] &(59)\\ &\leq \mathbb{E}\left[1 + \frac{(s x_i y_i)^2}{2!} + \frac{|s x_i y_i|^3}{3!} + \frac{(s x_i y_i)^4}{4!} + \frac{|s x_i y_i|^5}{5!} + \ldots\right] &(60)\end{aligned}$$

Now, since the x_i's and y_i's are Gaussian and independent, it is easy to verify that $\mathbb{E}\left[|x_i y_i|^p\right] = \mathbb{E}\left[|x_i|^p\right] \cdot \mathbb{E}\left[|y_i|^p\right] \leq p!\ \sigma^{2p}$ for $p \geq 2$, and so the expectation can be bounded by

$$\begin{aligned} \mathbb{E}\left[\exp\left(s x_i y_i\right)\right] &\leq 1 + s^2\sigma^4 + s^3\sigma^6 + s^4\sigma^8 + \dots & (61)\\ &= 1 + s^2\sigma^4 \sum_{j=0}^{\infty} (s\sigma^2)^j. & (62)\end{aligned}$$

Now, assume $s\sigma^2 = \nu < 1$ to obtain

$$\mathbb{E}\left[\exp\left(s x_i y_i\right)\right] \leq 1 + \frac{s^2\sigma^4}{1-\nu} \leq \exp\left(\frac{s^2\sigma^4}{1-\nu}\right). \tag{63}$$

Combining results, we have

$$\Pr\left(\sum_{i=1}^{k} x_i y_i \geq t\right) \leq \exp\left(-st + \frac{k s^2\sigma^4}{1-\nu}\right), \tag{64}$$

or equivalently,

$$\Pr\left(\sum_{i=1}^{k} x_i y_i \geq \frac{\gamma}{s} + \frac{k s\sigma^4}{1-\nu}\right) \leq \exp\left(-\gamma\right). \tag{65}$$

Now substitute $s = \nu/\sigma^2$, let $\alpha = k\nu\sigma^2/(1-\nu)$ and $\beta = \gamma\sigma^2/\nu$, and simplify to obtain

$$\Pr\left(Z \geq \alpha + \beta\right) \leq \exp\left(-\frac{\alpha\beta}{\sigma^2(k\sigma^2 + \alpha)}\right). \tag{66}$$

Letting $\alpha = \beta = t/2$, for $t < 2$, we obtain

$$\Pr\left(Z \geq t\right) \leq \exp\left(-\frac{t^2}{4\sigma^2(k\sigma^2 + t/2)}\right). \tag{67}$$

The other half of the bound can be obtained similarly using the fact that

$$\Pr\left(Z \leq -t\right) \leq \Pr\left(-sZ \geq st\right) \leq \exp\left(-st\right)\mathbb{E}\left[\exp\left(-sZ\right)\right], \tag{68}$$

and

$$\mathbb{E}\left[\exp\left(-sZ\right)\right] \leq \mathbb{E}\left[1 + \frac{(s x_i y_i)^2}{2!} + \frac{|s x_i y_i|^3}{3!} + \frac{(s x_i y_i)^4}{4!} + \frac{|s x_i y_i|^5}{5!} + \dots\right] \tag{69}$$

as above, making the bounds symmetric and identical. The result follows. ■

Leveraging the above lemmas, we can establish the following.

Theorem 7: Let $\{x_i\}_{i=1}^{n+k-1}$ be a sequence whose entries are i.i.d. Gaussian random variables with mean zero and variance $\mathbb{E}\left[x_i^2\right] = 1/k$. Let

$$\boldsymbol{X} = \begin{bmatrix} x_n & x_{n-1} & \dots & x_2 & x_1 \\ x_{n+1} & x_n & \dots & x_3 & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n+k-1} & x_{n+k-2} & \dots & x_{k+1} & x_k \end{bmatrix}, \tag{70}$$

be the $k \times n$ Toeplitz matrix generated by the sequence, and assume $n > 2$. Then, for any $\delta_S \in (0,1)$ there exist constants c_1 an c_2 depending only on δ_S, such that whenever $k \geq c_2 S^2 \log n$, $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability exceeding $1 - \exp(-c_1 k/S^2)$.

Proof: Following the proof method used in the previous subsection, we first use the symmetric bound in Lemma 8 to establish that

$$\Pr\left(\bigcup_{i=1}^{n}\{|G_{i,i}-1|\geq\delta_d\}\right)\leq 2n\exp\left(-\frac{k\delta_d^2}{16}\right). \tag{71}$$

The off-diagonal elements exhibit the same dependencies treated in the proofs of Theorems 5 and 6. Again splitting each sum into two sums over independent entries, we leverage Lemma 9 to obtain

$$\Pr\left(|G_{i,j}|\geq\frac{\delta_o}{S}\right)\leq 2\max\left\{2\exp\left(-\frac{k\delta_o^2}{4S^2(t_1/k+1/2)}\right),2\exp\left(-\frac{k\delta_o^2}{4S^2(t_2/k+1/2)}\right)\right\} \tag{72}$$

for any $0\leq\delta_d\leq 1$, where again t_1 and t_2 are the number of terms in each sum. Using the conservative upper bound $t_1\leq t_2\leq k$ we obtain

$$\Pr\left(\bigcup_{i=1}^{n}\bigcup_{\substack{j=1\\j\neq i}}^{n}\left\{|G_{i,j}|\geq\frac{\delta_o}{S}\right\}\right)\leq 2n^2\exp\left(-\frac{k\delta_o^2}{6S^2}\right). \tag{73}$$

Now, let $\delta_d=2\delta_S/3$ and $\delta_o=\delta_S/3$ and assume $n\geq 3$, to obtain

$$\Pr\left(\boldsymbol{X}\text{ does not satisfy }RIP(S,\delta_S)\right)\leq 3n^2\exp\left(-\frac{k\delta_S^2}{54S^2}\right). \tag{74}$$

For any $c_1<\delta_S^2/54$ and

$$k\geq\left(\frac{162}{\delta_S^2-54c_1}\right)S^2\log n, \tag{75}$$

the matrix $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability at least $1-\exp(-c_1k/S^2)$, proving the theorem. ■

For the full observation matrix, composed of entries from a Gaussian sequence, the following is true.

Theorem 8: Let $\{x_i\}_{i=1}^{p}$ be a sequence whose entries are i.i.d. realizations of zero-mean Gaussian random variables with variance $1/p$. Let

$$\boldsymbol{X}=\begin{bmatrix} x_1 & & 0\\ x_2 & \ddots & \\ \vdots & \ddots & x_1\\ x_p & & x_2\\ & \ddots & \vdots\\ 0 & & x_p \end{bmatrix} \tag{76}$$

be the $(n+p-1)\times n$ full Toeplitz matrix generated by the sequence, and assume $n>2$. Then, for any $\delta_S\in(0,1)$ there exist constants c_1 and c_2 depending only on δ_S, such that for any sparsity level $S\leq c_2\sqrt{p/\log n}$ $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability exceeding $1-\exp(-c_1p/S^2)$.

Proof: The proof is analogous to the proof of Theorem 6. The columns of $\boldsymbol{X}$ are identical (modulo an integer shift), so

$$\Pr\left(\bigcup_{i=1}^{n}\{|G_{i,i}-1|\geq\delta_d\}\right)\leq 2\exp\left(-\frac{p\delta_d^2}{16}\right). \tag{77}$$

and now,

$$\Pr\left(\bigcup_{i=1}^{n}\bigcup_{\substack{j=1\\j\neq i}}^{n}\left\{|G_{i,j}|\geq\frac{\delta_o}{S}\right\}\right)\leq 2n^2\exp\left(-\frac{p\delta_o^2}{4S^2}\right). \tag{78}$$

Letting $\delta_d = 2\delta_S/3$ and $\delta_o = \delta_S/3$ and assuming $n \geq 3$, we have that for any $c_1 < \delta_S^2/36$ and

$$S \leq \sqrt{\frac{\delta_S^2 - 36c_1}{108}} \cdot \sqrt{\frac{p}{\log n}}, \tag{79}$$

the matrix $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability at least $1 - \exp(-c_1 p/S^2)$, proving the theorem. ■

IV. Discussion

A. Generalizations and Dependency Tolerance using Graph Coloring

It is easy to see that the results of Theorems 5-8 also apply directly to Hankel matrices, which are Toeplitz-like matrices whose entries are identical along anti-diagonals. In addition, the proof techniques utilized to obtain the results of Theorems 5 and 7 also can be used to establish RIP for (left- or right-shifted) partial circulant matrices of the form

$$\boldsymbol{X} = \begin{bmatrix} x_n & x_{n-1} & \dots & \dots & \dots & x_3 & x_2 & x_1 \\ x_1 & x_n & \dots & \dots & \dots & x_4 & x_3 & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{k-1} & x_{k-2} & \dots & x_1 & x_n & \dots & x_{k+1} & x_k \end{bmatrix}, \tag{80}$$

generated by a random sequence of length n.

The techniques developed here can also be applied in more general settings where the observation matrices exhibit structured statistical dependencies. Recall that, in the above proofs, dependencies were tolerated by partitioning sums of dependent random variables into two component sums of fully independent random variables. The actual partitioning was not performed directly, rather the only facts required in the proof were that such partitions exist, and that the number of terms in each component sum was specified. If, for a given observation matrix, similar partitioning can be established, analogous results will follow.

We generalize the approach utilized in this paper using techniques from graph theory. See, for example, [22] for basic reference. Let

$$\Sigma = \sum_{i=1}^{k} x_i \tag{81}$$

be a sum of identically distributed random variables. We associate the sum Σ with an undirected graph $g(\Sigma) = (V, E)$ of degree Δ_g, by associating a vertex $i \in V = \{1, 2, \dots, k\}$ to each term x_i in the sum and creating an edge set E such that an edge $e = (i, j)$ between vertices is in the set if and only if x_i and x_j are statistically dependent. The degree of the graph Δ_g is defined to be the maximum number of edges originating from any of the vertices. Notice that any fully-disconnected subgraph of $g(\Sigma)$, by definition, represents a collection of i.i.d. random variables.

The goal, then, is to partition $g(\Sigma)$ into some number of fully-disconnected subgraphs. In graph theory terminology, any such partitioning—essentially a labeling of each vertex such that vertices sharing an edge are labeled differently—is called a (proper) coloring of the graph. Given a coloring of $g(\Sigma)$, the concentration behavior of each partial sum associated with each subgraph can be obtained in a straightforward manner by standard concentration inequalities, and the contribution of several such subgraphs can be quantified using the union bound. Note, however, that trivial partitions exist (let each subgraph contain only one vertex, for example), leading to particularly poor concentration bounds. We seek to partition $g(\Sigma)$ into as few fully-disconnected subgraphs as possible while ensuring that each subgraph contains as many vertices as possible.

To achieve this, we consider *equitable coloring* of $g(\Sigma)$. An equitable coloring is a proper graph coloring where the difference in size between the smallest and largest collections of vertices sharing the same color is at most one. Proving a conjecture of Paul Erdös, Hajnal and Szemerédi showed that equitable colorings of a graph with degree Δ exist for any number of colors greater or equal to $(\Delta+1)$ [23]. Along with the above argument, this shows that the concentration behavior of any sum Σ exhibiting limited statistical dependence, as defined by the degree Δ_g of the associated dependency graph $g(\Sigma)$, can be controlled using equitable graph coloring. This procedure was also used to extend Hoeffding's inequality to such graph-dependent random variables in [24].

Utilizing this framework, we can obtain results that apply to observation matrices with more general dependency structures. The following result is representative.

Theorem 9: Let $\boldsymbol{X}$ be a $k \times n$ matrix whose entries are identically distributed realizations of bounded zero-mean random variables with variance $\mathbb{E}\left[x_i^2\right] = 1/k$, satisfying $x_i^2 \leq c/k$ for some $c \geq 1$. Assume that the dependency degree among elements in any column of $\boldsymbol{X}$ is no greater than some integer $\Delta_d \geq 0$, and each inner product between columns exhibits dependency degree no greater than some integer $\Delta_o \geq 0$. Then, for any $\delta_S \in (0,1)$, there exist constants c_1 and c_2 depending on δ_S, the dependency degrees Δ_d and Δ_o, and c, such that whenever $k \geq c_2 S^2 \log n$, $\boldsymbol{X}$ satisfies RIP of order S with parameter δ_S with probability exceeding $1 - \exp(-c_1 k/S^2)$.

Proof: First consider the diagonal elements of the Gram matrix of $\boldsymbol{X}$, each of which satisfies

$$\Pr\left(|G_{i,i} - 1| \geq \delta_d\right) \leq 2(\Delta_d + 1)\exp\left(-\frac{2k\delta_d^2}{c^2}\left\lfloor\frac{k}{\Delta_d+1}\right\rfloor\right) \tag{82}$$

and by the union bound

$$\Pr\left(\bigcup_{i=1}^{n}\left\{|G_{i,i} - 1| \geq \delta_d\right\}\right) \leq 2n(\Delta_d + 1)\exp\left(-\frac{2\delta_d^2}{c^2}\left\lfloor\frac{k}{\Delta_d+1}\right\rfloor\right), \tag{83}$$

where $\lfloor\cdot\rfloor$ is the floor function, which returns the largest integer less than the argument. Similarly, the off-diagonal elements satisfy

$$\Pr\left(\bigcup_{i=1}^{n}\bigcup_{\substack{j=1\\ j\neq i}}^{n}\left\{|G_{i,j}| \geq \frac{\delta_o}{S}\right\}\right) \leq 2n^2(\Delta_o + 1)\exp\left(-\frac{\delta_o^2}{8c^2S^2}\left\lfloor\frac{k}{\Delta_o+1}\right\rfloor\right). \tag{84}$$

The result follows from suitable bounding of the overall error probability. ■

B. Connections with Other Works

To the best of our knowledge, this paper is the first work to establish the restricted isometry property for random Toeplitz matrices with bounded or Gaussian entries. Here we briefly describe connections between this paper and several related existing works.

In the compressed sensing literature, the first work to propose Toeplitz-structured observation matrices was [25], where observations were obtained by convolving the incoming unknown signal with a random filter—a filter whose taps were generated as realizations of certain random variables—followed by periodic downsampling of the output stream. While this approach was shown to be effective in practice, no theoretical guarantees were given. Without downsampling this initial approach is identical to the full observation model analyzed here, and in fact the techniques presented here could be utilized to establish conditions under which RIP would be satisfied for certain downsampled random filtering systems.

The first theoretical results for using random Toeplitz matrices in compresed sensing were established in [6]. Using an equitable graph coloring approach applied to the RIP proof of [5], we showed that $k \times n$ partial Toeplitz, Hankel, and left- or right-shifted circulant random matrices satisfy RIP of order S with high probability, provided $k = O\left(S^3 \log n\right)$. This sufficient condition is more restrictive than what we establish here, where we reduce the exponent on S by one order of magnitude.

In [26], techniques in sparse representation are applied to recover matrices that can be expressed as the superposition of a small number of component matrices—the so-called sparse matrix identification problem. When the component matrices are time-frequency shift matrices, the matrix identification problem becomes similar to the convolutional sparse signal recovery problem considered here. However, this work differs from our own in several significant ways. First, the work in [26] considers noise-free observations, and requires circular convolution with certain deterministic (Alltop) probe sequences. Second, and perhaps most notably, the theoretical analysis in [26] focuses on the coherence property of the dictionary (maximum absolute inner product between any two distinct columns), instead of RIP. Consequently, while we can establish that, for a given probe sequence, *any* sparse signal can be recovered with high probability from a collection of convolutional observations, the results in [26] only guarantee recovery of *most* signals defined on a randomly chosen signal support.

The problem of matrix identification was also studied in [27]. While this work did consider noisy observations, the recovery procedure proposed was the "bounded-noise" optimization (11). As explained in Section I-A the error bounds resulting from this optimization can be sufficiently weaker than the bounds that can be obtained using the Dantzig selector, given in (7). In fact, in addition to being restricted to circularly convolutional observations, and theoretical guarantees only on coherence (instead of RIP), the work in [27] provides no theoretical analysis to quantify the MSE of reconstructions obtained from noisy observations.

Our own previous work [4] was the first to use Geršgorin's Theorem to establish RIP for Toeplitz random matrices, achieving the less restrictive sufficient condition on the number of observations required, $k = O\left(S^2 \log n\right)$. While that work only treated matrices whose entries were drawn from a symmetric Bernoulli distribution, here we extend

the results to random matrices whose entries are bounded or Gaussian-distributed.

In addition to the directly-related contributions described above, we briefly mention several tangentially-related contributions. While this paper was in preparation we became aware of the work of [28], which also examines random convolution in compressed sensing. There are several significant differences between this work and our own. First, in [28], the probe sequence (while randomly generated) is assumed to have orthogonal shifts, while we can tolerate probe sequences that simply consist of independent realizations of certain random variables. Second, our results establish RIP for the observation matrices resulting from convolution with a random sequence, from which it follows that *any* sparse (or nearly sparse) signal can be recovered with high probability from the same set of observations. In contrast, for a given random probe, the results in [28] only guarantee recoverability of *most* sparse signals defined on a fixed support, albeit with a potentially less restrictive condition on the number of observations required, $k = O\left(\max\left\{S \log n, \log^3 n\right\}\right)$ compared to our requirement of $k = O(S^2 \log n)$. Perhaps the most significant difference, however, is that our observation model prescribes collecting a consecutive set of samples of the linear convolution between the unknown signal and a random probe sequence, while the observation model in [28] requires circular convolution of the unknown sparse signal with a random probe sequence, followed by either random subsampling or randomized "block averaging" of the full set of observations. Thus, while our observation model occurs naturally in the context of linear system identification, the methods proposed in [28] do not.

Finally, in [29] it was established that, subject to a similar condition as what we obtain here—namely that the number of rows of the matrix must be on the order of the square of the sparsity level of the target signal, certain deterministic matrices satisfy RIP. Among these was a special type of block-circulant matrix generated by a collection of $\ell > 1$ columns, where the elements of the matrix satisfy $X_{i+1,j+\ell} = X_{i,j}$, and the arithmetic on the indices is done modulo the signal length n. In contrast, the generalization of our Toeplitz results apply to "true" circulant matrices that are generated by a single (random) column.

C. Eigenvalues by Geršgorin's Theorem

The theory of sparse representation was an active area of research even before the advent of compressed sensing. The techniques that were developed in early works relied on the notion of coherence of a matrix, which is quantified by the largest (in magnitude) inner product between distinct columns of the matrix. The interesting point to note is that the notion of coherence can be parlayed into statements about RIP, the connection coming by way of Geršgorin's Theorem. Reminiscent constructs can be found, for example, in [30]. In addition, Geršgorin-like techniques arise in the proof of RIP for the deterministic constructions of [29], and are mentioned in [31] in the context of determining the eigenvalues of randomly chosen submatrices of a given dictionary matrix.

Using Geršgorin's Theorem to establish eigenvalues for general dictionaries is not without its limitations. For example, as noted in [31], the work of [32] shows that the minimum coherence between columns of any (generally overcomplete) finite Grassmanian frame cannot be too small. For large k and n, the coherence scales like $\sqrt{1/k}$, which would essentially imply a $k = O(S^2)$ requirement on the number of observations, similar to what we obtain

in our proofs. Applying Geršgorin's theorem to fully-independent random matrices leads to similar restrictions. For example, a simple application of Lemma 7 (analogous to the approach in the proof of Theorem 5, but without the dependency tolerance steps) shows that Geršgorin's Theorem leads to the requirement that $O(S^2 \log n)$ rows are needed in order for a fully random observation matrix to satisfy RIP of order S with some fixed success probability. On the other hand, we know from [1], [5], [7]–[9] that $k = O(S \log n)$ requirements suffice to establish RIP with the same probability of success.

Thus, while it is tempting to claim that the presence of dependencies in the Toeplitz-structured matrices amounts to an increase in the number of observations required for RIP to be satisfied, such a claim does not follow from the work presented here. Indeed, it is fully possible that the random matrices considered in this work do satisfy RIP when $k = O(S \log n)$, but the proof techniques utilized here are insufficient to establish that stronger result. The takeaway message here is that Geršgorin's Theorem provides a straightforward, but possibly suboptimal, approach to establishing RIP for general observation matrices.

V. Acknowledgments

The authors wish to thank Phil Schniter for pointing out a few minor errors in the initial version of the dependency tolerance arguments.

References

[1] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[2] ——, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[3] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," in *C. R. Acad. Sci., Ser. I*, Paris, 2008, vol. 346, pp. 589–592.

[4] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," in *Proc.* 42*nd Annu. Conf. Information Sciences and Systems (CISS '08)*, Princeton, NJ, Mar. 2008, pp. 5–10.

[5] R. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," in *Constructive Approximation*. New York: Springer, 2008.

[6] W. U. Bajwa, J. Haupt, G. Raz, S. J. Wright, and R. Nowak, "Toeplitz-structured compressed sensing matrices," in *Proc.* 14*th IEEE/SP Workshop on Statistical Signal Processing (SSP '07)*, Madison, WI, Aug. 2007, pp. 294–298.

[7] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[8] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[9] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[10] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, Sep. 2006.

[11] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Mar. 2006.

[12] D. B. Kilfoyle and A. B. Baggeroer, "The state of the art in underwater acoustic telemetry," *IEEE J. Oceanic Eng.*, vol. 25, no. 1, pp. 4–27, Jan. 2000.

[13] *Receiver Performance Guidelines*, ATSC Recommended Practices for Digital Television, 2004. [Online]. Available: http://www.atsc.org/standards/practices.html

[14] A. F. Molisch, "Ultrawideband propagation channels-Theory, measurement, and modeling," *IEEE Trans. Veh. Technol.*, vol. 54, no. 5, pp. 1528–1545, Sep. 2005.

[15] J. G. Proakis, *Digital Communications*, 4th ed. New York, NY: McGraw-Hill, 2001.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall, 1993.

[17] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2435–2476, Oct. 1998.

[18] M. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Select. Topics Signal Processing*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[19] S. J. Wright, R. D. Nowak, and M. T. Figueiredo, "Sparse reconstruction by separable approximation," in *Proc. IEEE Intl. Conf. Acoust., Speech and Signal Processing (ICASSP '08)*, Las Vegas, NV, Apr. 2008, pp. 3373–3376.

[20] R. S. Varga, *Geršgorin and His Circles*, ser. Springer Series in Computational Mathematics. Berlin, Germany: Springer-Verlag, 2004, no. 36.

[21] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, Oct. 2000.

[22] D. B. West, *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall, 2000.

[23] A. Hajnal and E. Szemerédi, "Proof of a conjecture of P. Erdős," in *Combinatorial Theory and its Application*, P. Erdős, A. Rényi, and V. T. Sós, Eds., North-Holland, Amsterdam, 1970, pp. 601–623.

[24] S. Pemmaraju, "Equitable coloring extends Chernoff-Hoeffding bounds," in *Proc. RANDOM-APPROX 2001*, Berkeley, CA, Aug. 2001, pp. 285–296.

[25] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. IEEE Intl. Conf. Acoust., Speech and Signal Processing (ICASSP '06)*, Toulouse, France, May 2006, pp. 872–875.

[26] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, 2007, in press.

[27] G. E. Pfander, H. Rauhut, and J. Tanner, "Identification of matrices having a sparse representation," *IEEE Trans. Signal Processing*, 2007, in press.

[28] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imaging Science*, July 2008, submitted.

[29] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, pp. 918–925, August 2007.

[30] D. L. Donoho and M. Elad, "Optimally sparse represetations in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, pp. 2197–2202, March 2003.

[31] J. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmonic Anal.*, vol. 25, no. 1, pp. 1–24, July 2008.

[32] T. Strohmer and R. Heath, "Grassmanian frames with applications to coding and communication," *Appl. Comput. Harmonic Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.