Toeplitz Compressed Sensing Matrices with Applications to Sparse Channel Estimation

Jarvis Haupt, *Member, IEEE,* Waheed U. Bajwa, *Member, IEEE,* Gil Raz, and Robert Nowak *Fellow, IEEE*

***Abstract*—Compressed sensing (CS) has recently emerged as a powerful signal acquisition paradigm. In essence, CS enables the recovery of high-dimensional sparse signals from relatively few linear observations in the form of projections onto a collection of test vectors. Existing results show that if the entries of the test vectors are independent realizations of certain zero-mean random variables, then with high probability the unknown signals can be recovered by solving a tractable convex optimization. This work extends CS theory to settings where the entries of the test vectors exhibit structured statistical dependencies. It follows that CS can be effectively utilized in linear, time-invariant system identification problems provided the impulse response of the system is (approximately or exactly) sparse. An immediate application is in wireless multipath channel estimation. It is shown here that time-domain probing of a multipath channel with a random binary sequence, along with utilization of CS reconstruction techniques, can provide significant improvements in estimation accuracy compared to traditional least-squares based linear channel estimation strategies. Abstract extensions of the main results are also discussed, where the theory of equitable graph coloring is employed to establish the utility of CS in settings where the test vectors exhibit more general statistical dependencies.**

***Index Terms*—circulant matrices, compressed sensing, Hankel matrices, restricted isometry property, sparse channel estimation, Toeplitz matrices, wireless communications.**

I. Introduction

A. Background

Consider the problem of recovering an unknown signal $\boldsymbol{\beta} \in \mathbb{R}^n$ from a collection of linear observations, $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} \in \mathbb{R}^k$. This very general observation model encompasses a wide variety of applications, including magnetic resonance imaging, digital imaging, and radio frequency surveillance. When the number of observations, k, equals or exceeds the dimension of the unknown signal, n, (the so-called overdetermined setting) 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.

The emerging theory of compressed sensing (CS) is primarily concerned with the regime where the number of observations is *less* than the dimension of the unknown signal, the so-called underdetermined setting. The seminal works in CS established that signals can still be recovered exactly from such incomplete observations using tractable recovery procedures such as convex optimization, provided the signals are sparse [1]–[3]. One concise way to specify for which matrices $\boldsymbol{X}$ this recovery is possible is using the restricted isometry property (RIP), which was first introduced in [4].

Definition 1 (Restricted Isometry Property)**.** *The observation matrix $\boldsymbol{X}$ is said to satisfy the restricted isometry property of order $S \in \mathbb{N}$ with parameter $\delta_S \in (0,1)$—or, in shorthand, $\boldsymbol{X}$ satisfies RIP(S,δ_S)—if*

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

holds for all $\boldsymbol{z} \in \mathbb{R}^n$ having no more than S nonzero entries.

In other words, $\boldsymbol{X}$ satisfies RIP(S,δ_S) if the singular values of all submatrices of $\boldsymbol{X}$ formed by retaining no more than S columns of $\boldsymbol{X}$ are in the range $(\sqrt{1-\delta_S}, \sqrt{1+\delta_S})$, and thus $\boldsymbol{X}$ acts almost like an isometry for sparse vectors having no more than S nonzero entries.

A variety of results are available in the CS literature for recovery procedures whose successes are contingent on observation matrices that satisfy the RIP [4]–[7]. Here, our primary interest will be in the recovery of sparse (or approximately sparse) signals in additive noise. That is, we are interested in the case where the observations are given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^k$ is a vector whose entries are realizations of some independent and identically distributed (i.i.d.) zero-mean random variables. The first work to establish theoretical results for CS in the stochastic noise setting was [8], 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 the *Dantzig selector*. The original specification of the result in [9] assumed a specific signal class, but the proof actually provides a more general oracle result which we state here.

Lemma 1 (**The Dantzig Selector** [9])**.** *Let $\boldsymbol{X}$ be an observation matrix satisfying RIP($2S$,δ_{2S}) with $\delta_{2S} < \sqrt{2}-1$ for some $S \in \mathbb{N}$, and let $\|\boldsymbol{X}\|_{\ell_1,\ell_2}$ denote the largest ℓ_2 norm of the columns of $\boldsymbol{X}$. Let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ be a vector of*

This research was supported in part by the National Science Foundation under grants CCF-0353079 and ECS-0529381, and 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.

J.H. is with the Department of Electrical and Computer Engineering at Rice University in Houston, TX. W.U.B. is with the Program in Applied and Computational Mathematics at Princeton University in Princeton, NJ. G.R. is with GMR Research and Technology in Concord, MA. R.N. is with the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison. E-mails: jdhaupt@rice.edu, wbajwa@math.princeton.edu, raz@gmrtech.com, nowak@engr.wisc.edu.

Submitted: August 29, 2008. Revised: March 17, 2010. Current version: June 19, 2010.

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 $\lambda_n = \|\boldsymbol{X}\|_{\ell_1,\ell_2} \cdot \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the estimator

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

satisfies

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

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

In particular, this result states that when $\boldsymbol{\beta}$ has fewer than S nonzero entries, the approximation error $\|\boldsymbol{\beta}_m - \boldsymbol{\beta}\|_{\ell_1} = 0$, and so ignoring log terms, we have that the overall mean-square estimation error of the Dantzig selector estimate is proportional to the number of nonzero entries S times the noise power σ^2.

In general, the RIP-based theory of CS provides a powerful toolbox that enables one to establish a variety of recovery claims for a given problem provided that (a) the problem can be cast into the canonical CS framework and (b) the resulting equivalent observation matrix can be shown to satisfy the RIP (with the problem-dependent parameter conditions). In particular, the utility of CS in *abstract* undersampled sparse recovery problems relies—to a large extent—on the fact that there exist many classes of $k \times n$ matrices that satisfy the RIP, despite the number of rows k being much smaller than the ambient dimension n. In this regard, some of the best known results in the existing literature correspond to the class of random observation matrices, which establish that certain probabilistic constructions of matrices satisfy the RIP of order $S \approx O(k)$ with high probability [1]–[4], [10]. For example, let $\boldsymbol{X}$ be a $k \times n$ matrix whose entries are i.i.d., taking the values $\pm 1/\sqrt{k}$ each with probability $1/2$. Then, for a specified $\delta_S \in (0,1)$, it is known that $\boldsymbol{X}$ satisfies RIP(S, δ_S) with probability at least $1 - \exp(-c_1 k)$ provided $k > c_2 S \log n$, where $c_1, c_2 > 0$ are functions of δ_S [10].

B. Our Contributions

The major theoretical contribution of the work presented in this paper is an extension of the CS theory to observation matrices that exhibit structured statistical dependencies across its rows and columns, such as random Toeplitz matrices, which arise naturally from the convolutional structure inherent to linear system identification problems. For example, we show in Section III of the paper that any $k \times n$ random Toeplitz matrix

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

whose entries $\{x_i\}_{i=1}^{n+k-1}$ are i.i.d. $\pm 1/\sqrt{k}$ each with probability $1/2$, satisfies RIP(S,δ_S) with probability at least $1 - \exp(-c_1' k/S^2)$, provided $k \geq c_2' S^2 \log n$, where $c_1', c_2' > 0$ are functions of δ_S [cf. Theorem 4].

In addition to the novelty of the results reported in the paper, our proof technique also differs markedly from the techniques used in the earlier literature to establish the RIP for random matrices, such as eigenanalysis of random matrices [2], [4] and concentration of measure inequalities [10]. Both of these approaches require the entries of the matrices to be fully-independent, and thus are not easily extendable to the settings considered here. Instead, our approach is to establish the RIP using bounds on the *coherence* of $\boldsymbol{X}$, which is defined to be the magnitude of the largest inner product between columns of a matrix and has been frequently utilized in related sparse recovery problems; see, e.g., [11]–[13]. Specifically, we parlay the coherence results into a statement about the RIP by using a result from classical eigenanalysis, known as Geršgorin's Disc Theorem. The key to accomplishing this is the development of concentration inequalities to establish bounds on the magnitudes of the inner products between columns of random matrices with structured statistical dependencies, which are of the form of sums of dependent random variables, and for this we employ a novel partitioning of the sum somewhat similar to that utilized in [14]. In particular, this enables us to build and improve upon our own previous works, which were the first to provide theoretical performance guarantees for CS using random Toeplitz-structured matrices [15], [16].

Besides establishing the RIP for random Toeplitz matrices, we also consider the application of our results in the problem of identification of discrete, linear, time-invariant (LTI) systems. In particular, we establish that random probing of LTI systems having sparse impulse responses, coupled with the utilization of CS reconstruction methods, results in significant improvements in estimation accuracy when compared with traditional least-squares based linear estimation strategies. 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 [16]–[19]. 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 potentially lead to significant reductions in transmission energy and improvements in spectral efficiency. Indeed, this observation has prompted a number of researchers in the recent past to propose various sparse-channel estimation schemes [17]–[19], some of which have in fact been inspired by the earlier literature on sparse signal approximation [17]. However, a major limitation of the previous investigations is the lack of a quantitative theoretical analysis of the performance of the proposed methods in terms of the reconstruction error. In contrast, we show in this paper that the reconstruction error of the Dantzig selector channel estimator comes within a logarithmic factor of the lower bound on the error of an (ideal, but unrealizable) oracle-based channel estimator.

C. Organization

The remainder of this paper is organized as follows. In Section II we describe the wireless multipath channel estimation problem. By casting the problem into the CS framework, we leverage the main results of the paper (that random Toeplitz matrices satisfy the RIP) to 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 theoretical 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.

II. Random Toeplitz Matrices and Sparse Channel Estimation

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., [16] 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 [20], digital television channels [21] and residential ultrawideband channels [22], however, tend to have sparse or approximately sparse impulse responses, and conventional linear channel estimation schemes such as the least-squares method fail to capitalize on the anticipated sparsity. 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 $\boldsymbol{\beta}$, with corruption by an additive noise vector $\boldsymbol{\eta}$; that is, $\boldsymbol{y} = \boldsymbol{x} * \boldsymbol{\beta} + \boldsymbol{\eta}$. 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,

$$y_j = \sum_{i=1}^{p} \beta_i \; x_{j+1-i} + \eta_j, \quad j = 1, \ldots, n+p-1,$$

and in what follows we assume the noise terms $\eta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. The resulting input-output relation can be expressed as a matrix-vector product

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n+p-2} \\ y_{n+p-1} \end{bmatrix} = \begin{bmatrix} x_1 & & 0 \\ x_2 & \ddots & \\ \vdots & \ddots & x_1 \\ x_p & & x_2 \\ & \ddots & \vdots \\ 0 & & x_p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \end{bmatrix}, \tag{2}$$

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 either the training signal or the impulse response could be rewritten as a convolution matrix in the above formulation, but this formulation casts the channel estimation problem into the canonical CS framework.

The purpose of channel estimation in communication systems is to assist in the reliable communication of data (information) from one point to another. Because of the dynamic nature of the wireless medium, the impulse response of a channel is bound to change over time [23]. 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. In this regard, we treat two facets of the sparse channel estimation problem. The first one 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 (2). The second case corresponds to the lack of a "guard interval" of length $n-1$ between the data and training sequence, and 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 (2) also contain contributions from the *unknown* data, rendering them useless for estimation purposes (the 0's in the convolution matrix in (2) 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 (1). Notice that when $p \geq n$, the partial Toeplitz matrix described in (1) 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, in Section III we develop a technique 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., [24]. 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 (1), LS solution requires that $k \geq n$ so as to obtain a meaningful channel estimate [24]. Under this assumption, the LS channel estimate is given by

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

where the observation matrix $\boldsymbol{X}$ corresponds to (1) 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*

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq \frac{n\sigma^2}{p}, \tag{3}$$

and

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

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

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

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

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

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. ∎

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 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 that one cannot expect to implement in practice. Nevertheless, its ideal performance is a useful target to consider for comparison.

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 $\boldsymbol{y}$

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

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

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

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 $\boldsymbol{\beta}$) $\cdot$ σ^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 (1)), the oracle estimator only requires that $k \geq S$ as opposed to $k \geq n$ for an LS channel estimate.

B. MSE of Dantzig Selector Channel Estimates

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. The appeal of the Dantzig selector 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 [22]. 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, and we denote by $\boldsymbol{\beta}_m$ the vector formed by retaining the m largest entries of $\boldsymbol{\beta}$ and setting the rest of the entries to zero. The following theorems describe the estimation performance that can be achieved using the Dantzig selector. The proofs are essentially a direct application of Lemma 1 and Theorems 4 and 5, and are therefore omitted for the sake of brevity. For the case of training without guard intervals, we have the following.

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 $k \geq 4c_2 S^2 \log n$. The Dantzig selector estimate $\widehat{\boldsymbol{\beta}}$, obtained by applying Lemma 1 with $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$, satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0' \min_{1\leq m\leq S} \left(\sigma \sqrt{\frac{2m(1+a)\log n}{k}} + \frac{\|\boldsymbol{\beta}_m - \boldsymbol{\beta}\|_1}{\sqrt{m}} \right)^2,$$

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 (1)*, c_0' is as defined in Lemma 1, and c_1 and c_2 are positive constants that depend only on S and are given in Theorem 4.*

Similarly, we obtain the following result for the case where guard intervals are used.

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 $p \geq 4c_2 S^2 \log n$. The Dantzig selector estimate $\widehat{\boldsymbol{\beta}}$, obtained by applying Lemma 1 with $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$, satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq c_0' \min_{1\leq m\leq S} \left(\sigma \sqrt{\frac{2m(1+a)\log n}{p}} + \frac{\|\boldsymbol{\beta}_m - \boldsymbol{\beta}\|_1}{\sqrt{m}} \right)^2,$$

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 (2)*, c_0' is as defined in Lemma 1, and c_1 and c_2 are positive constants that depend only on S and are given in Theorem 5.*

Remark 1. *To obtain the results above one must account for the scaling issues that arise because the entries of $\boldsymbol{X}$ have unit variance. For example, the MSE bound in Theorem 2 is obtained using $\sigma_{\mathit{eff}} = \sigma/\sqrt{k}$ in Lemma 1, and thus the condition in the Dantzig selector optimization becomes $\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z})\|_{\ell_\infty} \leq \sigma_{\mathit{eff}}\lambda_n k = \sigma\lambda_n\sqrt{k}$.*

While the immediate significance of these results may be obscured by the minimization over m, the implications can be better understood by considering two key regimes. First, in the case where the channel is S-sparse, the channel estimates satisfy $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq 2c_0'(1+a)\log n\left(S\sigma^2/k\right)$ with high probability for training without guard intervals, and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq 2c_0'(1+a)\log n\left(S\sigma^2/p\right)$ with high probability for training with guard intervals. In other words, the Dantzig selector 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).

Second, we consider a case with specific decay structure of the ordered entries of $\boldsymbol{\beta}$. One such decay structure, which is widely-studied in the literature [25], assumes that the j-th largest entry of $\boldsymbol{\beta}$ obeys

$$|\beta_{(j)}| \leq R \cdot j^{-\alpha - 1/2}\ , \tag{4}$$

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, we have $\|\boldsymbol{\beta}_m - \boldsymbol{\beta}\|_{\ell_1} \leq C_\alpha R\, m^{-\alpha+1/2}$, where $C_\alpha > 0$ is a constant that depends only on α. In this case, we have the following corollary of Theorem 3 (similar results can be obtained from Theorem 2).

Corollary 1. *Suppose that the channel impulse response $\boldsymbol{\beta} \in \mathbb{R}^n$ obeys* (4) *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 $p \geq C_2 (\log n)^{\frac{2\alpha-3}{2\alpha-1}} (\sigma^2)^{-\frac{2}{2\alpha-1}}$ and $\lambda_n = \sqrt{2(1+a)\log n}$ for any $a \geq 0$. Then the reconstruction error of the Dantzig selector channel estimate is upper bounded by*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq C_0 (\log n)^{\frac{2\alpha}{2\alpha+1}} \cdot \left(\frac{\sigma^2}{p}\right)^{\frac{2\alpha}{2\alpha+1}}, \tag{5}$$

with probability at least

$$1 - 2\max\left\{ \left(\sqrt{\pi(1+a)\log n}\cdot n^a\right)^{-1}, \exp\left(-C_1 (\log n \cdot \sigma^2)^{\frac{2}{2\alpha+1}} p^{\frac{2\alpha-1}{2\alpha+1}}\right) \right\}.$$

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 (5)) with that of the LS channel estimate. Notice that since the MSE lower bound of $n\sigma^2/p$ (given in (3)) holds for the LS channel estimate for all $\beta \in \mathbb{R}^n$, it remains valid under the decay condition (4). On the other hand, ignoring the $\log n$ factor in (5), 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) $n \cdot (\sigma^2/p)^{1/(2\alpha+1)}$ 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 (4) and hence, the performance of the DS channel estimator comes within a $\log n$ factor of a minimax estimator.

III. Random Toeplitz Matrices Satisfy the 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, similar results also hold in settings where the probe sequence consists of realizations of random variables drawn from any bounded zero-mean distribution, as well as 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 [26] and SpaRSA [27]. Depending on the computational resources available, this speedup can literally be the difference between *intractable* and *solvable* problems.

In this section we identify conditions under which Toeplitz matrices with entries drawn from either zero-mean bounded distributions or the zero-mean Gaussian distribution satisfy the restricted isometry property (RIP), as a function of the parameters S and δ_S. Recall the Restricted Isometry Property from Definition 1. 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, [28].

Lemma 2 (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*

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

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 2, 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 (2), 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 (1), 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 (2) as well as the partial matrices like (1), 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

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

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 3. *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,*

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

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

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

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 4. *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,*

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

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 4. *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*

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

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(S,δ_S) with probability exceeding $1 - \exp(-c_1 k/S^2)$. Specifically, for any $c_1 \leq \delta_S^2/32c^2$, it suffices to choose

$$c_2 \geq \left(\frac{96c^2}{\delta_S^2 - 32c_1c^2} \right).$$

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 3 we see that each diagonal element $G_{i,i} = \sum_{j=1}^{k} x_j^2$ satisfies

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

and by the union bound

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

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 4 is given by

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

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

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

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, note that t_1 and t_2 will differ by one, but each will be no greater than $(k+1)/2$.

Generically, we write $G_{i,j} = G^1_{i,j} + G^2_{i,j}$. Applying Lemma 4 with $a^2 = c/k$ to the component sums having t_1 and t_2 terms gives

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

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

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

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

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

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

Now, recall that the 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 4, we assume $n \geq 3$, let $\delta_d = \delta_o = \delta_S/2$ and use the union bound to obtain

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

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

$$\Pr\left(\boldsymbol{X} \text{ does not satisfy RIP}(S,\delta_S)\right) \leq \exp\left(-\frac{c_1k}{S^2}\right),$$

holds whenever

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

which proves the theorem. ∎

The same technique can be applied to the full observation matrices as in (2). This leads to the second main result of the paper.

Theorem 5. *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*

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

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(S,δ_S) with probability exceeding $1-\exp(-c_1p/S^2)$. Specifically, for any $c_1 < \delta_S^2/16c^2$, it suffices to choose

$$c_2 \leq \sqrt{\frac{\delta_S^2 - 16c_1c^2}{48c^2}}.$$

Remark 2. *Notice the difference in the statements of results in Theorems 4 and 5, which highlight an inherent difference in the respective observation models. In the setting of Theorem 4, the user is allowed the flexibility to obtain more measurements "on the fly," and the resulting (rescaled) matrices satisfy the RIP with higher orders S (or smaller parameters δ_S). Contrast that with the setting of Theorem 5, 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 the RIP is satisfied.*

Proof: The proof proceeds in a similar fashion to the proof of Theorem 4. Each column of the "full" observation matrix now contains p entries of the probe sequence, and is identical modulo an integer shift. From Lemma 3, the diagonal elements of the Gram matrix satisfy

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

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 4 and upper-bound the error for each term

by the worst-case behavior. This gives

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

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

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

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

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

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

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

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

B. Gaussian Entries

Similar results to those of Theorems 4 and 5 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 the 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 [29, Sec. 4, Lem. 1].

Lemma 5. *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*

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

and

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

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

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

follows.

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

Lemma 6. *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,*

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\exp(sx_iy_i)\right] \\
&= \mathbb{E}\left[1 + (sx_iy_i) + \frac{(sx_iy_i)^2}{2!} + \frac{(sx_iy_i)^3}{3!} + \ldots\right] \\
&\leq \mathbb{E}\left[1 + \frac{|sx_iy_i|^2}{2!} + \frac{|sx_iy_i|^3}{3!} + \frac{|sx_iy_i|^4}{4!} + \ldots\right].
\end{aligned}$$

Now, since the x_i's and y_i's are Gaussian and independent, we have

$$\begin{aligned}
\mathbb{E}\left[|x_iy_i|^p\right] &= \mathbb{E}\left[|x_i|^p\right]\cdot\mathbb{E}\left[|y_i|^p\right] \\
&= \left(\mathbb{E}\left[|x_i|^p\right]\right)^2.
\end{aligned}$$

Further, the absolute moments of x_i are given generally by $\mathbb{E}[|x_i|^p] = 2^{p/2}\Gamma\left((p-1)/2\right)\sigma^p/\sqrt{\pi}$ for integers $p \geq 1$. Simplifying the expression, we have

$$\mathbb{E}[|x_i|^p] = \begin{cases} 1\cdot 3\cdot 5\cdots(p-1)\cdot\sigma^p, & p \text{ even} \\ \sqrt{2/\pi}\cdot 2^{(p-1)/2}\cdot\left(\frac{p-1}{2}\right)!\cdot\sigma^p, & p \text{ odd} \end{cases}.$$

When $p \geq 2$ is even we have

$$\begin{aligned}
\left(\mathbb{E}[|x_i|^p]\right)^2 &\leq 1\cdot 1\cdot 3\cdot 3\cdots(p-1)\cdot(p-1)\cdot\sigma^{2p} \\
&\leq p!\ \sigma^{2p},
\end{aligned}$$

by inspection. When p is odd, say $p = 2\omega+1$ for some $\omega \geq 1$, we have

$$\begin{aligned}
\sqrt{\frac{2}{\pi}}\cdot 2^{(p-1)/2}\cdot\left(\frac{p-1}{2}\right)! &= \sqrt{\frac{2}{\pi}}\cdot 2^{\omega}\cdot\omega! \\
&\leq 2\cdot 4\cdots 2\omega,
\end{aligned}$$

and thus,

$$\begin{aligned}
\left(\mathbb{E}[|x_i|^p]\right)^2 &\leq 2\cdot 2\cdot 4\cdot 4\cdots 2\omega\cdot 2\omega\cdot\sigma^{2p} \\
&\leq 2\cdot 3\cdot 4\cdot 5\cdots 2\omega\cdot(2\omega+1)\cdot\sigma^{2p} \\
&\leq p!\ \sigma^{2p}.
\end{aligned}$$

In either case, $\left(\mathbb{E}[|x_i|^p]\right)^2 \leq p!\sigma^{2p}$, and so the expectation can be bounded by

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

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

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

Combining results, we have

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

or equivalently,

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

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

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

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

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

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

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

and

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-sZ\right)\right]\\ &\quad\leq\quad \mathbb{E}\left[1 + \frac{|sx_iy_i|^2}{2!} + \frac{|sx_iy_i|^3}{3!} + \frac{|sx_iy_i|^4}{4!} + \dots\right],\end{aligned}$$

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

Leveraging the above lemmas, we can establish the following.

Theorem 6. *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*

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

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 $k \geq c_2S^2\log n$, $\boldsymbol{X}$ satisfies RIP(S,δ_S) with probability exceeding $1 - \exp(-c_1k/S^2)$.

Proof: Following the proof method used in the previous subsection, we first use the symmetric bound in Lemma 5 to establish that

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

The off-diagonal elements exhibit the same dependencies treated in the proofs of Theorems 4 and 5. Again splitting each sum into two sums over independent entries, we leverage Lemma 6 to obtain

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

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

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

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

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

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

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

the matrix $\boldsymbol{X}$ satisfies RIP(S,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 7. *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*

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

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 5. The columns of $\boldsymbol{X}$ are identical (modulo an integer shift), so

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

and now,

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

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

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

the matrix $\boldsymbol{X}$ satisfies RIP(S,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 4-7 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 4 and 6 also can be used to establish the RIP for (left- or right-shifted) partial circulant matrices of the form

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

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 using an approach similar to that used in [14], 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, [30] for basic reference. Let $\Sigma = \sum_{i=1}^{k} x_i$ 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 contained in the edge 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 $(\Delta+1)$ [31]. 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 [32].

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

Theorem 8. *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(S,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

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

and by the union bound

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

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

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

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 works.

In the compressed sensing literature, the first work to propose Toeplitz-structured observation matrices was [33], 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 compressed sensing were established in [15]. Using an equitable graph coloring approach applied to the RIP proof of [10], we showed that $k \times n$ partial Toeplitz, Hankel, and left- or right-shifted circulant random matrices satisfy the RIP of order S with high probability, provided $k = \Omega\left(S^3 \log n\right)$. This sufficient condition is more restrictive than what we establish here, where we reduce the exponent on S by one order of magnitude.

Our own previous work [16] 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, $k = \Omega\left(S^2 \log n\right)$. 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 contrast to the *linear* convolution observation model utilized here, several recent works have also examined sparse recovery using measurements arising from *circular* convolution. In particular, the works [34], [35] considered the problem of recovering matrices that can be expressed as the superposition of a small number of component matrices from their action on a given probe signal—the so-called sparse matrix identification problem. Both of those works considered the use of deterministic (Alltop) probe sequences, while [34] also examined the use of random probes. This approach is somewhat reminiscent of the system identification problem considered here in cases where the component matrices are time-frequency shift matrices. However, the noisy recovery procedures proposed in each of those works were based on coherence measures and utilized a weaker "bounded-noise" optimization [5], [36], for which the error bounds can be sufficiently weaker than the bounds that can be obtained using the Dantzig selector and the RIP as here. A related effort, [37], established guarantees for the recovery of sparse signals using observations that arise from circular convolution with a specially-constructed probe sequence (designed to have exactly orthogonal circular shifts) followed by random downsampling or randomized "block averaging" of the output.

Finally, in [38] 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 the 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 applies to "true" circulant matrices that are generated by shifts of a single (random) row vector.

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 [39]. In addition, Geršgorin-like techniques arise in the proof of RIP for the deterministic constructions of [38], and are mentioned in [40] 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 [40], the work of [41] 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 $k = \Omega(S^2)$ 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 4 (analogous to the approach in the proof of Theorem 4, but without the dependency tolerance steps) shows that Geršgorin's Theorem leads to the requirement that $\Omega(S^2 \log n)$ rows are needed in order for a fully random observation matrix to satisfy the RIP of order S with some fixed success probability. On the other hand, we know from [1]–[4], [10] that $k = O(S \log n)$ measurements 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 the 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, and the anonymous reviewers for their time and for many helpful suggestions for improvement.

References

[1] 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.

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

[3] 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.

[4] ——, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[5] 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.

[6] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, vol. 26, no. 3, pp. 301–321, May 2008.

[7] D. Needell and R. Vershynin, "Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit," *IEEE J. Sel. Topics in Sig. Proc.*, vol. 4, no. 2, pp. 310–316, Apr. 2010.

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

[9] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[10] R. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008.

[11] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[12] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Appl. Comp. Harmonic Anal.*, vol. 22, no. 3, pp. 335–355, May 2007.

[13] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Comm. Pure Appl. Math*, vol. 59, no. 6, pp. 797–829, Jun. 2006.

[14] F. Paganini, "A set-based approach for white noise modeling," *IEEE Trans. Automatic Control*, vol. 41, no. 10, pp. 1453–1465, Oct. 1996.

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

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

[17] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 374–377, Mar. 2002.

[18] M. R. Raghavendra and K. Giridhar, "Improving channel estimation in OFDM systems for sparse multipath channels," *IEEE Signal Processing Lett.*, vol. 12, no. 1, pp. 52–55, Jan. 2005.

[19] C. Carbonelli, S. Vedantam, and U. Mitra, "Sparse channel estimation with zero tap detection," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1743–1753, May 2007.

[20] 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.

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

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

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

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

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

[33] 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.

[34] G. E. Pfander, H. Rauhut, and J. Tanner, "Identification of matrices having a sparse representation," *IEEE Trans. Signal Processing*, vol. 56, no. 11, pp. 5376–5388, Nov. 2008.

[35] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2275–2284, Jun. 2009.

[36] 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.

[37] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imaging Science*, vol. 2, no. 4, pp. 1098–1128, Nov. 2009.

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

[39] 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, Mar. 2003.

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

[41] 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.

Jarvis Haupt Jarvis Haupt received the B.S. (with highest distinction), M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin–Madison in 2002, 2003, and 2009, respectively.

He is currently a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering at Rice University in Houston TX, and will join the Department of Electrical and Computer Engineering at the University of Minnesota an Assistant Professor in Fall 2010. His research interests include high-dimensional statistical inference, adaptive sampling techniques, statistical signal processing and learning theory, and applications in the biological sciences, communications, imaging, and networks.

Dr. Haupt has completed internships at Georgia Pacific, Domtar Industries, Cray, and L-3 Communications/Integrated Systems. He is the recipient of several academic awards, including the Wisconsin Academic Excellence Scholarship, the Ford Motor Company Scholarship, the Consolidated Papers Tuition Scholarship, the Frank D. Cady Mathematics Scholarship, and the Claude and Dora Richardson Distinguished Fellowship. He served as Co-Chair of the Teaching Improvement Program at the University of Wisconsin–Madison for two semesters, and received Honorable Mention for the Gerald Holdridge Teaching Award for his work as a teaching assistant. Dr. Haupt is also a Certified Professional Locksmith.

Waheed U. Bajwa Waheed U. Bajwa received the BE (with Honors) degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan in 2001, and the MS and PhD degrees in electrical engineering from the University of Wisconsin-Madison, Madison, WI in 2005 and 2009, respectively.

He was affiliated with Communications Enabling Technologies, Islamabad, Pakistan - the research arm of Avaz Networks Inc., Irvine, CA (now Quartics LLC) - from 2000-2003, with the Center for Advanced Research in Engineering, Islamabad, Pakistan during 2003, and with the RF and Photonics Lab of GE Global Research, Niskayuna, NY during the summer of 2006. He is currently a Postdoctoral Research Associate in the Program in Applied and Computational Mathematics at Princeton University. His research interests include high-dimensional inference, statistical signal processing, wireless communications, learning theory, and applications in biological sciences, cyber-physical systems, radar and image processing, and cognitive/ad-hoc networks.

Dr. Bajwa received the Best in Academics Gold Medal and President's Gold Medal in Electrical Engineering from the National University of Sciences and Technology (NUST) in 2001, and the Morgridge Distinguished Graduate Fellowship from the University of Wisconsin-Madison in 2003. He was Junior NUST Student of the Year (2000), Wisconsin Union Poker Series Champion (Spring 2008), and President of the University of Wisconsin-Madison chapter of Golden Key International Honor Society (2009). He is currently a member of the IEEE, Pakistan Engineering Council, and Golden Key International Honor Society.

Gil Raz Dr. Gil Raz is the founder and president of GMR Research & Technology, Inc. Previously Dr. Raz was a staff member at MIT - Lincoln Laboratory. Prior to working at MIT he worked as a consultant for several firms implementing real-time signal processing algorithms as well as working at National Semiconductor as a software engineer implementing real-time embedded algorithms.

Dr. Raz earned his Ph.D. degree at the University of Wisconsin - Madison and his undergraduate degree at the Technion - Israel Institute of Technology.

Robert Nowak Robert Nowak received the B.S. (with highest distinction), M.S., and Ph.D. degrees in electrical engineering from the University of Wisconsin-Madison in 1990, 1992, and 1995, respectively.

He was a Postdoctoral Fellow at Rice University in 1995-1996, an Assistant Professor at Michigan State University from 1996-1999, held Assistant and Associate Professor positions at Rice University from 1999-2003, and was a Visiting Professor at INRIA, France, in 2001. He is now the McFarland-Bascom Professor of Engineering at the University of Wisconsin-Madison. His research interests include signal processing, machine learning, imaging and network science, and applications in communications, bioimaging, and systems biology.

Dr. Nowak has served as an Associate Editor for the IEEE Transactions on Image Processing, the Secretary of the SIAM Activity Group on Imaging Science, and is currently an Associate Editor for the ACM Transactions on Sensor Networks. He was General Chair for the 2007 IEEE Statistical Signal Processing workshop and Technical Program Chair for the 2003 IEEE Statistical Signal Processing Workshop and the 2004 IEEE/ACM International Symposium on Information Processing in Sensor Networks. Dr. Nowak received the General Electric Genius of Invention Award in 1993, the National Science Foundation CAREER Award in 1997, the Army Research Office Young Investigator Program Award in 1999, the Office of Naval Research Young Investigator Program Award in 2000, and IEEE Signal Processing Society Young Author Best Paper Award in 2000.