A Tutorial on Recovery Conditions for Compressive System Identification of Sparse Channels

Borhan M. Sanandaji,⋆ Tyrone L. Vincent,⋆ Kameshwar Poolla,⋄ and Michael B. Wakin⋆

***Abstract*— In this tutorial, we review some of the recent results concerning Compressive System Identification (CSI) (identification from few measurements) of *sparse channels* (and in general, Finite Impulse Response (FIR) systems) when it is known a priori that the impulse response of the system under study is sparse (high-dimensional but with few non-zero entries) in an appropriate basis. For the systems under study in this tutorial, the system identification problem boils down to an inverse problem of the form $A\boldsymbol{x} = \boldsymbol{b}$, where the vector $\boldsymbol{x} \in \mathbb{R}^N$ is high-dimensional but with $K \ll N$ non-zero entries and the matrix $A \in \mathbb{R}^{M \times N}$ is underdetermined (i.e., $M < N$). Over the past few years, several algorithms with corresponding recovery conditions have been proposed to perform such a recovery. These conditions provide the number of measurements sufficient for correct recovery. In this note, we review alternate approaches to derive such recovery conditions concerning CSI of FIR systems whose impulse response is known to be sparse.**

I. Introduction

A. High-dimensional but Sparse Dynamical Systems

The fundamental dimensionality of many systems of practical interest is often much less than what is suggested by the dimensionality of a particular choice for parameterization. Many systems are often either low-order or can be represented in a suitable basis or formulation in which the number of significant degrees of freedom is small. We refer to such systems simply as *sparse systems*. In terms of a difference equation, for example, a sparse system may be a high-dimensional system with only a few non-zero coefficients or it may be a system with an impulse response that is long but contains only a few non-zero terms. Multipath propagation [1]–[3], sparse channel estimation [4]–[6], large-scale interconnected systems with sparse graph flow [7]–[10], time varying systems with few piecewise-constant parameter changes [11], [12], and sparse initial state estimation [13], [14] are examples involving dynamical systems that are high-dimensional in terms of their ambient dimension but have a sparse (low-order) representation.

B. Exploiting Sparsity via Compressive System Identification

System identification [15] is a field in control theory that aims at identifying an equivalent model of a system based on its input-output observations. Classical system identification approaches have limited performance in cases when the number of available data samples is small compared to the order of the system [15]. These approaches usually require a large data set in order to achieve a certain performance due to the asymptotic nature of their analysis. On the other hand, there are many application fields where only limited data sets are available. Online estimation, Linear Time-Variant (LTV) system identification and setpoint-operated processes are examples of situations for which limited data samples are available. For some specific applications, the cost of the measuring process or the computational effort is also an issue. In such situations, it is *necessary* to perform the system identification from the smallest possible number of observations, although doing so leaves an apparently ill-conditioned identification problem. Solving such an ill-conditioned problem for a unique solution is impossible unless one has *extra* information about the true solution.

For high-dimensional but sparse dynamical systems and inspired by Compressive Sensing (CS), we aim at exploiting sparsity by performing system identification using a number of observations that is smaller than the ambient dimension. We categorize such approaches as Compressive System Identification (CSI). CSI is beneficial in applications when only a limited data set is available.

C. CSI of Sparse Channels

As mentioned earlier, many Linear Time-Invariant (LTI) systems of practical interest have a sparse impulse response [4]. Sparse channels are examples of such systems that have broad applications in signal processing, seismic imaging, radar imaging, and wireless multipass channel estimation [6]. The measurement process in such problems may involve the convolution of an (unknown) system impulse response with a (known) input (probe). In general, this problem can be considered as identifying the impulse response of a Finite Impulse Response (FIR) LTI system from its input-output observations.

Let $\{a_k\}_{k=1}^{N+M-1}$ be the applied input sequence to an LTI system characterized by its finite impulse response $\boldsymbol{x} = \{x_k\}_{k=1}^{N}$. Then the corresponding output is calculated from the time-domain convolution. Considering the a_k and x_k sequences to be zero-padded from both sides, each output sample b_k can be written as

$$b_k = \sum_{j=1}^{N} x_j a_{k-j}. \tag{1}$$

⋆B. M. Sanandaji, T. L. Vincent, and M. B. Wakin are with Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401, USA. {bmolazem, tvincent, mwakin}@mines.edu.

⋄K. Poolla is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering, University of California, Berkeley, CA 94720, USA. poolla@berkeley.edu.

This work was partially supported by AFOSR Grant FA9550-09-1-0465, NSF Grant CCF-0830320, NSF Grant CNS-0931748 and NWO Grant 680-50-0927.

If we only keep M consecutive observations of the system, $\boldsymbol{b} = \{b_k\}_{k=N+1}^{N+M}$, then (1) can be written in a matrix-vector multiplication format as

$$\boldsymbol{b} = A\boldsymbol{x}, \tag{2}$$

where

$$A = \begin{bmatrix} a_N & a_{N-1} & \cdots & a_1 \\ a_{N+1} & a_N & \cdots & a_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{N+M-1} & a_{N+M-2} & \cdots & a_M \end{bmatrix} \tag{3}$$

is an $M \times N$ Toeplitz matrix. Supposing the system has a K-sparse impulse response, we are interested in efficiently acquiring and retrieving $\boldsymbol{x}$ via its multiplication by the $M \times N$ compressive ($M < N$) Toeplitz matrix A. In addition to the application to convolution-based measurement scenarios, it is worth noting that while fully populated $M \times N$ random matrices require NM elements to be generated, Toeplitz matrices require only $N + M - 1$ distinct random entries, which may provide an advantage in other CS applications.

In this tutorial, we review and derive bounds on the sufficient number of measurements M (namely, recovery conditions) such that (2) (a highly underdetermined set of linear equations) has a *correct* solution by solving an ℓ_1-minimization problem. In what follows, we first review some of the basic concepts in CS and then explain its use in the context of channel estimation and FIR systems.

II. Compressive Sensing (CS)

CS, introduced by Candès, Romberg and Tao [16] and Donoho [17], is a powerful paradigm in signal processing which enables the recovery of an unknown vector from an underdetermined set of measurements under the assumption of sparsity of the signal and under certain conditions on the measurement matrix. The CS recovery problem can be viewed as recovery of a K-sparse signal $\boldsymbol{x} \in \mathbb{R}^N$ from its observations $\boldsymbol{b} = A\boldsymbol{x} \in \mathbb{R}^M$ where $A \in \mathbb{R}^{M \times N}$ is the measurement matrix with $M < N$ (in many cases $M \ll N$). A K-sparse signal $\boldsymbol{x} \in \mathbb{R}^N$ is a signal of length N with K non-zero entries where $K \ll N$. The notation $K := \|\boldsymbol{x}\|_0$ denotes the sparsity level of $\boldsymbol{x}$. Since the matrix $A \in \mathbb{R}^{M \times N}$ has a non-trivial nullspace when $M < N$, there exist infinitely many solutions to the equation $\boldsymbol{b} = A\boldsymbol{x}$, given $\boldsymbol{b}$. However, recovery of $\boldsymbol{x}$ is indeed possible from CS measurements if the true signal is known to be sparse. Recovery of the true signal can be accomplished by seeking a sparse solution among these candidates.

A. Recovery via ℓ_0-minimization

Supposing that $\boldsymbol{x}$ is exactly K-sparse, then recovery of $\boldsymbol{x}$ from $\boldsymbol{b}$ can be formulated as the ℓ_0-minimization

$$\widehat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_0 \qquad \text{subject to} \qquad \boldsymbol{b} = A\boldsymbol{x}. \tag{4}$$

When A is populated with random entries and given some technical conditions on A, then with high probability this optimization problem returns the proper K-sparse solution $\boldsymbol{x}$. We note that the recovery program (4) can be interpreted as finding a K-term approximation to $\boldsymbol{b}$ from the columns of the dictionary A [17], [18].

B. Recovery via ℓ_1-minimization

Unfortunately, solving the ℓ_0-minimization problem (4) is known to be NP-hard. Thanks to the results of CS in regards to sparse signal recovery, however, it has been discovered that it is not always necessary to solve the ℓ_0-minimization problem (4) to recover $\boldsymbol{x}$. In fact, a much easier problem often yields an equivalent solution: we only need to solve for the ℓ_1-sparsest $\boldsymbol{x}$ that agrees with the measurements $\boldsymbol{b}$ [16], [17], [19]–[22] by solving

$$\widehat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_1 \qquad \text{subject to} \qquad \boldsymbol{b} = A\boldsymbol{x}. \tag{5}$$

The ℓ_1-minimization problem (5), also known as Basis Pursuit (BP) [23], is significantly more approachable and can be solved with traditional linear programming techniques whose computational complexities are polynomial in N.

C. ℓ_0/ℓ_1 Equivalence and the Restricted Isometry Property

The Restricted Isometry Property (RIP), introduced by Candès and Tao [21], is one of the most fundamental recovery conditions that has been studied in the CS literature [24]. Establishing the RIP for a given matrix A guarantees that the solution to the ℓ_1-minimization problem is equivalent to the solution to the ℓ_0-minimization problem.

Definition 1: The matrix $A \in \mathbb{R}^{M \times N}$ has the RIP of order K if there exists a constant $\epsilon_K \in (0, 1)$ such that

$$(1 - \epsilon_K) \leq \frac{\|A\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \leq (1 + \epsilon_K) \tag{6}$$

holds for all K-sparse signals $\boldsymbol{x} \in \mathbb{R}^N$ ($\forall \boldsymbol{x}$ with $\|\boldsymbol{x}\|_0 \leq K$).

A useful interpretation of this condition is in terms of the singular values. Establishing the RIP of order K with isometry constant ϵ_K (namely, RIP(K, ϵ_K)) is equivalent to restricting all the eigenvalues of all the submatrices $A_S^T A_S$ to the interval $(1 - \epsilon_K, 1 + \epsilon_K)$. In this notation, A_S is an $M \times K$ submatrix of A whose columns are those columns of A indexed by the set S with cardinality $|S| \leq K$. Note that by definition of the RIP we need to satisfy this condition for all the $\binom{N}{K}$ submatrices $A_S^T A_S$ [4], [25], [26]. Therefore establishing the RIP for a given matrix is a combinatorial task. However, it has been shown that some specific matrix ensembles satisfy the RIP. For example, Baraniuk et al. [25] showed that a random $M \times N$ matrix with independent and identically distributed (i.i.d.) Gaussian entries drawn from $\mathcal{N}\left(0, \frac{1}{M}\right)$ will satisfy the RIP of order K with high probability if $M = \mathcal{O}\left(K \log \frac{N}{K}\right)$. For the ℓ_1-minimization problem (5), establishing the RIP of order $2K$ with constant $\epsilon_{2K} < 0.4652$ for a given A guarantees exact recovery of any K-sparse signal $\boldsymbol{x} \in \mathbb{R}^N$ [20], [26], [27]. Robust recovery is also possible in the presence of measurement noise and for compressible signals [27]. While other recovery guarantees including the Exact Recovery Condition (ERC) [28], and the mutual coherence [29], [30] have been proposed in the CS literature, in this tutorial we consider the RIP [21].

III. RIP for Toeplitz Matrices

As mentioned earlier, the sparse channel estimation problem boils down to recovery of a signal $\boldsymbol{x} \in \mathbb{R}^N$ from an underdetermined set of linear equations $\boldsymbol{b} = A\boldsymbol{x}$ where the matrix A is a Toeplitz ensemble of the form (3). In order to have exact unique recovery of $\boldsymbol{x}$ via the ℓ_1-minimization problem (5), it is sufficient to show that A satisfies the RIP.

Compressive Toeplitz (and circulant) matrices have been previously studied in the context of CS in [1]–[4], [31]–[33]. Recently Rauhut et al. [26] improved the best existing quadratic bounds [4], [34] on the number of measurements sufficient for correct recovery. Using more complicated mathematical tools such as Dudley's inequality for chaos and generic chaining, Rauhut et al. demonstrated[1] that random Toeplitz matrices (and in a more general case, partial random circulant matrices) satisfy the RIP of order K if the number of measurements $M \geq K^{1.5} (\log N)^{1.5}$.

In the rest of this tutorial, we review two alternate approaches that can be used for establishing the RIP for Toeplitz matrices. The first approach is based on deriving Concentration of Measure (CoM) inequalities for Toeplitz matrices [3], [34]. Although the RIP estimate based on CoM inequalities is quadratic in terms of the sparsity level (i.e., $M \sim K^2$) and falls short of the best known estimates mentioned above, deriving CoM inequalities for Toeplitz matrices is of its own importance. First, as we will see in the rest of this tutorial, concentration inequalities can be used to explain the signal-dependent behavior of the recovery results. Moreover, they are simpler to derive and also give insight to other applications such as Compressive Binary Detection (CBD) [3], [34], [36].

The second approach follows the work by Haupt et al. [4] and is based on Geršgorin's Disc Theorem [37]. In this tutorial, we basically follow the steps outlined by Haupt et al. [4] to establish the RIP for Toeplitz matrices. However, we derive new related tail probability bounds that result in smaller constants in the final recovery condition.

IV. Concentration of Measure Inequalities

A. Definition

CoM inequalities are one of the leading techniques used in the theoretical analysis of randomized compressive linear operators [38]. These inequalities quantify how well a random matrix will preserve the norm of a high-dimensional signal when mapping it to a low-dimensional space. A typical CoM inequality takes the following form [39]: For any fixed signal $\boldsymbol{x} \in \mathbb{R}^N$ and a suitable random $M \times N$ matrix A, the norm of the signal projected by A will be highly concentrated around the norm of the original signal with high probability. Formally, there exist constants c_1 and c_2 such that for any fixed $\boldsymbol{x} \in \mathbb{R}^N$,

$$\mathbf{P}\left\{\left|\|A\boldsymbol{x}\|_2^2 - \mathbf{E}\left[\|A\boldsymbol{x}\|_2^2\right]\right| \geq \epsilon \mathbf{E}\left[\|A\boldsymbol{x}\|_2^2\right]\right\} \leq c_1 e^{-c_2 M c_0(\epsilon)},$$

where $c_0(\epsilon)$ is a function of $\epsilon \in (0,1)$.

CoM inequalities have been well-studied and derived for *unstructured* random compressive matrices, populated with i.i.d. random entries [39]. However, in many practical applications, measurement matrices possess a certain structure. In particular, when linear dynamical systems are involved, Toeplitz matrices appear due to the convolution process.

B. CoM for Toeplitz Matrices

In [3] we derived CoM bounds for compressive Toeplitz matrices of the form (3) with entries $\{a_k\}_{k=1}^{N+M-1}$ drawn from an i.i.d. Gaussian random sequence. Our main result, detailed in Theorem 1, states that the upper and lower tail probability bounds depend on the number of measurements M and on the eigenvalues of the covariance matrix of the vector $\boldsymbol{x}$ defined as

$$P(\boldsymbol{x}) = \begin{bmatrix} \mathcal{R}_{\boldsymbol{x}}(0) & \mathcal{R}_{\boldsymbol{x}}(1) & \cdots & \mathcal{R}_{\boldsymbol{x}}(M-1) \\ \mathcal{R}_{\boldsymbol{x}}(1) & \mathcal{R}_{\boldsymbol{x}}(0) & \cdots & \mathcal{R}_{\boldsymbol{x}}(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{R}_{\boldsymbol{x}}(M-1) & \mathcal{R}_{\boldsymbol{x}}(M-2) & \cdots & \mathcal{R}_{\boldsymbol{x}}(0) \end{bmatrix},$$

where

$$\mathcal{R}_{\boldsymbol{x}}(\tau) := \sum_{i=1}^{N-\tau} x_i x_{i+\tau}$$

denotes the un-normalized sample autocorrelation function of $\boldsymbol{x} \in \mathbb{R}^N$.

Theorem 1 ([3]): Let $\boldsymbol{x} \in \mathbb{R}^N$ be fixed. Define two quantities $\rho(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ associated with the eigenvalues of the covariance matrix $P(\boldsymbol{x})$ as

$$\rho(\boldsymbol{x}) = \frac{\max_i \lambda_i}{\|\boldsymbol{x}\|_2^2}$$

and

$$\mu(\boldsymbol{x}) = \frac{\sum_{i=1}^M \lambda_i^2}{M\|\boldsymbol{x}\|_2^4},$$

where λ_i is the i-th eigenvalue of $P(\boldsymbol{x})$. Let $\boldsymbol{b} = A\boldsymbol{x}$, where A is a random compressive Toeplitz matrix of the form (3) populated with i.i.d. Gaussian entries having zero mean and unit variance. Noting that $\mathbf{E}\left[\|\boldsymbol{b}\|_2^2\right] = M\|\boldsymbol{x}\|_2^2$, then for any $\epsilon \in (0,1)$, the upper tail probability bound is

$$\mathbf{P}\left\{\|\boldsymbol{b}\|_2^2 - M\|\boldsymbol{x}\|_2^2 \geq \epsilon M\|\boldsymbol{x}\|_2^2\right\} \leq e^{-\frac{\epsilon^2 M}{8\rho(\boldsymbol{x})}} \quad (7)$$

and the lower tail probability bound is

$$\mathbf{P}\left\{\|\boldsymbol{b}\|_2^2 - M\|\boldsymbol{x}\|_2^2 \leq -\epsilon M\|\boldsymbol{x}\|_2^2\right\} \leq e^{-\frac{\epsilon^2 M}{8\mu(\boldsymbol{x})}}. \quad (8)$$

Observe that for all $\boldsymbol{x} \in \mathbb{R}^N$, $\mu(\boldsymbol{x}) \leq \rho(\boldsymbol{x})$. Thus,

$$\mathbf{P}\left\{\left|\|\boldsymbol{b}\|_2^2 - M\|\boldsymbol{x}\|_2^2\right| \geq \epsilon M\|\boldsymbol{x}\|_2^2\right\} \leq 2e^{-\frac{\epsilon^2 M}{8\rho(\boldsymbol{x})}}. \quad (9)$$

[1] After the initial submission of our manuscript, in their very recent work, Krahmer et al. [35] showed that the minimal number of measurements sufficient for correct recovery scales linearly with K, or formally $M \geq \mathcal{O}\left(K \log(K)^2 \log(N)^2\right)$ measurements are sufficient to establish the RIP. The recent linear RIP result confirms what is suggested by simulations.

C. Implications

Theorem 1 provides CoM inequalities for *any* (not necessarily sparse) signal $\boldsymbol{x} \in \mathbb{R}^N$. The significance of these results comes from the fact that the tail probability bounds are functions of the signal $\boldsymbol{x}$, where the dependency is captured in the quantities $\rho(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$. That is, the concentration measures are *signal-dependent*, which is not the case when A is unstructured. Indeed, allowing A to have $M \times N$ i.i.d. Gaussian entries with zero mean and unit variance (and thus, no Toeplitz structure) would result in the concentration bound (see, e.g., [39])

$$\mathbf{P}\left\{\left|\|\boldsymbol{b}\|_2^2 - M\|\boldsymbol{x}\|_2^2\right| \geq \epsilon M \|\boldsymbol{x}\|_2^2\right\} \leq 2e^{-\frac{\epsilon^2 M}{4}}. \quad (10)$$

Comparing (9) with (10) reveals that achieving the same probability bound for Toeplitz matrices requires choosing M larger by a factor of $2\rho(\boldsymbol{x})$. Typically, when using CoM bounds such as (7) and (8), we must set M large enough so that both bounds are sufficiently small over all signals $\boldsymbol{x}$ belonging to some class of interest. For example, we are often interested in signals that have a *sparse* representation. Because we generally wish to keep M as small as possible, it is interesting to try to obtain an upper bound for the quantities $\rho(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ over the class of signals of interest. For the class of K-sparse signals, it is trivial to show that $\mu(\boldsymbol{x}) \leq \rho(\boldsymbol{x}) \leq K$. This bound in fact can be used in establishing RIP for Toeplitz matrices. In a previous work, a CoM inequality of the form (10) has been used to show that an unstructured matrix A satisfies the RIP with high probability if $M \geq \mathcal{O}\left(K \log\left(\frac{N}{K}\right)\right)$ [25]. An approach identical to the one taken in [25] can be used to establish the RIP for Toeplitz matrices based on the CoM inequalities given in Theorem 1. In particular, since $\rho(\boldsymbol{a})$ is bounded by K for all K-sparse signals in the time domain, we have the following result.

Theorem 2 ([3], [34]): An $M \times N$ Toeplitz matrix A with i.i.d. Gaussian entries satisfies the RIP of order K with high probability if $M = \mathcal{O}\left(K^2 \log\left(\frac{N}{K}\right)\right)$.

D. Simulation Results

Although the result of Theorem 2 demonstrates a quadratic dependency of M in K for all K-sparse signals, simulations strongly suggest that the minimal number of measurements sufficient for correct recovery scales linearly with K. Moreover, simulation results depict different recovery performance for different K-sparse signal classes when measured by Toeplitz matrices. Fig. 1(a) depicts the required number of measurements M for having a recovery rate exceeding 0.99 over 1000 realizations of $\boldsymbol{b} = A\boldsymbol{x}$ for each sparsity level K for signals of length $N = 512$. In order to emphasize the signal-dependency of the recovery performance, we repeat the simulation for two different signal classes: 1) K-sparse signals with random support (i.e., location of the non-zero entries) and random sign entries (we refer to these signals as Scattered), and 2) K-sparse signals with block support and same sign entries (we refer to these signals as Block Same Sign). As can be seen, Block Same Sign signals have

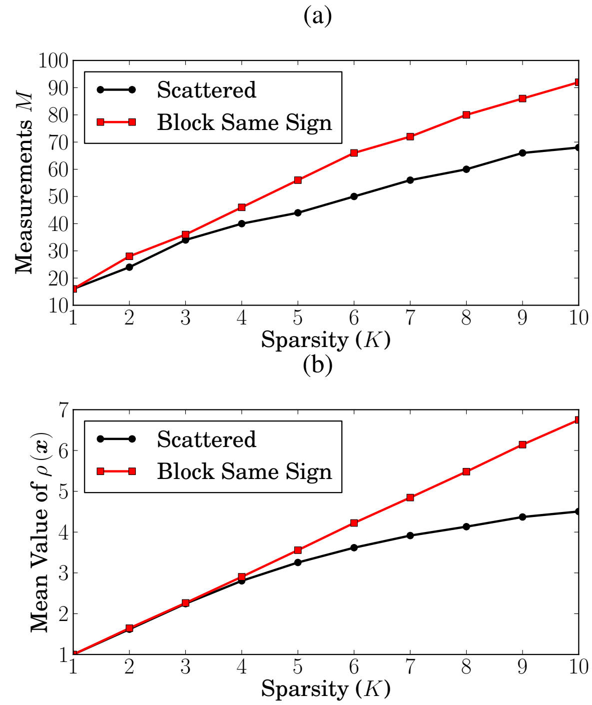

Fig. 1. Results regards to two different signal classes: 1) K-sparse signals with random support and random sign entries (we refer to these signals as Scattered), and 2) K-sparse signals with block support and same sign entries (we refer to these signals as Block Same Sign) of length $N = 512$. (a) Required number of measurements M to achieve a recovery rate exceeding 0.99 over 1000 realizations of $\boldsymbol{b} = A\boldsymbol{x}$ for each sparsity level K. (b) Mean value of $\rho(\boldsymbol{x})$ for 1000 realizations of $\boldsymbol{x}$ for each sparsity level K.

weaker recovery performance when measured by Toeplitz matrices. As a benefit of deriving CoM inequalities for Toeplitz matrices and Theorem 1 we can trace this signal-dependency behavior in the recovery performance back to the concentration performance. Fig. 1(b) shows the mean value of $\rho(\boldsymbol{x})$ for 1000 realizations for each sparsity level K for signals of length $N = 512$ and for the two above-mentioned signal classes. Observe that Block Same Sign signals have higher mean $\rho(\boldsymbol{x})$ value compared to Scattered signals. For a more detailed analysis on the signal-dependency behavior of $\rho(\boldsymbol{x})$ in different bases and for different signal classes see [34].

V. Eigenvalue Analysis via Geršgorin Theorem

Recall the eigenvalue interpretation of the RIP explained in II-C. As the second considered approach for establishing RIP for Topelitz matrices in this paper, we use this interpretation which requires deriving tail probability bounds on the pairwise products of the columns of a Toeplitz matrix. Our derivation follows the steps of the work of Haupt et al. [4]. However, we derive new tail probability bounds that result in smaller constants in the final recovery condition. We start by presenting Geršgorin's Disc Theorem.

Lemma 1 (Geršgorin's Disc Theorem [37]): Consider a matrix $P \in \mathbb{R}^{N \times N}$. For each row of P, calculate the absolute sum of the off-diagonal entries as $r_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} |P_{ij}|$. Then, all the eigenvalues of P lie in $\bigcup_{i=1}^{N} D(P_{ii}, r_i)$ where $D(P_{ii}, r_i)$ is a disc centered at P_{ii} with radius r_i.

Define the Grammian matrix $P = A^T A \in \mathbb{R}^{N \times N}$. Using the results of Lemma 1, it is trivial to see that if we assume that for all the N diagonal entries of P, we have $|P_{ii} - 1| < \delta_1$ and for all the $\frac{N(N-1)}{2}$ distinct off-diagonal entries of P, we have $|P_{ij}| < \frac{\delta_2}{K}$, then all the eigenvalues of any $K \times K$ submatrix of P formed by $A_S^T A_S$ lie in $\left(1 - \delta_1 - (K-1)\frac{\delta_2}{K}, 1 + \delta_1 + (K-1)\frac{\delta_2}{K}\right)$. Assuming $\delta_1 + \delta_2 = \delta_K$ this implies all the eigenvalues of $A_S^T A_S$ for any set $S \subset \{1, 2, \dots, N\}$ with $|S| \leq K$ lie in $(1 - \delta_K, 1 + \delta_K)$. We now derive tail probability bounds on the diagonal and off-diagonal entries of P.

Lemma 2 ([34]): Define the random variable $z \in \mathbb{R}$ as $z = \sum_{i=1}^{M} a_i^2$ where $\{a_i\}_{i=1}^{M}$ are i.i.d. Gaussian random variables with $\mathcal{N}\left(0, \frac{1}{M}\right)$. Then

$$\mathbf{P}\{z - 1 \geq \epsilon\} \leq \left((1+\epsilon)e^{-\epsilon}\right)^{\frac{M}{2}}, \qquad \forall \epsilon \geq 0,$$

and

$$\mathbf{P}\{z - 1 \leq -\epsilon\} \leq \left((1-\epsilon)e^{+\epsilon}\right)^{\frac{M}{2}}, \qquad \forall \epsilon \geq 0.$$

Using the fact that $(1+\epsilon)e^{-\epsilon} \leq e^{-\frac{\epsilon^2}{3}}$ for $0 \leq \epsilon \leq 0.787$ and $(1-\epsilon)e^{+\epsilon} \leq e^{-\frac{\epsilon^2}{2}}$ for $\forall \epsilon \geq 0$, we have

$$\mathbf{P}\{|z - 1| \geq \epsilon\} \leq 2e^{-\frac{\epsilon^2 M}{6}}, \qquad 0 \leq \epsilon \leq 0.787.$$

Lemma 3 ([3], [39]): Assume $a_1 \sim \mathcal{N}(0, \sigma_1)$ and $a_2 \sim \mathcal{N}(0, \sigma_2)$ are two independent Gaussian random variables. Then for $s > 0$

$$\mathbf{E}\left[e^{s a_1 a_2}\right] = \frac{1}{\sqrt{1 - s^2 \sigma_1^2 \sigma_2^2}}.$$

Lemma 4: Define the random variable $z \in \mathbb{R}$ as

$$z = \sum_{i=1}^{M} a_{2i-1} a_{2i} = [a_1 a_2 + a_3 a_4 + \cdots + a_{2M-1} a_{2M}]$$

where $\{a_i\}_{i=1}^{2M}$ are i.i.d. Gaussian random variables with $\mathcal{N}\left(0, \frac{1}{M}\right)$. Then

$$\mathbf{P}\{|z| \geq \epsilon\} \leq 2\left((1+\epsilon)e^{-\epsilon}\right)^{M}, \qquad \forall \epsilon \geq 0.$$

Using the fact that $(1+\epsilon)e^{-\epsilon} \leq e^{-\frac{\epsilon^2}{3}}$ for $0 \leq \epsilon \leq 0.787$,

$$\mathbf{P}\{|z| \geq \epsilon\} \leq 2e^{-\frac{\epsilon^2 M}{3}}, \qquad 0 \leq \epsilon \leq 0.787.$$

Proof See Appendix A. ■

Lemma 5: Define the random variable $z \in \mathbb{R}$ as a summation of M terms $a_i a_j$ $(i \neq j)$ where each a_i is an i.i.d. Gaussian random variable with $\mathcal{N}\left(0, \frac{1}{M}\right)$. Then

$$\mathbf{P}\{|z| \geq \epsilon\} \leq 4\left(e^{-\frac{\epsilon^2}{12}}\right)^{\lfloor \frac{M}{2} \rfloor}, \qquad 0 \leq \epsilon \leq 0.787.$$

Proof See Appendix B. ■

Note that the result of Lemma 2 applies to the tail probability bound on the diagonal entries of P while the result of Lemma 5 applies to the tail probability bound on the off-diagonal entries of P. Combining these results with Geršgorin's Disc Theorem and using the union bound we get the following RIP result.

Theorem 3: Fix $\nu \in (0,1)$. Then the matrix $A \in \mathbb{R}^{M \times N}$ satisfies the RIP of order K for some constant $\epsilon_K \in (0,1)$ with probability at least $1 - \nu^2$ if $M \geq \frac{192}{\epsilon_K^2} K^2 \log \frac{\sqrt{3}N}{\nu}$.

Proof See Appendix C. ■

VI. CONCLUSION

In this tutorial, we reviewed two alternate approaches that can be used for establishing the RIP for Toeplitz matrices. The first approach was based on deriving CoM inequalities for Toeplitz matrices. Although the RIP estimate based on CoM inequalities is quadratic in terms of the sparsity level (i.e., $M \sim K^2$) and falls short of the best known estimates, we showed that deriving CoM inequalities for Toeplitz matrices is of its own importance. First, such concentration inequalities can be used to explain the signal-dependent behavior of the recovery results. Moreover, they are simpler to derive and can be used in other applications such as the binary detection problem.

The second approach was based on Geršgorin's Disc Theorem and followed the existing steps outlined by Haupt et al. to establish the RIP for Toeplitz matrices. Deriving different tail probability bounds, however, we established the RIP using smaller constants.

APPENDIX

A. Proof of Lemma 4

Observe that $\mathbf{E}[z] = 0$. Using a Chernoff bound, we have

$$\mathbf{P}\{z - \mathbf{E}[z] \geq \epsilon\} = \mathbf{P}\{z \geq \epsilon\} \leq \min_{s>0} \frac{\mathbf{E}\left[e^{sz}\right]}{e^{s\epsilon}}.$$

By Lemma 3 and using the independenc of the M terms in z (i.e., $\{a_{2i-1}a_{2i}\}_{i=1}^{M}$), we have, for $s > 0$

$$\mathbf{E}\left[e^{sz}\right] = \left(\frac{1}{\sqrt{1 - \left(\frac{s}{M}\right)^2}}\right)^{M}.$$

Therefore,

$$\mathbf{P}\{z \geq \epsilon\} \leq \min_{s>0} \left(1 - \left(\frac{s}{M}\right)^2\right)^{-\frac{M}{2}} e^{-s\epsilon}.$$

Let $f(s) = \left(1 - \left(\frac{s}{M}\right)^2\right)^{-\frac{M}{2}} e^{-s\epsilon}$. We have

$$\frac{df(s)}{ds} = 0 \Leftrightarrow \epsilon s^2 + Ms - \epsilon M^2 = 0.$$

The function $f(s)$ has a minimum at

$$s^\star = M \frac{\sqrt{1 + 4\epsilon^2} - 1}{2\epsilon},$$

which equivalently satisfies

$$1-\left(\frac{s^\star}{M}\right)^2=\frac{s^\star}{M\epsilon}.$$

Therefore,

$$\begin{aligned}\mathbf{P}\left\{z\geq\epsilon\right\}\leq f(s^\star)&=\left(\left(\frac{\sqrt{1+4\epsilon^2}-1}{2\epsilon^2}\right)^{-\frac{1}{2}}e^{-\frac{\sqrt{1+4\epsilon^2}-1}{2}}\right)^M\\&=:\left(g\left(\epsilon\right)\right)^M\\&\leq\left(\left(1+\epsilon\right)e^{-\epsilon}\right)^M,\end{aligned}$$

where we use the inequality $g(\epsilon)\leq(1+\epsilon)\,e^{-\epsilon}$, $\forall\epsilon\geq 0$. Following the similar steps, we have for $s>0$

$$\begin{aligned}\mathbf{P}\left\{z\leq-\epsilon\right\}=\mathbf{P}\left\{-z\geq\epsilon\right\}&\leq\min_{s>0}\frac{\mathbf{E}\left[e^{-sz}\right]}{e^{s\epsilon}}\\&=\min_{s>0}\left(1-\left(\frac{s}{M}\right)^2\right)^{-\frac{M}{2}}e^{-s\epsilon}\\&\leq\left(\left(1+\epsilon\right)e^{-\epsilon}\right)^M.\end{aligned}$$

Combining the two tail probability bounds, we have

$$\begin{aligned}\mathbf{P}\left\{|z|\geq\epsilon\right\}&\leq\mathbf{P}\left\{z\geq\epsilon\right\}+\mathbf{P}\left\{z\leq-\epsilon\right\}\\&\leq 2\left(\left(1+\epsilon\right)e^{-\epsilon}\right)^M\qquad\forall\epsilon\geq 0.\end{aligned}$$

B. Proof of Lemma 5

Observe that compared to the definition of z in Lemma 4, the M elements in z are not necessarily independent anymore. However, Haupt et al. [4] showed that with a clever grouping of the M variables in z, we can always write z as a sum of two variables whose elements are independent. Assume without loss of generality that M is even. As an example, assume

$$z=\sum_{i=1}^M a_ia_{i+1}=\left[a_1a_2+a_2a_3+\cdots+a_Ma_{M+1}\right].$$

Apparently, the M terms in z are not independent. For example, a_1a_2 and a_2a_3 are dependent terms (both depend on a_2). However, we can write z as a sum of z_1 and z_2 where

$$z_1=\sum_{i=1}^{\frac{M}{2}}a_{2i-1}a_{2i}=\left[a_1a_2+a_3a_4+\cdots+a_{M-1}a_M\right],$$

and

$$z_2=\sum_{i=1}^{\frac{M}{2}}a_{2i}a_{2i+1}=\left[a_2a_3+a_4a_5+\cdots+a_Ma_{M+1}\right].$$

By this grouping, each term z_1 and z_2 has $\frac{M}{2}$ independent terms. Therefore, we can apply the result of Lemma 4 to each of z_1 and z_2. We have

$$\begin{aligned}\mathbf{P}\left\{|z|\geq\epsilon\right\}&=\mathbf{P}\left\{|z_1+z_2|\geq\epsilon\right\}\\&\leq\mathbf{P}\left\{|z_1|\geq\frac{\epsilon}{2}\right\}+\mathbf{P}\left\{|z_2|\geq\frac{\epsilon}{2}\right\}\\&\leq 4\left(e^{-\frac{\epsilon^2}{12}}\right)^{\frac{M}{2}}\end{aligned}$$

Moreover, it is always possible to write z as a sum of two variables whose elements are independent. It is trivial to show that in any situation we can choose elements in z_1 and z_2 such that they are a summation of at least $\lfloor\frac{M}{2}\rfloor$ terms.

C. Proof of Theorem 3

Without loss of generality, assume M is even. From Lemma 2 and Lemma 5, we have for $0\leq\epsilon\leq 0.787$

$$\begin{aligned}\mathbf{P}\left\{|P_{ii}-1|\geq\epsilon\right\}&\leq 2e^{-\frac{\epsilon^2M}{6}},\\\mathbf{P}\left\{|P_{ij}|\geq\epsilon\right\}&\leq 4e^{-\frac{\epsilon^2M}{24}}.\end{aligned}$$

Therefore for some ϵ_1 and ϵ_2 such that $\epsilon_1+\epsilon_2=\epsilon_K\in(0,1)$ and using the union bound we have

$$\mathbf{P}\left\{\bigcup_{i=1}^N|P_{ii}-1|\geq\epsilon_1\right\}\leq 2Ne^{-\frac{\epsilon_1^2M}{6}}\tag{11}$$

and

$$\begin{aligned}\mathbf{P}\left\{\bigcup_{i=1}^N\bigcup_{j=1,j\neq i}^N|P_{ij}|\geq\frac{\epsilon_2}{K}\right\}&\leq 4\frac{N\left(N-1\right)}{2}e^{-\frac{\epsilon_2^2M}{24K^2}}\\&\leq 2N^2e^{-\frac{\epsilon_2^2M}{24K^2}}.\end{aligned}\tag{12}$$

Now define event A as $\mathcal{E}_A\triangleq\left\{\bigcap_{i=1}^N|P_{ii}-1|\leq\epsilon_1\right\}$ and event B as $\mathcal{E}_B\triangleq\left\{\bigcap_{i=1}^N\bigcap_{j=1,j\neq i}^N|P_{ij}|\leq\frac{\epsilon_2}{K}\right\}$. As mentioned earlier, if $\mathcal{E}_A$ and $\mathcal{E}_B$ happen, then A satisfies the RIP of order K with constant ϵ_K. Using the complement events we have

$$\begin{aligned}&\mathbf{P}\left\{A\text{ does NOT satisfy RIP}\left(K,\epsilon_K\right)\right\}\\&\leq\mathbf{P}\left\{\mathcal{E}_A^c\ \text{ or }\ \mathcal{E}_B^c\right\}\\&\leq\mathbf{P}\left\{\bigcup_{i=1}^N|P_{ii}-1|\geq\epsilon_1\right\}+\mathbf{P}\left\{\bigcup_{i=1}^N\bigcup_{j=1,j\neq i}^N|P_{ij}|\geq\frac{\epsilon_2}{K}\right\}\\&\leq 2Ne^{-\frac{\epsilon_1^2M}{6}}+2N^2e^{-\frac{\epsilon_2^2M}{24K^2}}.\end{aligned}$$

Choosing $\epsilon_1=\epsilon_2=\frac{\epsilon_K}{2}$ and for $N\geq 3$ we have

$$\mathbf{P}\left\{A\text{ does NOT satisfy RIP}\left(K,\epsilon_K\right)\right\}\leq 3N^2e^{-\frac{\epsilon_K^2M}{96K^2}}.\tag{13}$$

Therefore for a given $\nu\in(0,1)$, we have

$$\mathbf{P}\left\{A\text{ does NOT satisfy RIP}\left(K,\epsilon_K\right)\right\}\leq\nu^2\tag{14}$$

for

$$M\geq\frac{192}{\epsilon_K^2}K^2\log\frac{\sqrt{3}N}{\nu},\tag{15}$$

which completes the proof.

References

[1] J. Romberg, "Compressive sensing by random convolution," *SIAM Journal on Imaging Sciences*, vol. 2, no. 4, pp. 1098–1128, 2009.

[2] W. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," *Proc. of 42nd Annual Conf. on Inform. Sciences and Systems (CISS08)*, pp. 5–10, 2008.

[3] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Concentration of measure inequalities for compressive Toeplitz matrices with applications to detection and system identification," *Proceedings of the* 49*th IEEE Conference on Decision and Control*, pp. 2922–2929, 2010.

[4] J. Haupt, W. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5862–5875, 2010.

[5] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Compressive system identification in the linear time-invariant framework," *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference*, pp. 783–790, 2011.

[6] J. Romberg, "An overview of recent results on the identification of sparse channels using random probes," 49*th IEEE Conference on Decision and Control*, pp. 2936–2941, 2010.

[7] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Exact topology identification of large-scale interconnected dynamical systems from compressive observations," *Proceedings of the* 2011 *American Control Conference*, pp. 649–656, 2011.

[8] A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[9] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Compressive topology identification of interconnected dynamic systems via clustered orthogonal matching pursuit," *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference*, pp. 174–180, 2011.

[10] ——, "A review on sufficient conditions for structure identification of interconnected systems," *Proceedings of the* 16*th IFAC Symposium on System Identification*, 2012.

[11] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tóth, and K. Poolla, "Compressive system identification of LTI and LTV ARX models," *Proceedings of the* 50*th IEEE Conference on Decision and Control and European Control Conference*, pp. 791–798, 2011.

[12] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automatica*, vol. 46, no. 6, pp. 1107–1111, 2010.

[13] M. B. Wakin, B. M. Sanandaji, and T. L. Vincent, "On the observability of linear systems from random, compressive measurements," *Proceedings of the* 49*th IEEE Conference on Decision and Control*, pp. 4447–4454, 2010.

[14] B. M. Sanandaji, M. B. Wakin, and T. L. Vincent, "Observability with random observations," *Arxiv preprint arXiv:1211.4077*, 2012.

[15] L. Ljung, *System Identification - Theory for the User*. Prentice-Hall, 2nd edition, 1999.

[16] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

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

[18] M. B. Wakin, "The geometry of low-dimensional signal models," *PhD Dissertation, Rice University*, 2006.

[19] E. 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, 2006.

[20] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

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

[22] E. J. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 227–254, 2006.

[23] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[24] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

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

[26] H. Rauhut, J. Romberg, and J. Tropp, "Restricted isometries for partial random circulant matrices," *Applied and Computational Harmonic Analysis*, 2011.

[27] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus-Mathématique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[28] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1030–1051, 2006.

[29] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[30] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[31] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, "Random filters for compressive sampling and reconstruction," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 872–875, 2006.

[32] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, "Toeplitz-structured compressed sensing matrices," *IEEE/SP* 14*th Workshop on Statistical Signal Processing*, pp. 294–298, 2007.

[33] H. Rauhut, "Circulant and Toeplitz matrices in compressed sensing," *Proc. SPARS*, vol. 9, 2009.

[34] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Concentration of measure inequalities for Toeplitz matrices with applications," *to appear in IEEE Transactions on Signal Processing*, 2011.

[35] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," 2012.

[36] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk, "Signal processing with compressive measurements," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 445–460, 2010.

[37] R. Varga, *Geršgorin and his circles*. Springer Series in Computational Mathematics, 2004.

[38] M. Ledoux, *The concentration of measure phenomenon*. Amer Mathematical Society, 2001.

[39] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.