Compressed Channel Sensing

Waheed U. Bajwa†, Jarvis Haupt†, Gil Raz‡ and Robert Nowak†
†Department of Electrical and Computer Engineering, University of Wisconsin-Madison
‡GMR Research and Technology, Concord, MA 01742-3819

***Abstract*—Reliable wireless communications often requires accurate knowledge of the underlying multipath channel. This typically involves probing of the channel with a known training waveform and linear processing of the input probe and channel output to estimate the impulse response. Many real-world channels of practical interest tend to exhibit impulse responses characterized by a relatively small number of nonzero channel coefficients. Conventional linear channel estimation strategies, such as the least squares, are ill-suited to fully exploiting the inherent low-dimensionality of these sparse channels. In contrast, this paper proposes sparse channel estimation methods based on convex/linear programming. Quantitative error bounds for the proposed schemes are derived by adapting recent advances from the theory of compressed sensing. The bounds come within a logarithmic factor of the performance of an ideal channel estimator and reveal significant advantages of the proposed methods over the conventional channel estimation schemes.**

I. Channel Sensing and Estimation

Optimal demodulation and decoding in wireless communication systems often requires accurate knowledge of channel impulse response functions. Typically, an impulse response is estimated by probing the channel with known training waveforms and comparing the transmitted and received signals. The importance of channel identification is underscored by the number of research papers dedicated to this topic. Searching for "channel estimation" turns up more than 36 000 articles in Google Scholar and over 700 IEEE journal articles in IEEE Xplore.

There are two salient aspects to this channel identification problem, namely, *sensing* and *estimation*. Sensing corresponds to the design of waveforms used to probe the channel. Estimation is the problem of processing the input probe and channel output to recover the impulse response. The ability to accurately identify a channel critically depends on both the design of appropriate probes and the application of effective estimation methods. In particular, probing schemes and estimation strategies that are tailored to the anticipated characteristics of the underlying channel yield better estimates than generic procedures. Grappling with these issues is central to most of the papers written on this topic.

Multipath wireless channels tend to exhibit impulse responses dominated by a relatively small number of clusters of significant paths, especially when operating at large bandwidths and signaling durations and/or with numbers of antenna elements [1]–[3]. These are often called "sparse" channels, since most of the channel coefficients are either zero or nearly zero. In the simplest setting, which is the primary focus of this paper, consider frequency-selective channels (channels with a large delay spread relative to the inverse of the communication bandwidth) with most of the impulse response energy localized to relatively small regions in delay. Sparse channel models of this type have received considerable attention lately [4]–[6].

This paper tackles the problem of sensing and estimating sparse (single-antenna) frequency-selective channels. A number of authors have recently addressed this problem [4]–[6]. Lacking from the previous investigations is a quantitative theoretical analysis of the performance of the proposed sparse channel estimation methods. The main results of this paper adapt recent advances from the theory of compressed sensing to devise quantitative error bounds for convex/linear programming based sparse channel estimation schemes. The bounds come within a logarithmic factor of the performance of an ideal channel estimator and clearly reveal the relationship between the input probes and the accuracy of the channel estimates.

II. Multipath Wireless Channel Modeling

In this section, we review mathematical models for multipath wireless channels. In particular, the so-called "virtual channel model" is a discrete approximation of the continuous-time channel at the delay resolution dictated by the channel bandwidth. The virtual channel model captures the relationship between clustering of physical paths and sparsity of dominant channel coefficients and sets the stage for the application of compressed sensing theory and methods.

We consider single-antenna communication channels, which are often characterized as linear, time-varying systems [7]. The corresponding (complex) basedband transmitted and received signals are related as

$$y(t) = \int_0^{T_m} h(t,\tau) x(t-\tau) \mathrm{d}\tau + z(t) \tag{1}$$

where $h(t,\tau)$ is the time-varying channel impulse response, and $x(t)$, $y(t)$ and $z(t)$ represent the transmitted, received and additive white Gaussian noise (AWGN) waveforms, respectively. The maximum delay spread T_m is defined as the maximum possible nonzero delay. The focus of this paper is the identification of frequency-selective channels. A channel is said to be (purely) frequency-selective if (i) the channel impulse response remains constant over the time duration of interest, i.e., $h(t,\tau) \approx h(0,\tau) = h(\tau)$; and (ii) the

This research was supported in part by the DARPA A2I Program. E-mails: {wubajwa, jdhaupt}@wisc.edu, raz@gmrtech.com, nowak@engr.wisc.edu

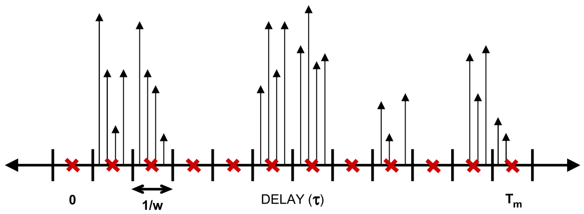

Fig. 1. A schematic illustrating the virtual representation of a single-antenna, frequency-selective channel. The virtual channel coefficients $\{\beta_j\}$ correspond to uniformly-spaced samples of a smoothed version of the channel impulse response taken at the virtual delays $\{\hat{\tau}_j = j/W\}$ in the delay space.

channel delay spread is large relative to the inverse of the communication bandwidth W, i.e., $T_m W \geq 1$.

A. Virtual Channel Modeling

Frequency-selective channels generate multiple delayed and attenuated copies of the transmitted waveform. For such "multipath" channels, $h(\tau)$ is modeled as

$$h(\tau) = \sum_{i=1}^{N_{\text{path}}} \alpha_i \delta(\tau - \tau_i) \tag{2}$$

and the transmitted and received waveforms are related by

$$y(t) = \sum_{i=1}^{N_{\text{path}}} \alpha_i x(t - \tau_i) + z(t) \tag{3}$$

which corresponds to signal propagation along N_{path} physical paths, where $\alpha_i \in \mathbb{C}$ and $\tau_i \in [0, T_m]$ are the complex path gain and the delay associated with the i-th physical path, respectively.

The discrete path model (2), while realistic, is difficult to analyze and identify due to nonlinear dependence on the real-valued delay parameters $\{\tau_i\}$. However, because the communication bandwidth W is limited, the continuous-time channel can be accurately approximated by a discrete counterpart, known as a virtual channel model, with the aid of sampling theorems and/or power series expansions—see, e.g., [7], [8]. The key idea behind virtual channel modeling is to provide a discrete approximation of frequency-selective channels by uniformly sampling the physical multipath environment in delay at a resolution commensurate with W, i.e.,

$$y(t) \approx \sum_{j=0}^{p-1} \beta_j x(t - j/W) + z(t) \tag{4}$$

$$\beta_j \approx \sum_{i \in \mathcal{S}_{\tau,j}} \alpha_i \operatorname{sinc}(j - W\tau_i) \tag{5}$$

where $p = \lceil T_m W \rceil + 1$, $\operatorname{sinc}(a) = \sin(\pi a)/\pi a$ and $\mathcal{S}_{\tau,j} = \{i : \tau_i \in [j/W - 1/2W, j/W + 1/2W)\}$ denotes the set of all physical paths whose delays lie within the delay resolution bin of width $\Delta\tau = 1/W$ centered around the j-th resolvable *virtual delay*, $\hat{\tau}_j = j/W$. Henceforth, $\{\beta_j\}$ are termed as the virtual channel coefficients in the delay space. The expression (5) states that the coefficient β_j approximately consists of the sum of gains of all paths whose delays lie within the j-th delay resolution bin, as illustrated in Fig. 1. Note that this approximation gets more accurate with increasing W, due to higher delay resolution.

B. Discrete-Time Representation

The continuous-time received signal $y(t)$ can be sampled at a rate of $1/W$ at the receiver to obtain an equivalent discrete-time representation of (4)

$$\boldsymbol{y} = \boldsymbol{x} * \boldsymbol{\beta} + \boldsymbol{z} \qquad \Longrightarrow \qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} \tag{6}$$

where $*$ denotes the discrete convolution operator, $\boldsymbol{\beta} \in \mathbb{C}^p$ is the vector of channel coefficients, and $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are the vectors of samples of $x(t)$, $y(t)$ and $z(t)$, respectively. The dimensions of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are dictated by the input signaling duration T_s, where $x(t) = 0$ for all $t \notin [0, T_s]$. We use n $(= \lceil T_s W \rceil)$ to denote the number of nonzero samples of $x(t)$, resulting in $\boldsymbol{x} \in \mathbb{C}^n$ and $\boldsymbol{y}$, $\boldsymbol{z} \in \mathbb{C}^{n+p-1}$. Finally, the matrix $\boldsymbol{X}$ is an $(n+p-1) \times p$ (Toeplitz-structured) convolution matrix formed from vector $\boldsymbol{x}$.

In the following, we shall model $\boldsymbol{\beta}$ as a deterministic but unknown vector. In practical communication systems, there is also a power constraint on the transmitted signal $x(t)$ that can be readily translated into an average power constraint on the entries of $\boldsymbol{x}$. Without loss of generality, we assume that $\max_k \mathbb{E}[|x_k|^2] \leq 1$ and the entries of $\boldsymbol{z}$ correspond to an independent and identically distributed (i.i.d.) complex Gaussian white noise sequence. We denote their distribution as $\mathcal{CN}(0, \sigma^2)$, where $\sigma^2 > 0$ is the noise power.

The virtual representation of a frequency-selective channel captures its essential characteristics in terms of the channel coefficients $\{\beta_j\}$. Identifying a frequency-selective channel, therefore, becomes equivalent to designing the (discrete) input probe $\boldsymbol{x}$ and estimating $\boldsymbol{\beta}$ from the output $\boldsymbol{y}$. Three types of input probes are commonly employed for channel sensing, namely, impulses, pseudo-random inputs, and multitone signals. Channel estimates are usually obtained by solving the least squares (LS) problem (or a variant of it)

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^H \boldsymbol{X})^{-1} \boldsymbol{X}^H \boldsymbol{y}. \tag{7}$$

It is easy to show that the mean squared error (MSE) of an LS channel estimator obeys

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] = trace((\boldsymbol{X}^H \boldsymbol{X})^{-1})\sigma^2. \tag{8}$$

III. Sparse Multipath Channels

Channel measurement results dating as far back as 1987 [1] and as recent as 2007 [9] suggest that multipath components tend to be distributed in clusters rather than uniformly over the channel delay spread. These clusters of paths physically correspond to large-scale objects in the scattering environment (e.g., buildings and hills in an outdoor propagation environment), while multipath components within a cluster arise as a result of scattering from small-scale structures of the corresponding large-scale reflector (e.g., windows of a building, trees on a hill).

Based on the interarrival times between different multipath clusters within the delay spread, wireless channels can be categorized as either “rich” or “sparse”. In a rich multipath channel, the interarrival times are smaller than the inverse of the communication bandwidth W. Sparse multipath channels, on the other hand, exhibit interarrival times that are larger than $1/W$. Therefore, similar to the setting in Fig. 1, not every delay bin of width $\Delta\tau = 1/W$ contains a multipath component in this case. In particular, since a channel coefficient consists of the sum of gains of all paths falling within its respective delay bin, sparse frequency-selective channels tend to have far fewer nonzero channel coefficients than p at any fixed (but large enough) bandwidth. We formalize this notion of multipath sparsity as follows.

Definition 1 (Sparse Multipath Channels): Let W be the operating bandwidth and $p = \lceil T_m W \rceil + 1$ be the number of resolvable paths (channel coefficients) within the channel delay spread. We say that a multipath channel is S-sparse if $\|\boldsymbol{\beta}\|_{\ell_0} = S < p$, where $\| \cdot \|_{\ell_0}$ counts the number of nonzero entries in a vector.

Many real-world channels of practical interest, such as underwater acoustic channels [10], digital television channels [11] and residential ultrawideband channels [2], in fact, tend to be sparse or approximately sparse, with $S \ll p$. However, conventional LS based channel estimation schemes, while appropriate for rich channels, fail to capitalize on the anticipated sparsity of the abovementioned channels. To get an idea of the potential MSE gains to be had by incorporating the sparsity assumption into the channel estimation strategy, we compare the performance of an LS based channel estimator to that of a channel estimation strategy that has been equipped with an *oracle*. The oracle does not reveal the true $\boldsymbol{\beta}$, but does inform us of the indices of nonzero entries of $\boldsymbol{\beta}$. Clearly this represents an ideal estimation strategy and one cannot expect to attain its performance level. Nevertheless, it is the target that one should consider.

We begin with an application of arithmetic-harmonic means inequality to the MSE expression in (8) and note that

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq \frac{p^2\sigma^2}{trace(\boldsymbol{X}^H\boldsymbol{X})} = \frac{p\,\sigma^2}{\|\boldsymbol{x}\|_{\ell_2}^2} \tag{9}$$

with equality if and only if $\boldsymbol{X}^H\boldsymbol{X} = \|\boldsymbol{x}\|_{\ell_2}^2 \boldsymbol{I}_p$. Now let $T_\star \subset \{1,\dots,p\}$ be the set of indices of nonzero entries of $\boldsymbol{\beta}$ and suppose that an oracle provides us with $T_\star$. Then an ideal channel estimator $\boldsymbol{\beta}^\star$ can be obtained from $\boldsymbol{y}$ by first forming a *restricted* LS estimator

$$\boldsymbol{\beta}_{T_\star} = (\boldsymbol{X}_{T_\star}^H \boldsymbol{X}_{T_\star})^{-1}\boldsymbol{X}_{T_\star}^H \boldsymbol{y} \tag{10}$$

where $\boldsymbol{X}_{T_\star}$ is a submatrix obtained by extracting the S columns of $\boldsymbol{X}$ corresponding to the indices in $T_\star$, and then setting $\boldsymbol{\beta}^\star$ to $\boldsymbol{\beta}_{T_\star}$ on the indices in $T_\star$ and zero on the indices in $T_\star^c$. Clearly, the MSE of this oracle channel estimator obeys

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}\|_{\ell_2}^2\right] &= trace((\boldsymbol{X}_{T_\star}^H \boldsymbol{X}_{T_\star})^{-1})\sigma^2 \\ &\geq \frac{S^2\sigma^2}{trace(\boldsymbol{X}_{T_\star}^H \boldsymbol{X}_{T_\star})} = \frac{S\,\sigma^2}{\|\boldsymbol{x}\|_{\ell_2}^2} \end{aligned}\tag{11}$$

with equality if and only if $\boldsymbol{X}_{T_\star}^H \boldsymbol{X}_{T_\star} = \|\boldsymbol{x}\|_{\ell_2}^2 \boldsymbol{I}_S$. Comparing the MSE lower bounds (9) and (11) shows that LS based conventional channel estimates may be at a significant disadvantage when it comes to identifying sparse channels. And while the oracle channel estimator $\boldsymbol{\beta}^\star$ is impossible to construct in practice, results from the theories of adaptive denoising and compressed sensing can be readily adapted to construct channel estimates using impulse probes and discrete multitone probes, respectively, which come within a logarithmic factor of the ideal MSE in (11). Below, we briefly summarize these results which are essentially direct applications of existing theory and methods.

A. Impulse Probing

The channel output corresponding to an impulse probe, $x_k = \delta_k$, is given by $\boldsymbol{y} = \boldsymbol{\beta} + \boldsymbol{z}$. Trivially, $\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{y}$ and $\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] = p\,\sigma^2$. On the other hand, let $\lambda(p,\sigma) = \sqrt{2\log p}\cdot\sigma$ and define a hard-threshold estimator $\widehat{\boldsymbol{\beta}}_\lambda$ by setting

$$\widehat{\beta}_{\lambda,i} = y_i 1_{\{|y_i|>\lambda\}}, \quad i = 1,\dots,p. \tag{12}$$

It is well-known that the MSE of this thresholded estimator obeys the following upper bound [12]:

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_{\ell_2}^2\right] \leq const\cdot\log p\cdot S\,\sigma^2 \tag{13}$$

which comes within a factor of $\log p$ to the ideal (oracle based) MSE of $S\,\sigma^2$ (note that $\|\boldsymbol{x}\|_{\ell_2}^2 = 1$ in this case), and shows an MSE improvement by a factor of (roughly) $O(p/S)$ over the LS MSE of $p\,\sigma^2$.

B. Discrete Multitone Probing

While the thresholded estimator $\widehat{\boldsymbol{\beta}}_\lambda$ performs significantly better than the LS estimator, power constraint on the entries of an input probe makes impulse probing highly unattractive for channel estimation purposes (since $\|\boldsymbol{x}\|_{\ell_2}^2 = 1$), except in very high signal-to-noise ratio (SNR) scenarios ($\sigma^2 \lll 1$). Instead, an attractive alternative is to employ discrete multitone probes of the form

$$x_k = \frac{1}{\sqrt{M}}\sum_{m=1}^{M} t_m e^{j\omega_m k}, \quad k = 0,\dots,n-1$$

where the input probe duration $n \geq p$, the number of tones $M \leq p$, the frequencies $\{\omega_m\}$ are randomly selected from $\left\{2\pi\frac{\ell}{n} : \ell = 0,\dots,n-1\right\}$, and the amplitudes $\{t_m\}$ are i.i.d. binary random variables taking the values ±1 with probability $1/2$ each. Stated in the language of orthogonal frequency division multiplexing (OFDM) communication systems, this is equivalent to assigning M OFDM tones (out of a possible n) as pilot tones.

It is straightforward to check that $\|\boldsymbol{x}\|_{\ell_2}^2 = n$ in this case and hence, from (9), $\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq p\,\sigma^2/n$. Recent results from compressed sensing, however, show that one can improve on the LS estimator using sparse estimation methods based on convex/linear programming. The crucial observation here is that a multitone probe essentially samples the frequency response of a multipath channel at M locations. This can be easily seen by transforming the time domain channel estimation problem into the frequency domain through an n point discrete Fourier transform of the channel output $\boldsymbol{y}$, which results in

$$Y(\omega_m) = \sqrt{n}\, t_m \beta(\omega_m) + Z(\omega_m), \quad m = 1, \dots, M. \quad (14)$$

Now suppose that the number of frequency samples $M \geq const \cdot (\log p)^4 \cdot S$. Then knowledge of the channel frequency response at M frequencies is sufficient to construct an estimator $\widehat{\boldsymbol{\beta}}_{CS}$ of the sparse channel by solving either an ℓ_1 penalized LS problem (the "lasso" estimator [13]) or a convex program called the "Dantzig selector" [14]; see (16). Further, it has been shown that, with high probability, the resulting estimator obeys the following upper bound [14], [15]:

$$\|\widehat{\boldsymbol{\beta}}_{CS} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq const \cdot \log p \cdot \left(\frac{S\,\sigma^2}{n}\right) \quad (15)$$

which is within a factor of $\log p$ to the oracle estimator's lower bound of $S\,\sigma^2/n$. Note that the key to understanding this result is the well-known time-frequency duality: signals concentrated in the time domain are spread out in the frequency domain and vice versa. Sampling an S-sparse multipath channel at roughly $O(S)$ locations in the frequency domain, therefore, suffices to capture its salient information.

IV. Compressed Channel Sensing

A pseudo-random sequence $\{x_k\}_{k=0}^{n-1}$ is another form of input that is generally used to probe a channel, where n is the input probe duration and x_k's are i.i.d. realizations from a zero mean, unit variance distribution $f(x)$. For the sake of this exposition, we limit ourselves to $f(x)$ being the Rademacher distribution, i.e., x_k's take values $+1$ or -1 with probability $1/2$ each, but extending the main results of the paper to other distributions is straightforward. The energy in the input probe in this case is again given by $\|\boldsymbol{x}\|_{\ell_2}^2 = n$, resulting in the LS lower bound of $\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|_{\ell_2}^2\right] \geq p\,\sigma^2/n$.

We now show that it is possible to obtain a more reliable estimator of $\boldsymbol{\beta}$ as a solution to the convex program

$$\widehat{\boldsymbol{\beta}} = \underset{\tilde{\boldsymbol{\beta}} \in \mathbb{C}^p}{\arg\min} \|\tilde{\boldsymbol{\beta}}\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{X}^H \boldsymbol{r}\|_{\ell_\infty} \leq \lambda \quad (16)$$

where $\lambda(p,\sigma) > 0$ and $\boldsymbol{r}$ is the $(n+p-1)$-dimensional vector of residuals: $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$. This optimization program goes by the name of Dantzig selector (DS) and is computationally tractable since it can be recast as a linear program [14]. We state our main results in terms of the DS primarily because it provides the cleanest and most interpretable error bounds that we know. Note, however, that similar bounds also hold for the lasso estimator [15] which can sometimes be more computationally attractive because of the availability of a wide array of efficient software packages for solving it [16]. The key to proving the efficacy of the DS is in showing that the Toeplitz matrix $\boldsymbol{X}$ generated by the pseudo-random input probe $\boldsymbol{x}$ satisfies the so-called "restricted isometry property" (RIP) with sufficiently small value of $3S$-restricted isometry constant.

Definition 2 (Restricted Isometry Constant): Suppose that the columns of $\boldsymbol{X}$ are normalized to unit ℓ_2 norm. The $3S$-restricted isometry constant of $\boldsymbol{X}$, denoted by δ_{3S}, is defined as the smallest value such that

$$(1 - \delta_{3S})\|\tilde{\boldsymbol{\beta}}\|_{\ell_2}^2 \leq \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|_{\ell_2}^2 \leq (1 + \delta_{3S})\|\tilde{\boldsymbol{\beta}}\|_{\ell_2}^2 \quad (17)$$

holds for all $3S$-sparse vectors $\tilde{\boldsymbol{\beta}} \in \mathbb{C}^p$. The matrix $\boldsymbol{X}$ is said to satisfy RIP of order $3S$ if $\delta_{3S} \in [0, 1)$.

Note that if any two columns of $\boldsymbol{X}$ happen to be linearly dependent then $\delta_{3S} \geq 1$. Loosely speaking, the RIP essentially requires that mutual coherence between the columns of $\boldsymbol{X}$ is sufficiently small so that $\boldsymbol{X}^H\boldsymbol{X}$ (approximately) behaves like a multiple of identity on the set of sparse vectors. The main result in [14] asserts that the DS solution is highly accurate in this case.

Theorem 1 ([14]): Suppose that the column-normalized version of $\boldsymbol{X}$ satisfies RIP of order $3S$ with $\delta_{3S} < 1/2$. Choose $\lambda(p,\sigma) = \sqrt{2(1+a)\log p} \cdot \sigma$ for any $a \geq 0$. Then, with probability exceeding $1 - (\sqrt{\pi \log p} \cdot p^a)^{-1}$, the DS estimator $\widehat{\boldsymbol{\beta}}$ obeys

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2 \leq const \cdot \log p \cdot \left(\frac{S\,\sigma^2}{n}\right). \quad (18)$$

Theorem 1 states that the DS based channel estimator can *potentially* achieve squared error within a factor of $\log p$ of the oracle based MSE lower bound of $S\,\sigma^2/n$. However, it remains to be seen whether the convolution matrix $\boldsymbol{X}$ formed from vector $\boldsymbol{x}$ satisfies the conditions of this theorem. The main thesis of this paper is that this is indeed the case.

Theorem 2: Let $\{x_k\}_{k=0}^{n-1}$ be a sequence of i.i.d. random variables drawn from the Rademacher distribution and $\boldsymbol{X}$ be the $(n+p-1) \times p$ Toeplitz matrix generated by this sequence (as described in Section II-B). Suppose that the duration of the input probe $n \geq c_1 \cdot \log p \cdot S^2$. Then, with probability exceeding $1 - \exp(-c_2 \cdot n)$, the column-normalized version of $\boldsymbol{X}$ satisfies RIP of order $3S$ with $\delta_{3S} < 1/2$. Here, c_1 and c_2 are constants that do not depend on n or p.

Theorem 2 is, in fact, a strengthened version of the results first reported in [17] for Toeplitz-structured matrices. A detailed proof of the theorem, which leverages a few key ideas from [17], is given in the Appendix. Theorem 2, combined with Theorem 1, shows that the DS estimator $\widehat{\boldsymbol{\beta}}$ corresponding to a pseudo-random input probe does remarkably better than the LS estimator in learning an S-sparse channel: assuming that the input probe duration $n \geq const \cdot \log p \cdot S^2$, the improvement is roughly by a factor of $O(p/S)$.

The appeal of the DS 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 [2]. One such scenario could be, for example, that the magnitudes of the ordered channel coefficients exhibit a power law decay, i.e., the ℓ-th largest coefficient obeys $|\beta_{(\ell)}| \leq R \cdot \ell^{-1/q}$ for some $R > 0$ and $q \leq 1$. Define $S = |\{j : |\beta_j| > \sigma\}|$. Then, using a

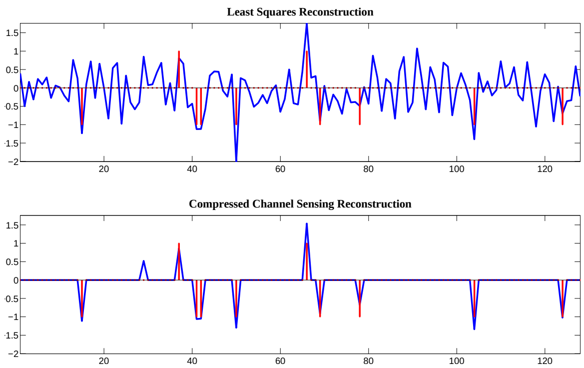

Fig. 2. An illustrative example contrasting the sparse reconstruction abilities of LS and lasso ($n = p = 128$, $S = 10$, SNR $= -10$ dB).

pseudo-random input probe with $n \geq const \cdot \log p \cdot S^2$, the DS estimator achieves the minimax rate over the class of objects exhibiting the power law decay [14, Th. 1.3].

V. Numerical Results and Discussion

Convex programming based channel estimators, such as the DS (or lasso) estimator, are inherently tuned to yield sparse solutions. This is in stark contrast to linear channel estimators, such as the LS estimator, wherein each coefficient of the channel estimate will be nonzero in general. To this end, an example contrasting the sparse reconstruction abilities of LS and lasso is illustrated in Fig. 2. The setup corresponds to using an $n = 128$ length pseudo-random input probe to sense a $p = 128$ length channel that has only $S = 10$ nonzero channel coefficients. The output of the channel is observed at an SNR of -10 dB (SNR $= 10\log_{10}(1/\sigma^2)$), and LS and lasso estimates are obtained by pseudo-inverting $\boldsymbol{X}$ and executing GPSR [16], respectively. As can be seen from Fig. 2, the lasso estimate is able to identify every nonzero channel coefficient (shown as an impulse), as well as reject all but one of the noise-only coefficients. In contrast, it is easy to see that even a clairvoyantly thresholded LS estimate would be unable to identify all the nonzero channel coefficients in this case.

We conclude this paper by noting that with the advent of modern-day wireless transceivers capable of communicating at large spatio-spectral-temporal dimensions, multipath sparsity is becoming more and more pronounced across a broad range of communication systems. As such, we expect similar performance gains by applying sparse estimation strategies to these systems and propose to extend the results of this and related works [18], [19] to estimating sparse channels in time, frequency, and space.

Appendix

Proof of Theorem 2: Suppose that the columns of $\boldsymbol{X}$ are normalized to unit ℓ_2 norm. For ease of notation, we index the generative sequence of $\boldsymbol{X}$ as $\{x_k\}_{k=1}^{n}$, where x_k's are now i.i.d. binary random variables taking the values $\pm 1/\sqrt{n}$ with probability $1/2$ each. Define $\delta = \delta_{3S} \in (0, 1/2)$ and $m = 3S$.

Let $T \subset \{1, 2, \ldots, p\}$ be a subset of indices of cardinality $|T|$, and let $\boldsymbol{X}_T$ be the $(n + p - 1) \times |T|$ submatrix of $\boldsymbol{X}$ formed by retaining the columns indexed by the entries of T. The RIP condition essentially requires that, for all subsets T satisfying $|T| = m$, the eigenvalues of the Gram matrix $\boldsymbol{G}(T) = \boldsymbol{X}_T'\boldsymbol{X}_T$ lie in the interval $[1 - \delta, 1 + \delta]$. For a *fixed* subset T, this condition can be established using Gershgorin's circle theorem, which states that the eigenvalues of an $m \times m$ matrix $\boldsymbol{G}$ all lie in the union of m discs, where the i-th disc is centered at the diagonal entry $G_{i,i}$ and has radius

$$R(i) = \sum_{j=1, j\neq i}^{m} |G_{i,j}|. \tag{19}$$

Notice that by choice of the x_k's, $G_{i,i}(T) = 1$ deterministically. Thus, to establish that the eigenvalues lie in $[1-\delta, 1+\delta]$ for a fixed T, it is sufficient to show that the off-diagonal entries of $\boldsymbol{G}(T)$ are all less than δ/m in absolute value, since this would imply $R(i) \leq (m-1)(\delta/m) < \delta$ for all i.

To guarantee the RIP condition for $\boldsymbol{X}$, however, the eigenvalue bounds must hold for all subsets T that satisfy $|T| = m$. To this end, we consider the full $p \times p$ Gram matrix of $\boldsymbol{X}$, $\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{X}$, and show that the off-diagonal entries of $\boldsymbol{G}$ are all bounded above by δ/m in absolute value. The implication is that, since the Gram matrix $\boldsymbol{G}(T)$ corresponding to *any* subset T satisfying $|T| = m$ is itself a submatrix of $\boldsymbol{G}$, $\boldsymbol{G}(T)$ has bounded off-diagonals and, therefore, the eigenvalues of all $\binom{p}{m}$ Gram matrices $\boldsymbol{G}(T)$ lie in $[1 - \delta, 1 + \delta]$.

To proceed, notice that each off-diagonal term of $\boldsymbol{G}$ is simply the inner product between i-th and j-th column of $\boldsymbol{X}$, and thus $G_{i,j} = G_{j,i}$. We can write an expression for the off-diagonal element $G_{i,j}$, assuming $p \geq j > i \geq 1$, as

$$G_{i,j} = \sum_{k=1}^{n} x_k x_{k+(j-i)} \ , \tag{20}$$

where we define $x_\ell = 0$ for $\ell > n$. Standard concentration inequalities are not directly applicable here because all of the entries in the sum are not mutually independent. For example, consider $i = 1$, $j = 2$, and $n = 5$. Then $G_{1,2} = x_1x_2 + x_2x_3 + x_3x_4 + x_4x_5$ and the first two terms are dependent (through x_2), as are the second and third (through x_3), etc. But notice that the first and third terms are independent as are the second and fourth, suggesting that each sum may be split into two sums of i.i.d. random variables.

This is true in general, and to establish the claim we leverage a result from the theory of graph coloring. A proper coloring of a graph is a coloring (or labeling) of vertices so that two vertices have different colors (labels) if they are connected by an edge; a proper coloring forms a partition of the vertices into color classes. An equitable coloring is a proper coloring where the difference in size between the smallest and largest color classes is at most one.

We begin by associating a graph to each sum $G_{i,j}$, where each term in the sum corresponds to a vertex in the graph. Two vertices are connected by an edge if the corresponding terms in the sum are statistically dependent. Notice that in this setting, the maximum number of edges originating from any vertex (the degree of the graph) is one, since each term is dependent with at most one other term.

The problem of splitting the original sum into sums of independent terms is equivalent to coloring this graph. To obtain sums of approximately the same size we consider equitable coloring, and we leverage a result of Hajnal and Szemerédi which states that a graph with degree Δ can be equitably colored with $\Delta + 1$ colors [20]. Here, this result guarantees that two colors are sufficient.

To proceed, we rewrite the expression in (20) to explicitly show the number of nonzero terms in the sum, and obtain

$$G_{i,j} = \left\{ \begin{array}{ll} 0 & n < p \text{ and } j - i \geq n \\ \sum_{k=1}^{n-(j-i)} g_k & \text{otherwise} \end{array} \right\}, \quad (21)$$

where $\{g_k\}$ are (dependent) random variables taking values $\pm 1/n$ each with probability $1/2$. Because we are interested in obtaining absolute upper bounds on $G_{i,j}$, only the nonzero situation requires further analysis. Partition $\{g_k\}$ into two mutually independent subsets according to the equitable coloring, and appropriately reindex the terms to obtain

$$G_{i,j} = \sum_{k_1=1}^{q_1 = \frac{n-(j-i)}{2}} g'_{k_1} + \sum_{k_2=1}^{q_2 = \frac{n-(j-i)}{2}} g'_{k_2} \quad (22)$$

when $n - (j - i)$ is even, and

$$G_{i,j} = \sum_{k_1=1}^{q_1 = \frac{n-(j-i)+1}{2}} g'_{k_1} + \sum_{k_2=1}^{q_2 = \frac{n-(j-i)-1}{2}} g'_{k_2} \quad (23)$$

when $n - (j - i)$ is odd. Generically, we write $G_{i,j} = G^1_{i,j} + G^2_{i,j}$. We analyze each component sum using Hoeffding's (two-sided) inequality for bounded random variables to obtain, for example,

$$\Pr\left(|G^1_{i,j}| > \epsilon\right) \leq 2 \exp\left(\frac{-\epsilon^2 n^2}{2q_1}\right), \quad (24)$$

and choosing $\epsilon = \delta/2m$ yields

$$\Pr\left(|G^1_{i,j}| > \delta/2m\right) \leq 2 \exp\left(\frac{-\delta^2 n^2}{8q_1 m^2}\right). \quad (25)$$

Considering both sums, we can write

$$\begin{aligned} &\Pr\left(|G_{i,j}| > \delta/m\right) \\ &\leq \Pr\left(\{|G^1_{i,j}| > \delta/2m\} \text{ or } \{|G^2_{i,j}| > \delta/2m\}\right) \\ &\leq 2 \max\left\{\Pr\left(|G^1_{i,j}| > \delta/2m\right), \Pr\left(|G^2_{i,j}| > \delta/2m\right)\right\} \\ &\leq 2 \max\left\{2 \exp\left(\frac{-\delta^2 n^2}{8q_1 m^2}\right), 2 \exp\left(\frac{-\delta^2 n^2}{8q_2 m^2}\right)\right\}. \quad (26) \end{aligned}$$

Notice that smaller values of q_1 and q_2 lead to tighter bounds, and thus the slowest rate of concentration occurs when the number of nonzero terms in $G_{i,j}$ is $n - 1$. In this case, we have $(n-1)/2 \leq q_1 \leq q_2 \leq (n+1)/2$, implying that the worst case sum has at least $(n-1)/2$ terms. As a result

$$\begin{aligned} \Pr\left(|G_{i,j}| > \delta/m\right) &\leq 4 \exp\left(\frac{-\delta^2 n^2}{4(n-1)m^2}\right) \\ &\leq 4 \exp\left(\frac{-\delta^2 n}{4m^2}\right). \quad (27) \end{aligned}$$

To establish RIP we require that *each* of the $p(p-1)/2$ unique off-diagonal terms $G_{i,j}$ satisfy this bound. Applying the union bound yields

$$\begin{aligned} \Pr\left(\text{any } |G_{i,j}| > \delta/m\right) &\leq 2p^2 \exp\left(\frac{-\delta^2 n}{4m^2}\right) \\ &\leq \exp\left(\frac{-\delta^2 n}{4m^2} + 3 \log p\right). (28) \end{aligned}$$

where the last step follows under the mild assumption that $p \geq 2$. Now, notice that whenever $\delta^2 n/4m^2 > 3 \log p$, or $n > \frac{12m^2 \log p}{\delta^2}$, RIP is satisfied with probability at least

$$1 - \exp\left(\frac{-\delta^2 n}{4m^2} + 3 \log p\right). \quad (29)$$

This success probability is nonzero and can be very close to one when n is large compared to m^2. ∎

References

[1] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Select. Areas Commun.*, pp. 128–137, Feb. 1987.

[2] A. F. Molisch, "Ultrawideband propagation channels-Theory, measurement, and modeling," *IEEE Trans. Veh. Technol.*, pp. 1528–1545, Sep. 2005.

[3] N. Czink, X. Yin, H. Ozcelik, M. Herdin, E. Bonek, and B. H. Fleury, "Cluster characteristics in a MIMO indoor propagation environment," *IEEE Trans. Wireless Commun.*, pp. 1465–1475, Apr. 2007.

[4] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, pp. 374–377, Mar. 2002.

[5] C. Carbonelli, S. Vedantam, and U. Mitra, "Sparse channel estimation with zero tap detection," *IEEE Trans. Wireless Commun.*, pp. 1743–1753, May 2007.

[6] Y. Selen and E. G. Larsson, "RAKE receiver for channels with a sparse impulse response," *IEEE Trans. Wireless Commun.*, pp. 3175–3180, Sep. 2007.

[7] P. Bello, "Characterization of randomly time-variant linear channels," *IEEE Trans. Commun.*, pp. 360–393, Dec. 1963.

[8] A. M. Sayeed and B. Aazhang, "Joint multipath-Doppler diversity in mobile wireless communications," *IEEE Trans. Commun.*, pp. 123–132, Jan. 1999.

[9] L. Vuokko, V.-M. Kolmonen, J. Salo, and P. Vainikainen, "Measurement of large-scale cluster power characteristics for geometric channel models," *IEEE Trans. Antennas Propagat.*, pp. 3361–3365, Nov. 2007.

[10] D. Kilfoyle and A. Baggeroer, "The state of the art in underwater acoustic telemetry," *IEEE J. Oceanic Eng.*, pp. 4–27, Jan. 2000.

[11] *Receiver Performance Guidelines*, ATSC Recommended Practices for Digital Television, 2004.

[12] D. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, pp. 425–455, 1994.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, pp. 267–288, 1996.

[14] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, pp. 2313–2351, Dec. 2007.

[15] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," pre-print.

[16] M. A. T. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Select. Topics Signal Processing*, pp. 586–597, Dec. 2007.

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

[18] G. E. Pfander, H. Rauhut, and J. Tanner, "Identification of matrices having a sparse representation," pre-print.

[19] M. Herman and T. Strohmer, "High resolution radar via compressed sensing," pre-print.

[20] A. Hajnal and E. Szemerédi, "Proof of a conjecture of P. Erdős," in *Combinatorial Theory and its Application*, 1970, pp. 601–623.