Learning Sparse Doubly-Selective Channels

Waheed U. Bajwa, Akbar M. Sayeed, and Robert Nowak
bajwa@cae.wisc.edu, akbar@engr.wisc.edu, nowak@engr.wisc.edu
Department of Electrical and Computer Engineering, University of Wisconsin-Madison

***Abstract*—Coherent data communication over doubly-selective channels requires that the channel response be known at the receiver. Training-based schemes, which involve probing of the channel with known signaling waveforms and processing of the corresponding channel output to estimate the channel parameters, are commonly employed to learn the channel response in practice. Conventional training-based methods, often comprising of linear least squares channel estimators, are known to be optimal under the assumption of rich multipath channels. Numerous measurement campaigns have shown, however, that physical multipath channels tend to exhibit a sparse structure at high signal space dimension (time-bandwidth product), and can be characterized with significantly fewer parameters compared to the maximum number dictated by the delay-Doppler spread of the channel. In this paper, it is established that traditional training-based channel learning techniques are ill-suited to fully exploiting the inherent low-dimensionality of sparse channels. In contrast, key ideas from the emerging theory of compressed sensing are leveraged to propose sparse channel learning methods for both single-carrier and multicarrier probing waveforms that employ reconstruction algorithms based on convex/linear programming. In particular, it is shown that the performance of the proposed schemes come within a logarithmic factor of that of an ideal channel estimator, leading to significant reductions in the training energy and the loss in spectral efficiency associated with conventional training-based methods.**

I. Introduction

Several coherent communication techniques have been developed in the last decade or so to maximally exploit the effects of time- and frequency-selectivity of doubly-selective channels—see, e.g., [1]–[4]. In particular, doubly-selective channels can offer large joint multipath-Doppler diversity gains when perfect channel state information (CSI) is available at the receiver [2], [3]. In many practical scenarios, however, the receiver has seldom access to the CSI and the channel needs to be learned either implicitly or explicitly to reap the benefits of coherent demodulation and decoding.

Two classes of methods are commonly employed to learn a channel at the receiver. In *training-based channel learning* methods, the transmitter multiplexes training signals that are known to the receiver with information bearing signals in time, frequency and/or code domain and CSI is obtained at the receiver from knowledge of the training and received signals. In *blind channel learning* methods, CSI is acquired at the receiver by making use of the statistics of information bearing signals only. Although theoretically efficient, blind learning methods typically require complex signal processing at the receiver and often entail inversion of large data-dependent matrices, which also makes them highly prone to error propagation in rapidly-varying channels. Training-based methods, on the other hand, require relatively simple processing at the receiver and lead to decoupling of the data detection module from the channel learning module at the receiver, which helps to reduce the receiver complexity even further. As such, despite the fact that training-based methods are known to be suboptimal from the spectral efficiency viewpoint, they are widely prevalent in modern communication systems [5].

One of the first analytical studies of training-based channel learning methods was authored by Cavers [6], who coined the term *pilot symbol assisted modulation* for these methods. Since then, there has been a continued interest in the design and analysis of training-based methods for various classes of channels; we refer the reader to [5] for a tutorial overview of related work. These works often highlight two salient aspects of training-based channel learning methods, namely, *sensing* and *estimation*. Sensing corresponds to the design of signaling waveforms (training signals) used to probe the channel and their placement within the transceiver signal space. Estimation is the problem of processing the corresponding channel output at the receiver to recover the channel response. The ability of a training-based method to accurately learn the channel response depends critically on both the design/placement of appropriate training signals and the application of effective estimation methods. In particular, training waveforms 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.

This paper presents a new approach for learning (single-antenna) doubly-selective channels through training-based methods. A number of authors have recently addressed this problem—see, e.g., [7]–[9]. The analysis carried out in these and similarly related works, however, is often based on the assumption of a *rich* underlying multipath environment in the sense that the number of degrees of freedom (DoF) in the channel scale linearly with the signal space dimension (product of signaling duration and bandwidth) [10]. In contrast, physical wireless channels encountered in practice tend to exhibit impulse responses dominated by a relatively small number of dominant resolvable paths, especially when operating at large bandwidths and signaling durations and/or with numbers of antenna elements [11]–[13]. These are often called “sparse” channels, since majority of the DoF in the channel are either zero or nearly zero. The primary focus of this paper is on learning *sparse* doubly-selective channels—channels with most of the multipath energy localized to relatively small regions within the delay-Doppler spread. Sparse channel mod-

els of this type have received considerable attention lately, both from a communication-theoretic perspective [14] and a channel learning perspective [15]–[17]. In the context of channel learning, the previous investigations [15]–[17] lack a quantitative theoretical analysis of the performance of the proposed sparse channel learning methods in terms of the mean squared error (MSE). In contrast, the main results of this paper adapt recent advances from the theory of compressed sensing to devise quantitative error bounds for single-carrier and multicarrier training waveforms and convex/linear programming based estimation schemes. The bounds come within a logarithmic factor of the performance of an ideal channel estimator and clearly reveal the relationship between the training signals and the accuracy of the channel estimates.

A. Relationship to Previous Work

In the channel learning context, the work in this paper is closely related to some of the earlier works by Cotter and Rao [15], Li and Preisig [16], and Tauböck and Hlawatsch [17]. Similar to the main results of this paper, the channel learning techniques proposed in [15]–[17] have been inspired by the literature on sparse signal representations, more commonly studied under the rubric of compressed sensing these days [18]. Both [15] and [16] limit themselves to single-carrier signaling and propose variants of the matching pursuit algorithm [19] for estimation purposes. The results, however, are primarily based on simulation and experimental implementations and, as such, fail to provide any theoretical justifications for the use of the proposed training-based methods. The channel learning technique proposed in [15] also suffers from the drawback that it fails to take into account the Doppler sparsity and limits itself to sparsity in the delay domain only.

In [17], Tauböck and Hlawatsch focus on the case of multicarrier signaling and propose the use of an optimization-based estimator that goes by the name of *basis pursuit with inequality constraint* (BPIC) [20], [21]. Although some theoretical guarantees are provided for the proposed technique, the paper lacks a formal MSE analysis. Also, while BPIC is nearly optimal under the adversarial noise model [20], it is known to be strictly suboptimal in the presence of stochastic noise [22]. Finally, the multicarrier training waveforms in [17] are comprised of the elements of an *incomplete* short-time Fourier (STF) basis [1], also referred to as a Gabor basis or a Weyl-Heisenberg basis in the time-frequency analysis literature. Signaling using an incomplete STF basis, however, results in a loss in spectral efficiency of the communication system, which directly translates into a linear decrease in the overall system capacity [4].

In contrast to the aforementioned references, this paper studies both single-carrier and multicarrier signaling for channel sensing purposes. In particular, single-carrier training is carried out in the paper using spread spectrum waveforms multiplexed in the code domain, whereas the multicarrier training waveforms are comprised of the elements of a *complete* orthogonal STF basis, which helps to maximize the spectral efficiency of the system [4]. The main results of the paper are stated in terms of a linear programming based (nonlinear) estimator that goes by the name of *Dantzig selector* [22] and last, but not least, the focus is on providing a formal comparison of the MSE performance of the proposed sparse channel learning techniques with that of more conventional strategies, which often comprise of linear least squares channel estimators.

Finally, with regards to the connections with compressed sensing related literature, some of the analysis carried out in the paper in the context of single-carrier training is related to the recent work of Pfander et al. [23] and Herman and Strohmer [24]. Both [23] and [24] study the problem of identifying matrices that have a sparse representation in the dictionary of time-frequency shift matrices; [23] looks at this problem in an abstract setting, while [24] studies it from a radar perspective. It can be seen from Section III that the use of single-carrier spread spectrum training waveforms, along with appropriate modeling of sparse doubly-selective channels, also reduces the channel learning problem to that of identifying a matrix which has a sparse representation in the dictionary of time-frequency shift matrices. The resulting time-shifts in the paper are linear, however, as opposed to the circular ones considered in [23], [24]. More importantly, though, both [23], which makes use of a BPIC estimator for matrix identification, and [24], which focuses only on the noiseless setting, lack a formal MSE analysis. Further, the emphasis in [23], [24] is on finding the *coherence* [21] of the dictionary of (circular) time-frequency shift matrices, while we focus on showing that the dictionary of time-frequency shift matrices satisfies the *restricted isometry property* [25], which allows for improved recoverability results; we refer the reader to Section III and the proof of Theorem 2 for further details.

Notation: Throughout this paper, the following notation is used. Vectors (matrices) are denoted by bold-faced lower case (upper case) letters and, unless otherwise stated, all the vectors are taken to be column vectors. Scalars are denoted by light-faced letters and constants are denoted by the letter 'c' or some sub/superscripted version of it. $\mathbf{I}$ and $\mathbf{0}$ are used to denote identity matrices and zero vectors of appropriate sizes, respectively. Superscripts $\overline{(\cdot)}$, $(\cdot)'$ and $(\cdot)^\dagger$ are used to denote complex conjugation, transposition and conjugate transposition, respectively. If $\mathbf{A}$ is a $p \times q$ matrix, then $\mathbf{a} = \mathrm{vec}(\mathbf{A})$ is used to denote the $pq \times 1$ vector obtained by stacking columns of $\mathbf{A}$. The inverse and trace of $\mathbf{A}$ are denoted by $\mathbf{A}^{-1}$ and $\mathrm{tr}(\mathbf{A})$, respectively. $\|\mathbf{a}\|_p$ is the usual ℓ_p norm of the vector $\mathbf{a}$, while $\|\mathbf{a}\|_0$ counts the number of nonzero entries in $\mathbf{a}$. Finally, $\otimes$ is used to denote a Kronecker product and $I_{[a,b)}(t)$ is the indicator function of $[a, b)$.

Organization: The rest of this paper is organized as follows. In Section II, a modeling framework for multipath wireless channels is reviewed and the notion of sparse doubly-selective channels is formally described. Section III considers the problem of learning sparse doubly-selective using single-carrier signaling waveforms, while Section IV studies this problem from the multicarrier signaling perspective. Finally, some numerical results and a discussion of the numerical and theoretical results are provided in Section V.

II. Multipath Wireless Channel Modeling

One of the most salient characteristics of wireless channels is signal propagation over multiple spatially distributed paths, which gives rise to a large number of propagation parameters. From a communication-theoretic perspective, however, we are only interested in characterizing the *interaction* between the physical propagation environment and the signal space of wireless transceivers. This interaction, which occurs in the multiple dimensions of time, frequency and space, is known to depend only coarsely on the exact values of the physical parameters and can be accurately described by a significantly smaller number of DoF [2], [26]–[28].

In this section, we review a virtual modeling framework for doubly-selective channels that captures the interaction between the physical paths and the signal space. Physically, each propagation path in a doubly-selective channel can be represented as a distinct point in the delay-Doppler domain. The virtual channel model [2], also sometimes referred to as the canonical channel model [26], constructs a low-dimensional approximation of the underlying multipath environment through uniform sampling of the delay-Doppler domain at a resolution commensurate with the signaling duration and bandwidth. It plays a key role in the subsequent development in this paper since it captures the relationship between the clustering of physical paths within the delay-Doppler domain and sparsity of *effective* DoF in the channel and sets the stage for the application of compressed sensing theory and methods.

A. Doubly-Selective Channels: Physical Channel Model

We consider single-antenna communication channels, which are often characterized as linear, time-varying systems [26]. The corresponding (complex) baseband transmitted and received signals in the absence of noise are related as

$$y(t) = \int_0^{\tau_{max}} h(t,\tau)x(t-\tau)d\tau = \int H(t,f)X(f)e^{j2\pi ft}df$$
$$= \int_0^{\tau_{max}} \int_{-\nu_{max}/2}^{\nu_{max}/2} C(\nu,\tau)x(t-\tau)e^{j2\pi\nu t}d\nu d\tau \qquad (1)$$

where $x(t)$ and $y(t)$ represent the transmitted and received waveforms, respectively, and $X(f)$ is the Fourier transform of $x(t)$. The channel is characterized by the time-varying impulse response, $h(t,\tau)$, or the time-varying frequency response, $H(t,f)$, or the delay-Doppler spreading function, $C(\nu,\tau)$. All three channel characterizations are equivalent and related to each other via Fourier transforms. The exact relationship between these channel representations is illustrated in Fig. 1.

The parameters τ_{max} and ν_{max} in (1) are the two key channel parameters: τ_{max}, the delay spread of the channel, is defined as the maximum possible nonzero delay introduced by the channel and $\nu_{max}/2$, the Doppler spread of the channel, is defined as the maximum possible (one-sided) Doppler shift caused by the channel. Throughout the paper, we implicitly consider communication using packets of duration T and (two-sided) bandwidth W. Thus, the dimension of the transceiver signal space is $N_o \approx TW$, the time-bandwidth product. The

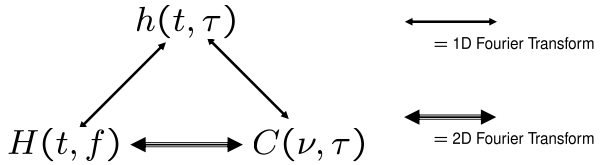


Fig. 1. Relationship between the time-varying impulse response, time-varying frequency response and delay-Doppler spreading function.

focus of this paper is on learning doubly-selective channels, which are characterized by the fact that the delay spread and Doppler spread of the channel are large relative to the inverse of the signaling bandwidth and duration, respectively, i.e., $W\tau_{max} \geq 1$ and $T\nu_{max} \geq 1$. We further limit ourselves to underspread channels, characterized by $\tau_{max}\nu_{max} \ll 1$, which is true of most radio channels [29], and assume that there is no interpacket interference in time and/or frequency, i.e., $T \gg \tau_{max}$ and $W \gg \nu_{max}$.

B. Doubly-Selective Channels: Virtual Representation

Doubly-selective channels generate multiple delayed, Doppler-shifted and attenuated copies of the transmitted waveform. A discrete path model is frequently used to capture the characteristics of these channels in terms of the physical propagation paths. In the discrete path model, the delay-Doppler spreading function of the channel is expressed as

$$C(\nu,\tau) = \sum_{i=1}^{N_{path}} \alpha_i \delta(\nu-\nu_i)\delta(\tau-\tau_i) \qquad (2)$$

and the transmitted and received waveforms are related by

$$y(t) = \sum_{i=1}^{N_{path}} \alpha_i e^{j2\pi\nu_i t}x(t-\tau_i) \qquad (3)$$

which corresponds to signal propagation along N_{path} physical paths, where $\alpha_i \in \mathbb{C}$, $\nu_i \in [-\nu_{max}/2, \nu_{max}/2]$ and $\tau_i \in [0, \tau_{max}]$ are the complex path gain, the delay and the Doppler shift associated with the i-th physical path, respectively.

The discrete path model (2), while realistic, is difficult to analyze and learn due to its nonlinear dependence on a potentially large number of physical parameters $\{(\alpha_i, \nu_i, \tau_i)\}$. However, because of the finite signaling duration and bandwidth, the discrete path model can be accurately approximated by a linear (in parameters) counterpart, known as a virtual channel model, with the aid of sampling theorems and/or power series expansions—see, e.g., [2], [26]. The key idea behind virtual channel modeling is to provide a low-dimensional approximation of the discrete path model by uniformly sampling the physical multipath environment in the delay-Doppler domain at a resolution commensurate with W and T ($\Delta\tau = 1/W, \Delta\nu = 1/T$). That is,

$$y(t) \approx \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} h_{\ell,k} e^{j2\pi\frac{k}{T}t}x(t-\ell/W) \qquad (4)$$

$$h_{\ell,k} \approx \sum_{i\in\mathcal{S}_{\tau,\ell}\cap\mathcal{S}_{\nu,k}} \alpha_i\, e^{-j\pi(k-T\nu_i)} \cdot \text{sinc}(k-T\nu_i)\text{sinc}(\ell-W\tau_i) \qquad (5)$$

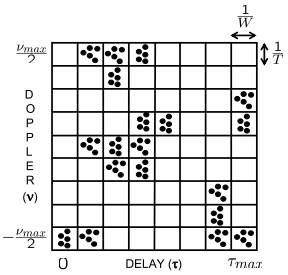


Fig. 2. A schematic illustrating the virtual representation of a single-antenna, doubly-selective channel. Each black dot denotes the contribution of a distinct physical path to the delay-Doppler spreading function and the virtual channel coefficients $\{h_{\ell,k}\}$ correspond to uniformly-spaced samples of a smoothed version of the spreading function taken at $\{(\hat{\tau}_\ell, \hat{\nu}_k) = (\ell/W, k/T)\}$.

where $\text{sinc}(a) = \sin(\pi a)/\pi a$, and $L = \lceil W\tau_{max} \rceil + 1$ and $K = \lceil T\nu_{max}/2 \rceil$ denote the maximum number of resolvable delays and (one-sided) Doppler shifts within the delay-Doppler spreading function, respectively. The set $\mathcal{S}_{\tau,\ell} = \{i : \tau_i \in [\ell/W - 1/2W, \ell/W + 1/2W)\}$ is the set of indices of all paths whose delays lie within the delay resolution bin of width $\Delta\tau = 1/W$ centered around the ℓ-th virtual delay, $\hat{\tau}_\ell = \ell/W$, while $\mathcal{S}_{\nu,k} = \{i : \nu_i \in [k/T - 1/2T, k/T + 1/2T)\}$ denotes the set of indices of all paths whose Doppler shifts lie within the Doppler resolution bin of width $\Delta\nu = 1/T$ centered around the k-th virtual Doppler shift, $\hat{\nu}_k = k/T$. The parameters $\{h_{\ell,k}\}$ are termed as the virtual channel coefficients in the delay-Doppler domain. The expression (5) states that the channel coefficient $h_{\ell,k}$ approximately consists of the sum of gains of all paths whose delays and Doppler shifts lie within the (ℓ, k)-th delay-Doppler resolution bin of size $\Delta\tau \times \Delta\nu$ centered around the sampling point $(\hat{\tau}_\ell, \hat{\nu}_k) = (\ell/W, k/T)$ in the delay-Doppler domain, as illustrated in Fig. 2. In essence, the virtual representation (4) effectively approximates a discrete path doubly-selective channel in terms of an N-dimensional parameter comprising of the virtual channel coefficients $\{h_{\ell,k}\}$,[1] where $N = L \cdot (2K + 1) = (\lceil W\tau_{max} \rceil + 1) \cdot (2\lceil T\nu_{max}/2 \rceil + 1) \approx \tau_{max}\nu_{max}N_o$.

C. Sparse Doubly-Selective Channels

Channel measurement results dating as far back as 1987 [11] and as recent as 2007 [13] suggest that multipath components tend to arrive at the receiver in clusters. 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 interspacings between different multipath clusters within the delay-Doppler domain, doubly-selective channels can be characterized as either "rich" or "sparse". In a rich multipath channel, the interspacings are smaller than $\Delta\tau = 1/W$ in delay and $\Delta\nu = 1/T$ in Doppler. Sparse multipath channels, on the other hand, exhibit interspacings that are larger than $\Delta\tau$ and/or $\Delta\nu$. Similar to the setting in Fig. 2, not every delay-Doppler bin of size $\Delta\tau \times \Delta\nu$ contains a physical path in this case. In particular, since a channel coefficient consists of the sum of gains of all paths falling within its respective delay-Doppler resolution bin, sparse doubly-selective channels tend to have far fewer than N nonzero channel coefficients at any fixed (but large enough) signaling duration and/or bandwidth. We formalize this notion of delay-Doppler sparsity as follows.

Definition 1 (D-Sparse Channels): Let D denote the number of *effective* DoF in a doubly-selective channel, that is, $D = |\{(\ell, k) : h_{\ell,k} > 0\}|$. We say that the channel is D-sparse if $D \ll N$, where $N = L \cdot (2K + 1) \approx \tau_{max}\nu_{max}N_o$ is the total number of resolvable delays and Doppler shifts (channel coefficients) within the delay-Doppler spread.

III. Learning Sparse Doubly-Selective Channels: Single-Carrier Signaling

Since the virtual representation of a doubly-selective channel captures its essential characteristics in terms of the channel coefficients $\{h_{\ell,k}\}$, the channel learning problem is equivalent to the design and placement of the training waveform $x(t)$ within the N_o-dimensional signal space and estimation of $h_{\ell,k}$'s from the (noisy) received waveform $y(t)$. The signaling waveforms commonly employed for channel sensing purposes can be broadly categorized as either single-carrier or multicarrier. We begin our treatment of the sensing and estimation of sparse doubly-selective channels by focusing on the case of single-carrier signaling in this section.

A. Sensing Phase

We consider binary phase-shift keying as the modulation scheme and propose the use of a single-carrier spread spectrum waveform corresponding to a particular spreading code for training purposes. The resulting training waveform $x(t)$ can be represented as

$$x(t) = \sum_{n=0}^{N_o-1} x_n I_{[0,T_c)}(t - nT_c), \quad 0 \le t < T \tag{6}$$

where $I_{[0,T_c)}(t)$ is the chip waveform, $T_c \approx 1/W$ is the chip duration and $\{x_n \in \mathbb{R}\}$ is the spreading code corresponding to the training waveform. The output of the channel corresponding to $x(t)$ is given by (cf. (4))

$$y(t) \approx \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} h_{\ell,k} e^{j2\pi \frac{k}{T} t} x(t - \ell/W) + z(t), \quad 0 \le t < T + \tau_{max} \tag{7}$$

where $z(t)$ is a zero-mean, circularly symmetric, complex additive white Gaussian noise (AWGN) waveform. For spread spectrum waveforms, chip-rate sampling of $y(t)$ at the receiver yields an equivalent discrete-time representation

$$y_n = \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} h_{\ell,k} e^{j2\pi \frac{k}{N_o} n} x_{n-\ell} + z_n, \quad n = 0, 1, \ldots, N_o + L - 2 \tag{8}$$

[1] Note that the approximation gets more accurate with increasing T and W, due to higher delay-Doppler resolution.

where $\{z_n\}$ corresponds to a zero-mean, circularly symmetric, complex AWGN sequence and $N_o \approx TW$ is the dimension of the transceiver signal space.

Now let $\tilde{N}_o = N_o + L - 1$ and define an $\tilde{N}_o$-length sequence of vectors $\{\mathbf{x}_n \in \mathbb{C}^L\}$ comprising of the spreading code $\{x_n\}$ as follows

$$\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-1} & \dots & x_{n-(L-1)} \end{bmatrix}', \quad n = 0, 1, \dots, \tilde{N}_o - 1$$

where the notational understanding is that $x_i = 0$ for $i \notin \{0, 1, \dots, N_o - 1\}$. Further, let

$$\mathbf{H} = \begin{bmatrix} h_{0,-K} & h_{0,-K+1} & \dots & h_{0,K} \\ h_{1,-K} & h_{1,-K+1} & \dots & h_{1,K} \\ \vdots & \vdots & & \vdots \\ h_{L-1,-K} & h_{L-1,-K+1} & \dots & h_{L-1,K} \end{bmatrix} \tag{9}$$

be the $L \times (2K+1)$ matrix of channel coefficients. Note that each column of the channel matrix $\mathbf{H}$ represents the impulse response of the channel corresponding to some *fixed* Doppler shift. Finally, let $\{\mathbf{u}_n \in \mathbb{C}^{2K+1}\}$ be an $\tilde{N}_o$-length sequence of phase vectors given by

$$\mathbf{u}_n = \begin{bmatrix} \omega_{N_o}^{Kn} & \omega_{N_o}^{(K-1)n} & \dots & \omega_{N_o}^{-Kn} \end{bmatrix}'$$

where $\omega_{N_o} = e^{-j\frac{2\pi}{N_o}}$ and $n = 0, 1, \dots, \tilde{N}_o - 1$. Then the sequence $\{y_n\}$ in (8) can be written as

$$\begin{aligned} y_n &= \mathbf{x}_n' \mathbf{H} \mathbf{u}_n + z_n = (\mathbf{u}_n' \otimes \mathbf{x}_n') \operatorname{vec}(\mathbf{H}) + z_n \\ &= (\mathbf{u}_n' \otimes \mathbf{x}_n')\, \mathbf{h} + z_n, \quad n = 0, 1, \dots, \tilde{N}_o - 1 \end{aligned} \tag{10}$$

where $\mathbf{h} = \operatorname{vec}(\mathbf{H}) \in \mathbb{C}^N$ is the vector of channel coefficients, and stacking y_n's into an $\tilde{N}_o$-dimensional vector $\mathbf{y}$ yields the following system of equations

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{z} \tag{11}$$

where the $\tilde{N}_o \times N$ "sensing matrix" $\mathbf{X}$ is comprised of $\{\mathbf{u}_n' \otimes \mathbf{x}_n'\}$ as its rows: $\mathbf{X} = \begin{bmatrix} \mathbf{u}_0 \otimes \mathbf{x}_0 & \dots & \mathbf{u}_{\tilde{N}_o - 1} \otimes \mathbf{x}_{\tilde{N}_o - 1} \end{bmatrix}'$. In the following, we shall treat $\mathbf{h}$ as a *deterministic but unknown* vector. It is further assumed that the communication system has a transmit energy budget of $\mathcal{E}$ for training purposes, i.e, $\sum_{n=0}^{N_o-1} \mathbb{E}[|x_n|^2] = \mathcal{E}$. Finally, without loss of generality, we assume that the spreading code $\{x_n\}$ is generated from a Rademacher distribution, i.e., x_n's independently take values $+\sqrt{\mathcal{E}/N_o}$ or $-\sqrt{\mathcal{E}/N_o}$ with probability $1/2$ each, and $\mathbf{z}$ is distributed as $\mathcal{CN}(\mathbf{0}_{\tilde{N}_o}, \mathbf{I}_{\tilde{N}_o})$.

B. Estimation Phase

The model (11) is a linear observation model with $N = L \cdot (2K+1)$ unknowns and it can be shown that the sensing matrix $\mathbf{X}$ has full column rank. In this case, and under no a priori sparsity assumption, the least squares (LS) estimator of the channel vector $\mathbf{h}$

$$\widehat{\mathbf{h}}_{LS} = (\mathbf{X}^\dagger \mathbf{X})^{-1} \mathbf{X}^\dagger \mathbf{y} \tag{12}$$

is known to be optimal in the sense that (i) it is also the maximum likelihood estimate of $\mathbf{h}$, and (ii) it achieves the Cramer-Rao lower bound [30].

Many real-world channels of practical interest, such as underwater acoustic channels [16], digital television channels [31] and residential ultrawideband channels [12], however, tend to be either sparse or approximately sparse, with $D = \|\mathbf{h}\|_0 \ll N$. Unfortunately, conventional LS channel estimators, while appropriate for rich channels, fail to capitalize on the anticipated sparsity of the aforementioned 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 channel estimator to that of a channel estimator that has been equipped with an *oracle*. The oracle does not reveal the true $\mathbf{h}$, but does inform us of the *sparsity pattern* (locations of nonzero entries) of $\mathbf{h}$. Clearly this represents an ideal estimation strategy and one cannot expect to attain its performance level. Nevertheless, it is the benchmark that one should consider. We begin this comparison with the following lemma.

Lemma 1: Given the observation model (11), the MSE of an LS channel estimator is lower bounded as

$$\mathbb{E}\left[\|\widehat{\mathbf{h}}_{LS} - \mathbf{h}\|_2^2\right] \geq \frac{N}{\mathcal{E}} \tag{13}$$

with equality if and only if $\mathbf{X}$ has orthogonal columns.

Sketch of Proof: Given the observation model (11), it is easy to see that

$$\mathbb{E}\left[\|\widehat{\mathbf{h}}_{LS} - \mathbf{h}\|_2^2\right] = \operatorname{tr}((\mathbf{X}^\dagger \mathbf{X})^{-1})$$

and since the trace of a matrix is equal to the sum of its eigenvalues, an application of arithmetic-harmonic means inequality yields

$$\operatorname{tr}((\mathbf{X}^\dagger \mathbf{X})^{-1}) \geq \frac{N^2}{\operatorname{tr}(\mathbf{X}^\dagger \mathbf{X})} = \frac{N}{\mathcal{E}}$$

with equality if and only if $\mathbf{X}^\dagger \mathbf{X} = \mathcal{E}\mathbf{I}_N$. ■

On the other hand, let $\mathcal{I}_* \subset \{1, \dots, N\}$ be the set of indices of the D nonzero entries of $\mathbf{h}$ and suppose that an oracle provides us with $\mathcal{I}_*$. Then an ideal estimator $\mathbf{h}^*$ can be obtained from $\mathbf{y}$ by first forming a *restricted* LS estimator

$$\mathbf{h}_{\mathcal{I}_*} = (\mathbf{X}_{\mathcal{I}_*}^\dagger \mathbf{X}_{\mathcal{I}_*})^{-1} \mathbf{X}_{\mathcal{I}_*}^\dagger \mathbf{y} \tag{14}$$

where $\mathbf{X}_{\mathcal{I}_*}$ is a submatrix obtained by extracting the D columns of $\mathbf{X}$ corresponding to the indices in $\mathcal{I}_*$, and then setting $\mathbf{h}^*$ equal to $\mathbf{h}_{\mathcal{I}_*}$ on the indices in $\mathcal{I}_*$ and zero on the indices in $\mathcal{I}_*^c$. Appealing to the proof of Lemma 1, the MSE of this oracle based channel estimator obeys

$$\mathbb{E}\left[\|\mathbf{h}^* - \mathbf{h}\|_2^2\right] = \operatorname{tr}((\mathbf{X}_{\mathcal{I}_*}^\dagger \mathbf{X}_{\mathcal{I}_*})^{-1}) \geq \frac{D}{\mathcal{E}} \tag{15}$$

with equality if and only if $\mathbf{X}_{\mathcal{I}_*}$ has orthogonal columns. Comparison of the MSE lower bounds (13) and (15) shows that conventional LS channel estimators may be at a significant disadvantage when it comes to identifying sparse channels.

While the ideal estimator $\mathbf{h}^*$ is impossible to construct in practice, we now show that it is possible to obtain a more reliable estimate of $\mathbf{h}$ as a solution to the convex program

$$\widehat{\mathbf{h}} = \underset{\tilde{\mathbf{h}} \in \mathbb{C}^N}{\arg\min} \|\tilde{\mathbf{h}}\|_1 \quad \text{subject to} \quad \|\mathbf{X}^\dagger \mathbf{r}\|_\infty \leq \lambda \tag{16}$$

where $\lambda(N,\mathcal{E}) > 0$ and $\mathbf{r}$ is the $\tilde{N_o}$-dimensional vector of residuals: $\mathbf{r} = \mathbf{y} - \mathbf{X}\tilde{\mathbf{h}}$. This optimization program goes by the name of Dantzig selector (DS) and is computationally tractable since it can be recast as a linear program [22]. 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 [32] which can sometimes be more computationally attractive because of the availability of a wide array of efficient software packages for solving it [33], [34].

The key to proving the efficacy of the DS estimator is in showing that $\mathbf{X}$ satisfies the so-called "restricted isometry property" (RIP) with sufficiently small value of $2D$-restricted isometry constant.

Definition 2 (Restricted Isometry Constant): The $2D$ restricted isometry constant of $\mathbf{X}$, denoted by δ_{2D}, is defined as the smallest value such that

$$\mathcal{E}(1-\delta_{2D})\|\tilde{\mathbf{h}}\|_2^2 \leq \|\mathbf{X}\tilde{\mathbf{h}}\|_2^2 \leq \mathcal{E}(1+\delta_{2D})\|\tilde{\mathbf{h}}\|_2^2 \qquad (17)$$

holds for all $2D$-sparse vectors $\tilde{\mathbf{h}} \in \mathbb{C}^N$. The matrix $\mathbf{X}$ is said to satisfy RIP of order $2D$ if $\delta_{2D} \in [0,1)$.

Note that if any two columns of $\mathbf{X}$ happened to be linearly dependent then $\delta_{2D} \geq 1$. Loosely speaking, RIP of order $2D$ essentially requires that mutual coherence between the columns of $\mathbf{X}$ is sufficiently small so that $\mathbf{X}/\sqrt{\mathcal{E}}$ (approximately) behaves like an isometry on the space of $2D$-sparse vectors. The following theorem asserts that the DS solution is highly accurate in this case.

Theorem 1: Suppose that $\mathbf{X}$ satisfies RIP of order $2D$ with $\delta_{2D} < \sqrt{2}-1$. Choose $\lambda(N,\mathcal{E}) = \sqrt{2\mathcal{E}(1+a)\log N}$ for any $a \geq 0$. Then, with probability exceeding $1 - 2(\sqrt{\pi(1+a)\log N}\cdot N^a)^{-1}$, the DS estimator $\widehat{\mathbf{h}}$ obeys

$$\|\widehat{\mathbf{h}} - \mathbf{h}\|_2^2 \leq c_1^2 \cdot \log N \cdot \left(\frac{D}{\mathcal{E}}\right) \qquad (18)$$

where the constant $c_1 = 4\sqrt{2(1+a)}/\left(1-(\sqrt{2}+1)\delta_{2D}\right)$.

Theorem 1, which is a slight variation on Theorem 1.1 in [22],[2] states that the DS estimator can *potentially* achieve squared error within a factor of $\log N$ of the oracle based MSE lower bound of $D/\mathcal{E}$. However, it remains to be seen whether the sensing matrix $\mathbf{X}$ satisfies RIP with $\delta_{2D} < \sqrt{2}-1$. We now state the key result of this section which shows that this is indeed the case.

Theorem 2: Let $\{x_n\}_{n=0}^{N_o-1}$ be a sequence of independent and identically distributed Rademacher variables taking values $+\sqrt{\mathcal{E}/N_o}$ or $-\sqrt{\mathcal{E}/N_o}$ with probability $1/2$ each. Further, let $\{\mathbf{x}_n \in \mathbb{C}^L\}$ and $\{\mathbf{u}_n \in \mathbb{C}^{2K+1}\}$ be as defined in Section III-A and suppose that the signal space dimension $N_o \geq c_2 \cdot \log N \cdot D^2$. Then, with probability exceeding $1-\exp(-c_3 \cdot N_o)$, the $\tilde{N_o} \times N$ matrix $\mathbf{X}$ given by

$$\mathbf{X} = \begin{bmatrix}\mathbf{u}_0 \otimes \mathbf{x}_0 & \mathbf{u}_1 \otimes \mathbf{x}_1 & \dots & \mathbf{u}_{\tilde{N_o}-1} \otimes \mathbf{x}_{\tilde{N_o}-1}\end{bmatrix}' \qquad (19)$$

satisfies RIP of order $2D$ with $\delta_{2D} \in (0, \sqrt{2}-1)$, where $\tilde{N_o} = N_o + L - 1$ and $N = L\cdot(2K+1)$. Here, $c_2, c_3 > 0$ are constants that do not depend on N or N_o.

The proof of this theorem is provided in the Appendix. Note that the main condition of the theorem $N_o \geq c_2 \cdot \log N \cdot D^2$ is trivially satisfied for sufficiently underspread doubly-selective channels since, by definition, $D \ll N \approx \tau_{max}\nu_{max}N_o \ll N_o$. Therefore Theorem 2, along with Theorem 1, shows that the DS estimator (16) does remarkably better than the LS estimator (12) in learning a D-sparse doubly-selective channel: using single-carrier spread spectrum training waveforms, the MSE improvement is roughly by a factor of $O(N/D)$.

IV. Learning Sparse Doubly-Selective Channels: Multicarrier Signaling

In this section, we consider multicarrier signaling for sensing and estimation of sparse doubly-selective channels. In particular, owing to the fact that orthogonal short-time Fourier (STF) basis functions serve as approximate eigenfunctions for underspread doubly-selective channels [1], [4], we investigate the use of training waveforms that consist of the elements of a complete orthogonal STF basis whose time-frequency support is matched to the channel characteristics.

A. Sensing Phase

A complete orthogonal STF basis for the N_o-dimensional signal space is generated via time and frequency shifts of a fixed prototype pulse $g(t)$: $\gamma_{\ell,k}(t) = g(t-\ell T_o)e^{j2\pi k W_o t}$, $(\ell,k) \in \mathcal{S} = \{0,1,\dots,N_t-1\}\times\{0,1,\dots,N_f-1\}$, where $N_t = T/T_o$ and $N_f = W/W_o$. The prototype pulse is assumed to have unit energy, $\int |g(t)|^2 dt = 1$, and completeness of $\{\gamma_{\ell,k}\}$ stems from the underlying assumption that $T_o W_o = 1$, which results in a total of $N_t N_f = TW/T_oW_o = N_o$ basis elements. Therefore, as opposed to signaling over an incomplete STF basis [1] (corresponding to $T_oW_o > 1$), signaling using a complete STF basis [4] does not lead to an inherent loss in spectral efficiency.[3]

We propose the use of a training waveform that randomly dedicates N_r of the N_o STF basis elements as "pilot tones". That is,

$$x(t) = \sqrt{\frac{\mathcal{E}}{N_r}} \sum_{(n,m)\in\mathcal{S}_r} \gamma_{n,m}(t), \quad 0 \leq t < T \qquad (20)$$

where the set of indices of pilot tones, $\mathcal{S}_r$, consists of N_r elements randomly selected from $\mathcal{S}$ and $\mathcal{E}$ is the transmit energy budget available for training purposes. At the receiver, assuming that the basis parameters T_o and W_o are matched to the channel parameters τ_{max} and ν_{max} so that $\gamma_{\ell,k}$'s serve as approximate eigenfunctions for sufficiently underspread channels [4],[4] projecting the (noisy) received signal $y(t)$ onto

[2]The variation is primarily due to the presence of complex-valued noise as opposed to the real-valued noise in [22, Th. 1.1] and noticing the fact that $\theta_{D,2D} < \sqrt{2}\delta_{2D}$; we refer the reader to [22] for further details.

[3]Note that signaling over a complete orthogonal STF basis can be thought of as block orthogonal frequency division multiplexing (OFDM) signaling with OFDM symbol duration T_o and block length $N_t = T/T_o$.

[4]Two necessary matching conditions are: (i) $\tau_{max} \leq T_o \leq 1/\nu_{max}$ and (ii) $\nu_{max} \leq W_o \leq 1/\tau_{max}$; we refer the reader to [4] for further details.

the STF basis waveforms yields

$$y_{n,m} = \langle y, \gamma_{n,m} \rangle \approx \sqrt{\frac{\mathcal{E}}{N_r}} H_{n,m} + z_{n,m}, \ (n,m) \in \mathcal{S}_r \quad (21)$$

where $\langle y, \gamma_{n,m} \rangle = \int y(t) \overline{\gamma_{n,m}(t)} dt$, $\{z_{n,m}\}$ corresponds to an AWGN sequence and the STF channel coefficients are given by $H_{n,m} \approx H(t,f)\big|_{(t,f)=(nT_o, mW_o)}$ [4].

Now recall from Section II that the time-varying frequency response $H(t,f) = \iint C(\nu,\tau) e^{j2\pi\nu t} e^{-j2\pi\tau f} d\nu d\tau$. The virtual representation of a doubly-selective channel therefore implies that $H(t,f) \approx \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} h_{\ell,k} e^{j2\pi \frac{k}{T} t} e^{-j2\pi \frac{\ell}{W} f}$. Consequently, the STF channel coefficients $\{H_{n,m}\}$ can be written as

$$\begin{aligned} H_{n,m} &= \sum_{\ell=0}^{L-1} \sum_{k=-K}^{K} h_{\ell,k} e^{j2\pi \frac{k}{N_t} n} e^{-j2\pi \frac{\ell}{N_f} m} = \mathbf{u}'_{f,m} \mathbf{H} \mathbf{u}_{t,n} \\ &= (\mathbf{u}'_{t,n} \otimes \mathbf{u}'_{f,m}) \text{vec}(\mathbf{H}) = (\mathbf{u}'_{t,n} \otimes \mathbf{u}'_{f,m}) \mathbf{h} \end{aligned} \quad (22)$$

where $\mathbf{H}$ is the $L \times (2K+1)$ matrix of channel coefficients defined earlier in (9), $\mathbf{h} = \text{vec}(\mathbf{H}) \in \mathbb{C}^N$, $\mathbf{u}_{f,m} = \begin{bmatrix} 1 & \omega_{N_f}^{1m} & \dots & \omega_{N_f}^{(L-1)m} \end{bmatrix}' \in \mathbb{C}^L$ and $\mathbf{u}_{t,n} = \begin{bmatrix} \omega_{N_t}^{Kn} & \omega_{N_t}^{(K-1)n} & \dots & \omega_{N_t}^{-Kn} \end{bmatrix}' \in \mathbb{C}^{2K+1}$. It is worth noting at this point that under the assumption of STF basis parameters being matched to the channel parameters (specifically, $T_o \leq 1/\nu_{max}$ and $W_o \leq 1/\tau_{max}$), one can easily ensure that $N_t \geq 2K+1$ and $N_f \geq L$. Finally, stacking the received training symbols $\{y_{n,m}\}$ into an N_r-dimensional vector $\mathbf{y}$ yields the following system of equations

$$\mathbf{y} = \mathbf{U}\mathbf{h} + \mathbf{z} \quad (23)$$

where the $N_r \times N$ sensing matrix $\mathbf{U}$ is comprised of $\{\sqrt{\mathcal{E}/N_r}(\mathbf{u}'_{t,n} \otimes \mathbf{u}'_{f,m}) : (n,m) \in \mathcal{S}_r\}$ as its rows and the AWGN vector $\mathbf{z}$ is distributed as $\mathcal{CN}(\mathbf{0}_{N_r}, \mathbf{I}_{N_r})$.

B. Estimation Phase

Similar to (11), the model (23) is a linear observation model with $N = L \cdot (2K+1)$ unknowns. To obtain reasonable channel estimates in this multicarrier setting, conventional channel estimators based on the LS criterion rely on the assumption that the number of pilot tones $N_r \geq N$ [30], [35]. It can be shown in this case that $\mathbf{U}$ has full column rank and the resulting LS channel estimator is of the form

$$\widehat{\mathbf{h}}_{LS} = (\mathbf{U}^\dagger \mathbf{U})^{-1} \mathbf{U}^\dagger \mathbf{y}. \quad (24)$$

As noted earlier, however, a LS channel estimator (while known to be optimal for nonsparse channels) is ill-suited for the purposes of estimating a sparse channel. To see this, note that the MSE of the LS estimator (24) is lower bounded by $N/\mathcal{E}$ (cf. Lemma 1). On the other hand, using arguments similar to the ones made in Section III-B, an ideal channel estimator having access to an oracle can be shown to have the MSE lower bound of $D/\mathcal{E}$. Equally importantly, the ideal estimator also does not require $N_r \geq N$ pilot tones and can provide reasonable estimates as long as $N_r \geq D$ (cf. (14)). This is especially important from the system efficiency viewpoint since one extra dimension allocated for training purposes is one less dimension available for data transmission.

The main thesis of this section is that it is in fact possible to come within a logarithmic factor of the performance of an ideal estimator, both in terms of the MSE and the minimum number of pilots needed. The proposed estimator is once again given as the solution to the Dantzig selector (DS)

$$\widehat{\mathbf{h}} = \underset{\tilde{\mathbf{h}} \in \mathbb{C}^N}{\arg\min} \|\tilde{\mathbf{h}}\|_1 \quad \text{subject to} \quad \|\mathbf{U}^\dagger \mathbf{r}\|_\infty \leq \lambda \quad (25)$$

where $\lambda(N, \mathcal{E}) = \sqrt{2\mathcal{E}(1+a)\log N}$ for some $a \geq 0$ and $\mathbf{r}$ is the N_r-dimensional vector of residuals: $\mathbf{r} = \mathbf{y} - \mathbf{U}\tilde{\mathbf{h}}$. Theorem 1, with $\mathbf{X}$ replaced by $\mathbf{U}$, is still applicable in this setting, which implies that the DS estimator obeys

$$\|\widehat{\mathbf{h}} - \mathbf{h}\|_2^2 \leq c_1^2 \cdot \log N \cdot \left(\frac{D}{\mathcal{E}}\right) \quad (26)$$

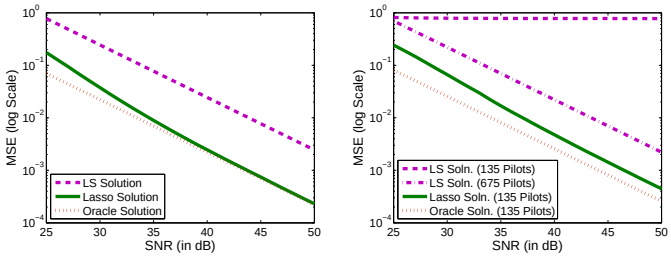
with high probability as long as $\mathbf{U}$ satisfies RIP of order $2D$ with $\delta_{2D} < \sqrt{2} - 1$. The goal, then, is to determine the number of pilot tones N_r for which (if any) $\mathbf{U}$ satisfies the aforementioned RIP condition. The key result of this section, which helps address this question, is stated in terms of the following theorem.

Theorem 3: Let $\mathcal{S} = \{0, 1, \dots, N_t - 1\} \times \{0, 1, \dots, N_f - 1\}$ and $\mathcal{S}_r$ be a random set of N_r ordered pairs sampled uniformly at random from $\mathcal{S}$. Further, let $\{\mathbf{u}_{f,m} \in \mathbb{C}^L\}$ and $\{\mathbf{u}_{t,n} \in \mathbb{C}^{2K+1}\}$ be as defined in Section IV-A and suppose that $N_r \geq c_4 \cdot \log^5 N_o \cdot D$. Then, with probability exceeding $1 - c_5 N_0^{-c_6}$, the $N_r \times N$ matrix $\mathbf{U}$ comprising of the vectors $\{\sqrt{\mathcal{E}/N_r}(\mathbf{u}'_{t,n} \otimes \mathbf{u}'_{f,m}) : (n,m) \in \mathcal{S}_r\}$ as its rows satisfies RIP of order $2D$ with $\delta_{2D} \in (0, \sqrt{2} - 1)$. Here, c_4, c_5 and c_6 are strictly positive constants that do not depend on N or N_o.

The proof of this theorem, which leverages some key ideas from [36], [37], is provided in the Appendix. Theorems 1 and 3 show that, even in the multicarrier setting, the DS estimator (25) comes remarkably close to matching the performance of an ideal estimator. And as for a comparison with the LS estimator (24), ignoring the $\log$ factors, the DS estimator roughly results in a decrease in the number of pilot tones and the MSE by a factor of $O(N/D)$. Finally, note that while Theorem 3 requires the number of pilot tones to satisfy $N_r \geq c_4 \cdot \log^5 N_o \cdot D$, it is conjectured that the true lower bound on N_r is along the lines of $N_r \geq c_7 \cdot \log N_o \cdot D$ for some constant $c_7 > 0$; see, e.g., [37].

V. Numerical Results and Discussion

We begin this section by numerically comparing the MSE performance of the sparse channel learning techniques proposed in Sections III and IV with that of conventional strategies comprising of linear LS channel estimators. The simulation parameters are chosen to be depictive of a communication system with (i) *Channel Parameters*: $\tau_{max} = 250$ μs and $\nu_{max} = 350$ Hz (corresponding to, e.g., a carrier frequency of 1.89 GHz and maximum speed of 100 km/h), and (ii) *Signaling Parameters*: $T = 45$ ms and $W = 45$ kHz, which



(a) (b)

Fig. 3. Numerical results comparing the performance of a lasso estimator with that of a LS estimator. The MSEs of the channel estimates are plotted on a log scale against the SNR in dB corresponding to (a) Spread Spectrum Training Waveforms, and (b) STF Training Waveforms.

result in $N_o = TW = 2025$ and $N = L \cdot (2K+1) = 221$. For the case of multicarrier signaling, the STF basis parameters are chosen to be $T_o = 1$ ms and $W_o = 1$ kHz, which correspond to $N_t = N_f = 45$.

The simulations are carried out under the assumption that only 10% of the channel coefficients are nonzero, i.e., $D = 22$. The simulation setup corresponds to realizing the channel matrix $\mathbf{H}$ given in (9) by first randomly selecting the locations of 22 nonzero channel coefficients and then generating their values from independent realizations of $\mathcal{CN}(0, 1/22)$. The output of the channel is observed at different values of signal-to-noise ratio (SNR), and LS and lasso estimates are obtained by pseudo-inverting the sensing matrices and executing SpaRSA [34], respectively.[5] Same (randomly generated) spreading code is used for both LS and lasso estimates in the case of single-carrier training. Multicarrier training is carried out by randomly designating N_r of the N_o STF basis functions as pilot tones in the case of lasso estimate, and by using a comb-type pilot arrangement in the case of LS estimate. That is, $x_{LS}(t) = \sqrt{\mathcal{E}/N_r}\sum_{(n,m)\in\mathcal{P}} \gamma_{n,m}(t)$, $\mathcal{P} = \{(n,m) : n = 0, 1, \ldots, N_t - 1,\ m = 0, N_f/p, \ldots, (p-1)N_f/p\}$, where it is assumed that $N_r = pN_t$ for some p that is a factor of N_f. This is because of the fact that comb-type pilot arrangements are known to be optimal for LS channel estimators [35].

The MSEs of the channel estimates, corresponding to averaging over 1000 independent trials, are plotted against the SNR in Fig. 3. As expected, the lasso estimator substantially outperforms the LS estimator and comes very close to matching the performance of the oracle based ideal channel estimator. In particular, the gap between the MSEs of the LS and lasso estimates corresponding to the spread spectrum training waveform is on the order of 7 dB at low SNR and 10 dB at high SNR—see Fig. 3(a). As for the case of STF training, the LS estimator fails to yield a consistent estimate when $N_r = 135 < N = 221$, and severely underperforms the lasso estimator with 135 pilots even when it itself utilizes 675 pilots; the loss in spectral efficiency is about 7 dB and the gap between the MSEs is on the order of 4.5 dB at low SNR and 7 dB at high SNR—see Fig. 3(b).

A few concluding remarks are in order now. Firstly, notice that the gap between the MSEs of the lasso estimate and the ideal estimate corresponding to spread spectrum training is much smaller than the one corresponding to STF training. This observation is attributable to the fact that the probability of the sensing matrix $\mathbf{X}$ (corresponding to the spread spectrum training waveform) not satisfying the RIP condition goes to zero exponentially in N_o, whereas for the sensing matrix $\mathbf{U}$ (corresponding to the STF training waveform) it goes to zero only polynomially in N_o (cf. Theorems 2 and 3). However, as to the question of which of the two training waveforms is best suited for channel sensing purposes, the answer depends on how the channel learning module integrates with the data transmission module, a detailed discussion of which is beyond the scope of this exposition.

Secondly, recall that $N \approx \tau_{max}\nu_{max}N_o$. Therefore, assuming that $D \sim N_o^{\mu_1}$ for some $\mu_1 \in [0, 1)$, the training-based schemes proposed in this paper yield estimates for which the MSE per channel coefficient scales as $\mathbb{E}[\|\widehat{\mathbf{h}} - \mathbf{h}\|_2^2]/N \sim N_o^{-1+\mu_1}/\mathcal{E}$. Hence, as long as the training energy $\mathcal{E} \sim N_o^{-\mu_2}$ for some $\mu_2 \in (0, 1-\mu_1)$, both the MSE per channel coefficient and the training energy would go to zero asymptotically in N_o. This shows that sublinearly sparse doubly-selective channels are asymptotically coherent—an observation that was made earlier in [14], albeit under the restrictive assumption of known channel sparsity pattern (the oracle setting).

Lastly, note that the appeal of the training-based methods proposed in this paper goes beyond the identification of truly sparse doubly-selective channels. Indeed, certain propagation environments might yield channels that are only *approximately* sparse. One such class of channels could be, for example, that the magnitudes of the ordered channel coefficients exhibit a power law decay in the sense that the j-th absolutely largest entry in $\mathbf{h} = \text{vec}(\mathbf{H})$ satisfies $|h_{(j)}| \leq \alpha j^{-1/s}$ for some $\alpha > 0$ and $s \leq 1$. Then, redefining the sparsity parameter D as $D = |\{j : |h_{(j)}| > \mathcal{E}^{-1/2}\}|$, it is easy to show that employing either spread spectrum training waveforms with $N_o \geq c_2 \cdot \log N \cdot D^2$ or STF training waveforms with $N_r \geq c_4 \cdot \log^5 N_o \cdot D$ and making use of the DS estimator yield channel estimates that achieve, with high probability, the minimax error rate over this class of approximately sparse channels—see [22, Th. 1.3].

APPENDIX

A. Proof of Theorem 2

We begin by noting that it is sufficient to prove the theorem with $\mathcal{E} = 1$, since the general case would follow from a simple rescaling argument. Therefore, we assume $\mathcal{E} = 1$ from now on. For ease of notation, define $\delta = \delta_{2D} \in (0, \sqrt{2} - 1)$ and $S = 2D$. Proving the RIP condition in the theorem requires showing that, for all subsets $\mathcal{I} \subset \{1, 2, \ldots, N\}$ which satisfy $|\mathcal{I}| = S$, the eigenvalues of the Gram matrix $\mathbf{G}(\mathcal{I}) = \mathbf{X}_{\mathcal{I}}^{\dagger}\mathbf{X}_{\mathcal{I}}$ lie in the interval $[1 - \delta, 1 + \delta]$. Here, $\mathbf{X}_{\mathcal{I}}$ denotes the $\tilde{N_o} \times S$ submatrix obtained by retaining the columns of $\mathbf{X}$ corresponding to the indices in $\mathcal{I}$.

The above condition can be established for a *fixed* $\mathcal{I}$ using the eigenvalue perturbation theory. In particular, Geršgorin

[5] As noted earlier, lasso is expected to perform as well as the DS [32].

disc theorem states that every eigenvalue of the Hermitian matrix $\mathbf{G}(\mathcal{I})$ lies in the union of S intervals given by

$$\mathcal{R}_i(\mathcal{I}) = \Big\{ z \in \mathbb{R} : |z - g_{i,i}(\mathcal{I})| \leq \sum_{\substack{j=1 \\ j \neq i}}^{S} |g_{i,j}(\mathcal{I})| \Big\}.$$

That is, $\big\{\text{eigenvalues of } \mathbf{G}(\mathcal{I})\big\} \subseteq \cup_{i=1}^{S} \mathcal{R}_i(\mathcal{I})$ [38]. Notice that $g_{i,i}(\mathcal{I}) = 1$ for every subset $\mathcal{I}$. Therefore, to establish that the eigenvalues of $\mathbf{G}(\mathcal{I})$ lie in $[1-\delta, 1+\delta]$, it is sufficient to show that $|g_{i,j}(\mathcal{I})| \leq \delta/S \; \forall \; i,j, i \neq j$, since this would imply that $\mathcal{R}_i(\mathcal{I}) \subset [1-\delta, 1+\delta]$ for all i.

Next, to guarantee that the eigenvalues of $\mathbf{G}(\mathcal{I})$ lie in $[1-\delta, 1+\delta]$ for *every* $\mathcal{I}$, consider the full $N \times N$ Gram matrix of $\mathbf{X}$, $\mathbf{G} = \mathbf{X}^{\dagger}\mathbf{X}$. Since the $S \times S$ Gram matrix $\mathbf{G}(\mathcal{I})$ corresponding to any $\mathcal{I}$ is a submatrix of $\mathbf{G}$, showing that every off-diagonal entry of $\mathbf{G}$ is bounded above by δ/S in absolute value is sufficient to guarantee the boundedness of the off-diagonals of $\mathbf{G}(\mathcal{I})$, $|g_{i,j}| \leq \delta/S \; \forall \; i,j, i \neq j \Rightarrow |g_{i,j}(\mathcal{I})| \leq \delta/S \; \forall \; \mathcal{I}, \forall \; i,j, i \neq j$, which in turn would imply that the eigenvalues of all $\binom{N}{S}$ Gram matrices $\mathbf{G}(\mathcal{I})$ lie in $[1-\delta, 1+\delta]$.

We proceed with our goal of showing that $|g_{i,j}| \leq \delta/S \; \forall \; i,j, i \neq j$ by writing the sensing matrix $\mathbf{X} = \big[\mathbf{u}_0 \otimes \mathbf{x}_0 \quad \ldots \quad \mathbf{u}_{\tilde{N}_o-1} \otimes \mathbf{x}_{\tilde{N}_o-1}\big]'$ in the form of a block matrix

$$\mathbf{X} = \big[\mathbf{X}_{-K} \quad \ldots \quad \mathbf{X}_0 \quad \ldots \quad \mathbf{X}_K\big].$$

Here, each $\tilde{N}_o \times L$ block $\mathbf{X}_r$ is of the form $\mathbf{X}_r = \mathbf{W}_r\mathbf{T}$, where $\mathbf{W}_r = \text{diag}(\omega_{N_0}^{-r0}, \omega_{N_0}^{-r1}, \ldots, \omega_{N_0}^{-r(\tilde{N}_o-1)}) \in \mathbb{C}^{\tilde{N}_o \times \tilde{N}_o}$ and $\mathbf{T} \in \mathbb{R}^{\tilde{N}_o \times L}$ is a Toeplitz matrix having $\big[x_0 \quad \mathbf{0}_{L-1}\big]$ as its first row and $\big[x_0 \quad \ldots \quad x_{N_o-1} \quad \mathbf{0}_{L-1}\big]'$ as its first column. Since each $g_{i,j}$ is simply the inner product between the i-th and j-th column of $\mathbf{X}$, we can alternatively bound $|\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle|$ corresponding to $-K \leq r, r' \leq K, \; 1 \leq p, p' \leq L, \; p = p' \Leftrightarrow r \neq r'$, where $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^{\dagger}\mathbf{b}$ and $\mathbf{x}_{r,p}$ denotes the p-th column of $\mathbf{X}_r$. Note that since $|\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle| = |\langle \mathbf{x}_{r',p'}, \mathbf{x}_{r,p} \rangle|$, there are mainly two cases that need to be considered here: (i) $p = p'$ (possible only when $r \neq r'$), and (ii) $p < p'$.

For case (i), we have $\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p} \rangle = \sum_{n=0}^{N_o-1} \omega_{N_0}^{rn} \omega_{N_0}^{-r'n} = 0$ (since $r \neq r'$). For case (ii), define $\Delta = p' - p$ and write

$$\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle = \sum_{n=0}^{N_o-1-\Delta} x_n x_{n+\Delta} \, \omega_{N_o}^{(r-r')(n+p'-1)}. \tag{27}$$

Observe that $|\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle|$ cannot be bounded through the use of standard concentration inequalities since the terms in the above sum are not mutually independent. For example, consider the case of $p = 1, p' = 2, r = r' = 0$, and $N_o = 5$. Then $\langle \mathbf{x}_{0,1}, \mathbf{x}_{0,2} \rangle = x_0x_1 + x_1x_2 + x_2x_3 + x_3x_4$, and the first two terms are dependent (due to x_1), as are the second and third terms (due to x_2), etc. Notice, however, that the first and third terms and the second and fourth terms are independent, which suggests that the entire sum can be written as two sums of mutually independent terms.

We now prove that this is true in general. That is, $\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle$ for any r, r' and $p < p'$ can always be written as two sums having mutually independent terms. To establish this claim, rearrange the summands in (27) as follows

$$\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle = \sum_{n=0}^{\Delta-1} \sum_{i=0}^{\lfloor \frac{N_o-1-\Delta-n}{\Delta} \rfloor} x_{n+i\Delta} \, x_{n+(i+1)\Delta} \cdot \omega_{N_o}^{(r-r')(n+i\Delta+p'-1)}. \tag{28}$$

Notice that (i) each term in an inner sum is only dependent with its adjacent terms in the sum, and (ii) the inner sums are mutually independent. Consequently, indexing the terms in (28) from 1 to $N_o - \Delta$, it is easy to see that all the odd-indexed terms are mutually independent, as are all the even-indexed ones. Finally, partitioning the above sum into odd- and even-indexed terms and appropriately reindexing the terms yield

$$\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle = \sum_{n_1=1}^{\lceil \frac{N_o-\Delta}{2} \rceil} x'_{n_1} e^{j\phi_{n_1}} + \sum_{n_2=1}^{\lfloor \frac{N_o-\Delta}{2} \rfloor} x'_{n_2} e^{j\phi_{n_2}} \tag{29}$$

where $\{x'_{n_1}\}$, $\{x'_{n_2}\}$ consist of mutually independent Rademacher variables that are distributed as $\pm 1/N_o$ with probability 1/2 each, and $\{\phi_{n_1}\}$, $\{\phi_{n_2}\}$ are the deterministic phase factors.

To proceed further, write $\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle$ in (29) as $\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle = S_{q_1} + S_{q_2}$, where $q_1 = \lceil \frac{N_o-\Delta}{2} \rceil$, $q_2 = \lfloor \frac{N_o-\Delta}{2} \rfloor$, and note that

$$\begin{aligned} &\Pr\Big(|\langle \mathbf{x}_{r,p}, \mathbf{x}_{r',p'} \rangle| > \delta/S \Big) \\ &\overset{(a)}{\leq} \; 2 \max\Big\{ \Pr\big(|S_{q_1}| > \delta/2S\big), \Pr\big(|S_{q_2}| > \delta/2S\big) \Big\} \\ &\overset{(b)}{\leq} \; 2 \max\left\{ 4\exp\left(\frac{-\delta^2 N_o^2}{16 S^2 q_1} \right), \, 4\exp\left(\frac{-\delta^2 N_o^2}{16 S^2 q_2} \right) \right\} \\ &\overset{(c)}{\leq} \; 8\exp\left(\frac{-\delta^2 N_o}{8S^2} \right). \end{aligned} \tag{30}$$

Here, (a) follows from a simple union bounding argument, (b) follows from an application of Hoeffding's inequality (adapted to bounded complex random variables), and (c) follows from the fact that $N_o/2 \geq q_1 \geq q_2$ for any $\Delta \geq 1$.

We have now established that the probability that $|g_{i,j}| > \delta/S$ does not exceed $8\exp\left(\frac{-\delta^2 N_o}{8S^2}\right) \; \forall \; i,j, i \neq j$. To satisfy the RIP condition, however, we need to upper bound the probability that $\max_{i,j,i\neq j} |g_{i,j}| = \max_{i<j} |g_{i,j}| > \delta/S$. To this end, we apply another union bounding argument to yield

$$\begin{aligned} \Pr\Big(\max_{i<j} |g_{i,j}| > \delta/S \Big) &\leq 4N(N-1)\exp\left(\frac{-\delta^2 N_o}{8S^2} \right) \\ &\leq \exp\left(\frac{-\delta^2 N_o}{8S^2} + 2\log 2N \right) \end{aligned} \tag{31}$$

which completes the proof of the theorem.

Remark 1: It is worth noting at this point that (i) if $K = 0$ (corresponding to a purely frequency-selective channel) then Theorem 2 reduces to [39, Th. 2], and (ii) if the blocks of $\mathbf{X}$ were to have a circulant structure (along with $\tilde{N}_o = N$) then we could have used [23, Th. 5.1] to upper bound the probability that $\max_{i,j,i\neq j} |g_{i,j}| > \delta/S$.

B. Proof of Theorem 3

Similar to the proof of Theorem 2, we define $\delta = \delta_{2D}$ and $S = 2D$, and assume that $\mathcal{E} = 1$. Further, without loss of generality, we assume that N_t is odd and define $\tilde{N}_t = \frac{N_t-1}{2}$. Next, define $\mathbf{U}^t$ and $\mathbf{U}^f$ to be the N_t- and N_f-point discrete Fourier transform (DFT) matrices, respectively, with entries

$$U^t_{i,j} = \omega_{N_t}^{(\tilde{N}_t-j+1)(i-1)}, \quad i,j \in \{1,2,\ldots,N_t\}$$
$$U^f_{i,j} = \omega_{N_f}^{(j-1)(i-1)}, \quad i,j \in \{1,2,\ldots,N_f\}$$

and let $\mathbf{U}^{t,f} = \mathbf{U}^t \otimes \mathbf{U}^f$ be the $N_o \times N_o$ Kronecker product of the two DFT matrices (recall that $N_tN_f = N_o$). The key thing to note here is that since the Kronecker product of two orthogonal matrices is orthogonal, $\mathbf{U}^{t,f}$ is an orthogonal matrix ($\mathbf{U}^{t,f\,\dagger}\mathbf{U}^{t,f} = N_o\mathbf{I}_{N_o}$).

To proceed further, define $\mathcal{R} \subset \{1,2,\ldots,N_o\}$ as follows:

$$\mathcal{R} = \{i : i = nN_t + m + 1,\ (n,m) \in \mathcal{S}_r\}.$$

Notice that, by construction, $\mathcal{R}$ is a random set (due to the fact that $\mathcal{S}_r$ consists of N_r elements randomly selected from $\{0,1,\ldots,N_t-1\}\times\{0,1,\ldots,N_f-1\}$). Further, the cardinality of this set is $|\mathcal{R}| = N_r$ and it is equivalent in distribution to a random set of N_r points sampled uniformly at random from $\{1,2,\ldots,N_o\}$. Therefore, the matrix $\mathbf{U}^{t,f}_{|\mathcal{R}}$ obtained by retaining the *rows* of $\mathbf{U}^{t,f}$ corresponding to the indices in $\mathcal{R}$ is equivalent in distribution to a matrix obtained by randomly sampling N_r rows of $\mathbf{U}^{t,f}$. Consequently, from [37, Th. 3.3] (see also [36, Lem. 4.3]) and under the assumption that $N_r \geq c_4' \cdot \log^5 N_o \cdot S$, we have that $\frac{1}{\sqrt{N_r}}\mathbf{U}^{t,f}_{|\mathcal{R}}$ satisfies RIP of order S corresponding to any $\delta \in (0, \sqrt{2}-1)$ with probability exceeding $1 - c_5' N_0^{-c_6'}$.

Finally, it can be seen from the definition of RIP that if $\frac{1}{\sqrt{N_r}}\mathbf{U}^{t,f}_{|\mathcal{R}}$ satisfies RIP of order S for some δ then all of its *column* submatrices with number of columns $\geq S$ also satisfy RIP of order S with the same δ. The sensing matrix $\mathbf{U}$, however, is just a column submatrix of $\frac{1}{\sqrt{N_r}}\mathbf{U}^{t,f}_{|\mathcal{R}}$ (with $N > S$ columns) and this completes the proof of the theorem.

REFERENCES

[1] W. Kozek and A. Molisch, "Nonorthogonal pulseshapes for multicarrier communications in doubly dispersive channels," *IEEE J. Select. Areas Commun.*, pp. 1579–1589, Oct. 1998.

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

[3] X. Ma and G. Giannakis, "Maximum-diversity transmissions over doubly selective wireless channels," *IEEE Trans. Inform. Theory*, pp. 1832–1840, Jul. 2003.

[4] K. Liu, T. Kadous, and A. Sayeed, "Orthogonal time-frequency signaling over doubly dispersive channels," *IEEE Trans. Inform. Theory*, pp. 2583–2603, Nov. 2004.

[5] L. Tong, B. Sadler, and M. Dong, "Pilot-assisted wireless transmissions," *IEEE Signal Processing Mag.*, pp. 12–25, Nov. 2004.

[6] J. Cavers, "An analysis of pilot symbol assisted modulation for Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, pp. 686–693, Nov. 1991.

[7] M.-A. Baissas and A. Sayeed, "Pilot-based estimation of time-varying multipath channels for coherent CDMA receivers," *IEEE Trans. Signal Processing*, pp. 2037–2049, Aug. 2002.

[8] X. Ma, G. Giannakis, and S. Ohno, "Optimal training for block transmissions over doubly selective wireless fading channels," *IEEE Trans. Signal Processing*, pp. 1351–1366, May 2003.

[9] A. Kannu and P. Schniter, "Design and analysis of MMSE pilot-aided cyclic-prefixed block transmissions for doubly selective channels," *IEEE Trans. Signal Processing*, pp. 1148–1160, Mar. 2008.

[10] A. Sayeed, "Sparse multipath wireless channels: Modeling and implications," in *Proc. ASAP 2006*.

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

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

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

[14] V. Raghavan, G. Hariharan, and A. Sayeed, "Capacity of sparse multipath channels in the ultra-wideband regime," *IEEE J. Select. Topics Signal Processing*, pp. 357–371, Oct. 2007.

[15] S. Cotter and B. Rao, "The adaptive matching pursuit algorithm for estimation and equalization of sparse time-varying channels," in *Proc. 34th Asilomar Conf. Signals, Systems and Computers*, 2000.

[16] W. Li and J. Preisig, "Estimation of rapidly time-varying sparse channels," *IEEE J. Oceanic Eng.*, pp. 927–939, Oct. 2007.

[17] G. Tauböck and F. Hlawatsch, "A compressed sensing technique for OFDM channel estimation in mobile environments: Exploiting channel sparsity for reducing pilots," in *Proc. ICASSP 2008*.

[18] *IEEE Signal Processing Mag., Special Issue on Compressive Sampling*, vol. 25, no. 2, Mar. 2008.

[19] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, pp. 3397–3415, Dec. 1993.

[20] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, pp. 1207–1223, Mar. 2006.

[21] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, pp. 6–18, Jan. 2006.

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

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

[24] M. Herman and T. Strohmer, "High resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, submitted.

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

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

[27] A. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Processing*, pp. 2563–2579, Oct. 2002.

[28] A. Sayeed and V. Veeravalli, "The essential degrees of freedom in space-time fading channels," in *Proc. PIMRC 2002*.

[29] J. Proakis, *Digital Communications*. McGraw-Hill, 2001.

[30] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

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

[32] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," *Ann. Statist.*, in press.

[33] M. 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.

[34] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," in *Proc. ICASSP 2008*.

[35] R. Negi and J. Cioffi, "Pilot tone selection for channel estimation in a mobile OFDM system," *IEEE Trans. Consumer Electron.*, pp. 1122–1128, Aug. 1998.

[36] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, pp. 5406–5425, Dec. 2006.

[37] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, pp. 1025–1045, Aug. 2008.

[38] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[39] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," in *Proc. CISS 2008*.