A Restricted Isometry Property for Structurally-Subsampled Unitary Matrices

Waheed U. Bajwa[†,‡], Akbar M. Sayeed[‡], and Robert Nowak[‡]
[†]The Program in Applied and Computational Mathematics, Princeton University
[‡]Department of Electrical and Computer Engineering, University of Wisconsin-Madison
wbajwa@math.princeton.edu, akbar@engr.wisc.edu, nowak@engr.wisc.edu

Abstract—Subsampled (or partial) Fourier matrices were originally introduced in the compressive sensing literature by Candès et al. Later, in papers by Candès and Tao and Rudelson and Vershynin, it was shown that (random) subsampling of the rows of many other classes of unitary matrices also yield effective sensing matrices. The key requirement is that the rows of U, the unitary matrix, must be highly incoherent with the basis in which the signal is sparse. In this paper, we consider acquisition systems that—despite sensing sparse signals in an incoherent domain— cannot randomly subsample rows from U. We consider a general class of systems in which the sensing matrix corresponds to subsampling of the rows of matrices of the form $\mathbf{\Phi} = \mathbf{RU}$ (instead of U), where R is typically a low-rank matrix whose structure reflects the physical/technological constraints of the acquisition system. We use the term "structurally-subsampled unitary matrices" to describe such sensing matrices. We investigate the restricted isometry property of a particular class of structurally-subsampled unitary matrices that arise naturally in application areas such as multiple-antenna channel estimation and sub-nyquist sampling. In addition, we discuss an immediate application of this work in the area of wireless channel estimation, where the main results of this paper can be applied to the estimation of multiple-antenna orthogonal frequency division multiplexing channels that have sparse impulse responses.

I. INTRODUCTION

A. Background

In a nutshell, the theory of sparse signal recovery—or compressive sensing (CS), as it is commonly called today— deals with recovering a signal $\mathbf{x} \in \mathbb{C}^p$ from linear observations of the form

$$\mathbf{y} = \mathbf{Ax} \quad : \quad \|\mathbf{x}\|_0 \leq S \tag{1}$$

where $\|\mathbf{x}\|_0$ counts the number of nonzero entries in $\mathbf{x}$ and $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a known matrix. This mathematical model corresponds to a nonadaptive measurement process that senses an *S-sparse signal* $\mathbf{x}$ by taking n linear measurements of the signal and the goal here is to reliably recover $\mathbf{x}$ from knowledge of the *observation vector* $\mathbf{y}$ and the *sensing matrix* $\mathbf{A}$. Fundamentally, however, the theory of CS deals with the special case of $n \ll p$—which arises in many data-starved inverse problems in a number of application areas—and attempts to answer the following questions [1], [2]: (i) What conditions does $\mathbf{A}$ need to satisfy to ensure successful recovery of a sparse $\mathbf{x}$? (ii) Can (1) be reliably solved for $\mathbf{x}$ in practice using polynomial-time solvers? and (iii) What performance guarantees can be given for various practical solvers when $\mathbf{y}$ is corrupted by either stochastic noise or deterministic perturbation?

A number of researchers have successfully addressed these questions, and their extensions to less restrictive notions of sparsity, over the past 30 years or so. In particular, the celebrated success of CS theory can primarily be attributed to some of the recent research breakthroughs that established that a signal $\mathbf{x}$ that is either S-sparse or approximately S-sparse can be reliably and efficiently reconstructed from (noiseless or noisy) $\mathbf{y}$ by making use of (i) an appropriately designed sensing matrix with a relatively small number of rows—typically much smaller than p—and (ii) tractable linear optimization programs, efficient greedy algorithms, or fast iterative thresholding methods. Proofs of these remarkable results all rely in some sense on the same property of the sensing matrix, namely that every collection of $2S$ columns of (appropriately normalized) $\mathbf{A}$ should behave almost like an isometry. One concise way to state this condition is through the *restricted isometry property* (RIP), first introduced in [3].

Definition 1: An $n \times p$ matrix $\mathbf{A}$ having unit ℓ_2-norm columns is said to have the RIP of order S with parameter δ_S if there exists some $\delta_S \in (0, 1)$ such that

$$(1 - \delta_S)\|\tilde{\mathbf{x}}\|_2^2 \leq \|\mathbf{A}\tilde{\mathbf{x}}\|_2^2 \leq (1 + \delta_S)\|\tilde{\mathbf{x}}\|_2^2 \tag{2}$$

holds for all S-sparse vectors $\tilde{\mathbf{x}}$. In this case, we sometimes make use of the shorthand notation $\mathbf{A} \in RIP(S, \delta_S)$ to state that $\mathbf{A}$ satisfies the RIP of order S with parameter δ_S.

It is easy to see from (2) that the RIP of order S is essentially a statement about the singular values of all $n \times S$ submatrices of $\mathbf{A}$. And while no algorithms are known to date that can check the RIP for a given matrix in polynomial time, one of the reasons that has led to the widespread applicability of CS theory in various application areas is the revelation that certain probabilistic constructions of matrices satisfy the RIP with high probability. For example, it has been established in [4] that if the entries of an $n \times p$ matrix $\mathbf{A}$ are drawn independently from a $\mathcal{N}(0, \frac{1}{n})$ distribution then $\mathbf{A} \in RIP(S, \delta_S)$ with probability exceeding $1 - e^{-O(n)}$ for every $\delta_S \in (0, 1)$ provided $n = \Omega(S \log \frac{p}{S})$. Similarly, consider an $n \times p$ subsampled unitary matrix $\mathbf{A}$ obtained by first randomly selecting n rows of a $p \times p$ unitary matrix $\mathbf{U}$ and then normalizing them so that the resulting columns of $\mathbf{A}$ have unit ℓ_2-norms. Then it has been shown in [5], [6] that $\mathbf{A} \in RIP(S, \delta_S)$ with probability exceeding $1 - p^{-O(\delta_S^2)}$

for every $\delta_S \in (0, 1)$ provided $n = \Omega(\mu_{\mathbf{U}}^2 S \log^5 p)$, where $\mu_{\mathbf{U}} \overset{def}{=} \sqrt{p} \max_{i,j} |u_{i,j}|$ is termed as the *coherence* of the unitary matrix $\mathbf{U}$.

B. Structurally-Subsampled Unitary Matrices

Subsampled unitary matrices were originally introduced—and partially analyzed—in the modern CS literature in [7]. Initially, the focus in [7] was on sensing matrices that corresponded to subsampled (or partial) Fourier matrices. Later, the analysis was advanced further by Candès and Tao and Rudelson and Vershynin in [5] and [6], respectively, where they established—among other things—that (random) subsampling of the rows of many other classes of unitary matrices also yield effective sensing matrices. Together, the results of [5]–[7] form the basis of the so-called *principle of incoherent measurements*, stated as follows.

> It is best to acquire samples of a sparse signal $\mathbf{x}$ in a maximally incoherent transform domain $\mathbf{U}$, where the incoherence is measured by the coherence parameter $\mu_{\mathbf{U}}$—the smaller the coherence parameter, the greater the incoherence.[1]

This principle is made mathematically precise in [5], [6] by stating that a (randomly) subsampled unitary matrix $\mathbf{A}$ needs to have $n = \Omega(\mu_{\mathbf{U}}^2 S \log^5 p)$ rows in order for it to satisfy the RIP of order S with high probability. Note that since $\mu_{\mathbf{U}} = \sqrt{p} \max_{i,j} |u_{i,j}|$, we have that (i) the coherence of a unitary matrix cannot be smaller than 1, and (ii) unitary matrices with entries of magnitude $O(1/\sqrt{p})$ are maximally incoherent. In other words, transform domains such as Fourier, composition of Fourier and wavelet, and Hadamard are all maximally incoherent and are, therefore, particularly well-suited for acquisition of sparse signals.

It turns out that a number of real-world signal acquisition systems already adhere to the principle of incoherent measurements due to various physical/technological reasons. For example, data acquired by magnetic resonance imaging scanners naturally correspond to Fourier-domain samples of the object being imaged [8]. Similarly, channel measurements collected by a communications receiver using multicarrier modulation inherently correspond to Fourier-domain samples of the single-antenna channel being estimated [9]. As such, there is a natural fit between the theory of subsampled unitary matrices and these two applications, as noted in, e.g., [8], [9].

Contrary to these examples, however, our interest in this paper is in acquisition systems that—despite sensing sparse signals in an incoherent domain—cannot sample *individual* coefficients in the transform domain. This indeed happens in a number of real-world systems because of a multitude of physical constraints and/or technological limitations. For example, the impulse response of a multiple-antenna channel generally lives in a three-dimensional (3-d) space but a communications receiver using multicarrier modulation can only acquire

2-d projections of its 3-d Fourier-domain samples (physical constraint) [10]. Similarly, it is generally desirable to project an ultrawideband signal with limited spectral content onto a smaller spectral band before sampling it since randomized sampling to acquire the signal can be very sensitive to timing errors (technological constraint) [11].

In the parlance of CS, the sensing matrices in both the aforementioned cases now correspond to subsampling of the rows of matrices of the form $\mathbf{\Phi} \overset{def}{=} \mathbf{RU}$ (instead of $\mathbf{U}$), where $\mathbf{R}$ is typically a low-rank matrix whose structure reflects the physical and/or technological constraints of the acquisition system and $\mathbf{U}$ is the transform domain (unitary) matrix. We use the term *structurally-subsampled unitary matrices* for such sensing matrices so as to distinguish them from the canonical subsampled unitary matrices studied in [5]–[7] and formally define these matrices as follows.

Definition 2: Let $\mathbf{U}$ be a $p \times p$ unitary matrix, $m \leq p$ be an integer parameter, and $\mathbf{R}$ be an $m \times p$ *row-mixing matrix*, given by

$$\mathbf{R} \overset{def}{=} \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \dots & \mathbf{r}_m \end{bmatrix}^T. \tag{3}$$

Next, choose a subset Ω of cardinality $n \overset{def}{=} |\Omega|$ uniformly at random (without replacement) from the set $\{1, \dots, m\}$. Then the structurally-subsampled unitary matrix $\mathbf{A}$ generated from $(\mathbf{R}, \mathbf{U})$ is a submatrix of $\mathbf{\Phi} = \mathbf{RU}$ obtained by selecting n rows of $\mathbf{\Phi}$ corresponding to the indices in Ω and normalizing the resulting columns so that they have unit ℓ_2-norms.

Remark 1: A structurally-subsampled unitary matrix $\mathbf{A}$ generated from $(\mathbf{R}, \mathbf{U})$ can also be thought of as a twice-subsampled version of the unitary matrix $\mathbf{U}$. Here, the first subsampling step corresponds to obtaining an $m \times p$ matrix $\mathbf{\Phi}$ from $\mathbf{U}$ by projecting the p columns of $\mathbf{U}$ onto the m rows of $\mathbf{R}$. In contrast, the second subsampling step corresponds to obtaining $\mathbf{A}$ from $\mathbf{\Phi}$ by randomly selecting (and appropriately normalizing) n rows of $\mathbf{\Phi}$.

C. Main Result

It is easy to see from Definition 2 that the theory of subsampled unitary matrices is not easily extendable to structurally-subsampled unitary matrices, except for the trivial case of $\mathbf{R}$ being a (square) diagonal matrix. In this paper, we investigate the restricted isometry property of a particular class of structurally-subsampled unitary matrices that arise naturally in application areas such as multiple-antenna channel estimation and sub-nyquist sampling. Specifically, let $k \in \{1, \dots, p\}$ be a parameter that is an integer factor of p and define $m \overset{def}{=} p/k$. Further, let $\mathcal{A}_p \overset{def}{=} \{a_i \in \mathbb{C}\}_{i=1}^p$ denote a p-length generating sequence and define the m rows of $\mathbf{R}$ to be generated from the sequence $\mathcal{A}_p$ as follows

$$\mathbf{r}_i^T \overset{def}{=} \begin{bmatrix} \underbrace{0 \ \dots \ 0}_{(i-1)k \text{ terms}} & a_{ik-k+1} & \dots & a_{ik} & \underbrace{0 \ \dots \ 0}_{p-ik \text{ terms}} \end{bmatrix}. \tag{4}$$

In other words, the row-mixing matrix $\mathbf{R}$ has a block-diagonal structure. Then the main result of this paper—stated in terms

---

[1]The coherence parameter gets its name from the fact that we can write $\mu_{\mathbf{U}} = \sqrt{p} \max_{i,j} |\langle \mathbf{u}_i, \mathbf{e}_j \rangle|$, where $\mathbf{u}_i$ denotes the i-th column of $\mathbf{U}^{\mathsf{H}}$ and $\mathbf{e}_j$ denotes the j-th column of the canonical basis $\mathbf{I}_p$.

of the following theorem—asserts that structurally-subsampled unitary matrices generated from $(\mathbf{R}, \mathbf{U})$ satisfy RIP for the nontrivial case of $k > 1$ when $\mathcal{A}_p$ is a Rademacher sequence.[2]

Theorem 1: Let the elements of the generating sequence $\mathcal{A}_p = \{a_i\}_{i=1}^p$ be independent realizations of Rademacher random variables taking values ± 1 with probability $1/2$. Further, let $\mathbf{R}$ be the $m \times p$ row-mixing matrix whose rows are generated by the sequence $\mathcal{A}_p$ according to (4), where $m = p/k$ for a parameter $k \in \{1, \ldots, p\}$ that is an integer factor of p. Choose a subset Ω of cardinality $n = |\Omega|$ uniformly at random (without replacement) from the set $\{1, \ldots, m\}$. Finally, let $\mathbf{U}$ be any $p \times p$ unitary matrix, and let $\mathbf{A}$ be the $n \times p$ matrix obtained by sampling n rows of $\mathbf{\Phi} = \mathbf{RU}$ corresponding to the indices in Ω and normalizing the resulting columns by $\sqrt{m/n}$. Then for each integer $p, S > 2$, and for any $z > 1$ and any $\delta_S \in (0, 1)$, there exist absolute (positive) constants c_1 and c_2 such that whenever

$$n \geq c_1 z \mu_{\mathbf{U}}^2 S \log^3 p \log^2 S \qquad (5)$$

the matrix $\mathbf{A} \in RIP(S, \delta_S)$ with probability exceeding $1 - 20 \max \left\{ \exp\left(-c_2 \delta_S^2 z\right), p^{-1} \right\}$.

Remark 2: One of the main advantages of describing the structurally-subsampled unitary matrix $\mathbf{A}$ of Theorem 1 as a subsampled version of $\mathbf{\Phi}$ is that it allows us to borrow some of the mathematical techniques used by Rudelson and Vershynin in [6] to establish the RIP of canonical subsampled unitary matrices. Nevertheless, construction of the sensing matrix $\mathbf{A}$ of Theorem 1 can equivalently be understood in the followsing sense: Divide the $p \times p$ unitary matrix $\mathbf{U}$ into m contiguous blocks of $k = p/m$ rows each and select n of these blocks corresponding to the indices in the set Ω. Then every row of $\mathbf{A}$ corresponds to a (random) superposition of the k rows of one of these selected blocks. In fact, this interpretation of the structurally-subsampled unitary matrices of Theorem 1 is what enables us to use the results of this theorem later in the context of estimation of multiple-antenna orthogonal frequency division multiplexing channels.

Before proceeding with the proof of this theorem, however, let us introduce some notation (originally used in [6]) that will greatly facilitate the mathematical analysis in the sequel. Specifically, it can be easily verified from (2) that an $n \times p$ matrix $\mathbf{A} \in RIP(S, \delta_S)$ if the following inequality holds for some constant $\delta_S \in (0, 1)$

$$\max_{\substack{T \subset \{1, \ldots, p\} \\ |T| \leq S}} \left\| \mathbf{A}_T^{\mathsf{H}} \mathbf{A}_T - \mathbf{I}_{|T|} \right\|_2 \leq \delta_S \qquad (6)$$

where $\| \cdot \|_2$ denotes the spectral norm of a matrix (the largest singular value of the matrix), and $\mathbf{A}_T$ denotes an $n \times |T|$ submatrix of $\mathbf{A}$ obtained by collecting all the columns of $\mathbf{A}$ corresponding to the indices in set T. This expression can be

further written in a compact form with the help of a non-negative function $\| \cdot \|_{T,S} : \mathbb{C}^{p \times p} \to [0, \infty)$ defined as follows

$$\|\mathbf{M}\|_{T,S} \stackrel{def}{=} \max_{\substack{T \subset [1 \ldots p] \\ |T| \leq S}} \left\| \mathbf{M}_{T \times T} \right\|_2 \qquad (7)$$

where $\mathbf{M}_{T \times T}$ denotes a $|T| \times |T|$ submatrix of $\mathbf{M}$ obtained by collecting all the entries of $\mathbf{M}$ corresponding to the indices in set $T \times T$. Going back to the definition of RIP, we can therefore alternatively state that an $n \times p$ matrix $\mathbf{A} \in RIP(S, \delta_S)$ for some constant $\delta_S \in (0, 1)$ if

$$\|\mathbf{A}^{\mathsf{H}} \mathbf{A} - \mathbf{I}_p\|_{T,S} \leq \delta_S \qquad (8)$$

and we will prove this inequality in the sequel for the structurally-subsampled unitary matrices of Theorem 1.

D. Organization

The rest of this paper is organized as follows. In Section II, we provide a proof of Theorem 1 using tools from the classical theory of probability in Banach spaces. In Section III, we discuss an application of Theorem 1 in the area of estimation of multiple-antenna channels that have sparse impulse responses. Finally, in Section IV, we heuristically compare the performance of structurally-subsampled unitary matrices to that of canonical subsampled unitary matrices and discuss the connections between the results of this paper and some existing works.

II. PROOF OF THE MAIN RESULT

It is a trivial exercise to verify that $\| \cdot \|_{T,S}$ defines a norm—which we term as (T, S)-norm—on the vector space $\mathbb{C}^{p \times p}$. Therefore, the matrix $(\mathbf{A}^{\mathsf{H}} \mathbf{A} - \mathbf{I}_p)$ lives in the Banach space $\mathcal{B} \stackrel{def}{=} (\mathbb{C}^{p \times p}, \| \cdot \|_{T,S})$ and the main tools that we use to establish the inequality in (8) come from the classical theory of probability in Banach spaces [12]. The general roadmap for our proof is very similar to [6, Theorem 3.3], which is now a well-established technique in the CS literature for establishing RIP of subsampled matrices [11], [13]. In particular, the proof relies heavily on an upper bound on expected (T, S)-norm of sum of independent rank-one matrices that was established in [6, Lemma 3.8]. In the following, we describe the basic steps taken to establish a formal proof of our stated claim.

First, we initially assume that—instead of the uniformly-at-random sampling model—the structurally-subsampled unitary matrix $\mathbf{A}$ in Theorem 1 is generated from $(\mathbf{R}, \mathbf{U})$ according to a Bernoulli sampling model. That is, let $\zeta_1, \ldots, \zeta_m$ be independent Bernoulli random variables taking the value 1 with probability n/m. Then,

$$\Omega \stackrel{def}{=} \{i : \zeta_i = 1\} \qquad (9)$$

and $\mathbf{A}$ is a (normalized) $|\Omega| \times p$ submatrix of $\mathbf{\Phi} = \mathbf{RU}$ obtained by sampling $|\Omega|$ rows of $\mathbf{\Phi}$ corresponding to the indices in Ω and normalizing the resulting columns by $\sqrt{m/n}$. We then have the following lemma that shows that under this assumption the Gram matrix $\mathbf{A}^{\mathsf{H}} \mathbf{A} = \mathbf{I}_p$ in expectation.

Lemma 1: Let the structurally-subsampled unitary matrix $\mathbf{A}$ in Theorem 1 be generated from $(\mathbf{R}, \mathbf{U})$ according to a Bernoulli sampling model (as described above). Then,

$$\mathbb{E}[\mathbf{A}^{\mathsf{H}}\mathbf{A}] = \mathbf{I}_p. \tag{10}$$

Proof: See the Appendix. ∎

Second, we use Lemma 1 to establish that $\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S}$ cannot be too large in expectation for large-enough values of n in the case of a Bernoulli sampling model. The proof of this result, however, is a little more involved and makes use of a number of auxiliary lemmas. The most important lemma that we will need in this regard is the following one, due to Rudelson and Vershynin [6, Lemma 3.8].

Lemma 2 (Rudelson–Vershynin): Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$, $r \leq p$, be vectors in $\mathbb{C}^p$ with uniformly bounded entries, $\|\mathbf{v}_i\|_\infty \leq K$ for all i. Further, let $\{\varepsilon_i\}$ be independent Rademacher random variables taking values ± 1 with probability $1/2$. Then

$$\mathbb{E}\left[\left\|\sum_{i=1}^{r}\varepsilon_i \mathbf{v}_i \mathbf{v}_i^{\mathsf{H}}\right\|_{T,S}\right] \leq B(r) \cdot \left\|\sum_{i=1}^{r}\mathbf{v}_i \mathbf{v}_i^{\mathsf{H}}\right\|_{T,S}^{1/2} \tag{11}$$

where $B(r) \stackrel{def}{=} c_3 K\sqrt{S}\log(S)\sqrt{\log p}\sqrt{\log r}$ for some absolute constant $c_3 > 0$.

In order to make use of Lemma 2, however, we require the entries of $\mathbf{A}$ to be uniformly bounded by some number K. To this end, we will make use of the classical Khintchine inequality for independent and identically distributed (i.i.d.) Rademacher random variables [12, Lemma 4.1].

Lemma 3 (Khintchine Inequality): Let $\{\varepsilon_i\}$ be independent Rademacher random variables taking values ± 1 with probability $1/2$. For any $s \in (0, \infty)$, there exist a positive finite constant C_s depending on s only such that for any finite sequence $\{\alpha_i\}$ of complex numbers

$$\left(\mathbb{E}\left[\left|\sum_i \varepsilon_i \alpha_i\right|^s\right]\right)^{1/s} \leq C_s\left(\sum_i |\alpha_i|^2\right)^{1/2}. \tag{12}$$

For the case of real numbers, it has been established by Haagerup in [14] that the best constant C_s in (12) is

$$C_s^* \stackrel{def}{=} \begin{cases} 1, & \text{if } 0 < s \leq 2, \\ 2^{1/2}\left(\frac{\Gamma((s+1)/2)}{\sqrt{\pi}}\right)^{1/s}, & \text{if } 2 < s < \infty, \end{cases} \tag{13}$$

where $\Gamma(z) \stackrel{def}{=} \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function. However, it is trivial to verify that C_s^* is also a valid constant in the case of complex numbers, since if the upper bound in the Khintchine inequality holds for real numbers with some constant then it also holds for complex numbers with the same constant. We are now ready to prove that the entries of the structurally-subsampled unitary matrix $\mathbf{A}$ cannot be too large in the case of a Bernoulli sampling model.

Lemma 4: Let the structurally-subsampled unitary matrix $\mathbf{A}$ in Theorem 1 be generated from $(\mathbf{R}, \mathbf{U})$ according to a Bernoulli sampling model (as described earlier). Then for any

integer $p > 2$ and any $r \in [2, 2\log p]$, we have

$$\left(\mathbb{E}\left[\|\mathbf{A}\|_{\max}^r\right]\right)^{1/r} \leq \sqrt{\frac{m}{n}}\left(\mathbb{E}\left[\|\mathbf{\Phi}\|_{\max}^r\right]\right)^{1/r} \leq \sqrt{\frac{16\mu_{\mathbf{U}}^2\log p}{n}} \tag{14}$$

where $\|\cdot\|_{\max}$ denotes the max norm of a matrix (absolute value of the largest-magnitude entry of the matrix).

Proof: The proof of this lemma is very similar to that of [11, Lemma 5] and is, therefore, omitted here for the sake of brevity (alternatively, see [15, Lemma 3.13]). ∎

All the pieces are now in place to bound $\mathbb{E}\left[\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S}\right]$ in the case of a Bernoulli sampling model using Lemma 2 and techniques developed in probability in Banach spaces [12].

Lemma 5: Let the structurally-subsampled unitary matrix $\mathbf{A}$ in Theorem 1 be generated from $(\mathbf{R}, \mathbf{U})$ according to a Bernoulli sampling model. Then for any integer $p > 2$ and any $\epsilon \in (0, 1)$, we have

$$\mathbb{E}\left[\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S}\right] \leq \epsilon \tag{15}$$

provided the number of rows $n \geq c_4\,\epsilon^{-2}\mu_{\mathbf{U}}^2\,S\log^3 p\log^2 S$ for some absolute constant $c_4 > 0$.

Proof: The proof of this lemma can be found in [15, Lemma 3.14]. ∎

Finally, we show that $\|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S}$ concentrates around its mean with high probability. To establish this fact, however, we need one additional classical result from the theory of probability in Banach spaces. The following result is originally due to Ledoux and Talagrand [12, Theorem 6.17] and appears in the following form in [6, Theorem 3.10].

Theorem 2 (Ledoux–Talagrand): Let $\mathcal{B} \stackrel{def}{=} (X, \|\cdot\|_X)$ be a Banach space. Further, let $\{Y_i\}_{i=1}^N$ be independent, symmetric random variables in $\mathcal{B}$ such that $\|Y_i\|_X \leq B$ for every i almost surely. Finally, define $Y \stackrel{def}{=} \|\sum_{i=1}^N Y_i\|_X$. Then for any integers $r \geq q$, any $t > 0$, and some absolute constant $c_5 > 0$, Y satisfies

$$\Pr\left(Y \geq 8q\mathbb{E}[Y] + 2rB + t\right) \leq \left(\frac{c_5}{q}\right)^r + \\ + 2\exp\left(-\frac{t^2}{256q\mathbb{E}[Y]^2}\right). \tag{16}$$

We are now ready to establish the RIP for structurally-subsampled unitary matrices described in Theorem 1.

Proof of Theorem 1: We begin by recalling the result established in [7, Section 2.3], which states that if it can be shown that subsampled matrices in a particular class satisfy the RIP with probability exceeding $1 - \eta$ for the Bernoulli sampling model, then it follows that subsampled matrices belonging to the same class satisfy the RIP with probability exceeding $1 - 2\eta$ for the uniformly-at-random sampling model. As such, we begin by assuming that the structurally-subsampled unitary matrix $\mathbf{A}$ is generated from $(\mathbf{R}, \mathbf{U})$ according to a Bernoulli sampling model.

Next, consider the Banach space $\mathcal{B} \stackrel{def}{=} (\mathbb{C}^{p \times p}, \|\cdot\|_{T,S})$ and define random variables $\{\mathbf{Y}_i\}_{i=1}^p$ and $\{\widehat{\mathbf{Y}}_i\}_{i=1}^p$ that take

values in $\mathcal{B}$ as follows

$$\mathbf{Y}_i \stackrel{def}{=} \frac{m}{n}\zeta_i\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathsf{H}} - \frac{1}{p}\mathbf{I}_p,$$

$$\widetilde{\mathbf{Y}}_i \stackrel{def}{=} \frac{m}{n}\Big(\zeta_i\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathsf{H}} - \zeta_i'\boldsymbol{\phi}_i'\boldsymbol{\phi}_i'^{\mathsf{H}}\Big), \quad i = 1, \ldots, p \quad (17)$$

where $\{\zeta_i\}$ are the Bernoulli random variables arising in the Bernoulli sampling model, $\{\boldsymbol{\phi}_i^{\mathsf{H}}\}$ denote the rows of $\boldsymbol{\Phi} = \mathbf{RU}$, and $\{\zeta_i'\}$ and $\{\boldsymbol{\phi}_i'^{\mathsf{H}}\}$ are independent copies of $\{\zeta_i\}$ and $\{\boldsymbol{\phi}_i^{\mathsf{H}}\}$, respectively. In other words, each random variable $\widetilde{\mathbf{Y}}_i \stackrel{def}{=} \mathbf{Y}_i - \mathbf{Y}_i'$ is a symmetric version of the corresponding random variable $\mathbf{Y}_i$, where $\mathbf{Y}_i'$ denotes an independent copy of $\mathbf{Y}_i$. In particular, we have that $\sum_{i=1}^{p}\widetilde{\mathbf{Y}}_i$ is a symmetric version of $\sum_{i=1}^{p}\mathbf{Y}_i$ and, therefore, the following symmetrization inequalities hold for all $u > 0$ [12, Chapter 6]

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^{p}\widetilde{\mathbf{Y}}_i\Big\|_{T,S}\Big] \leq 2\mathbb{E}\Big[\Big\|\sum_{i=1}^{p}\mathbf{Y}_i - \mathbb{E}\Big[\sum_{i=1}^{p}\mathbf{Y}_i\Big]\Big\|_{T,S}\Big], \quad (18)$$

$$\Pr\left(\Big\|\sum_{i=1}^{p}\mathbf{Y}_i\Big\|_{T,S} > 2\mathbb{E}\Big[\Big\|\sum_{i=1}^{p}\mathbf{Y}_i\Big\|_{T,S}\Big] + u\right) \leq$$
$$2\Pr\left(\Big\|\sum_{i=1}^{p}\widetilde{\mathbf{Y}}_i\Big\|_{T,S} > u\right). \quad (19)$$

There are two key observations that can be made here. First, we can bound the expected value of $\widetilde{\mathbf{Y}} \stackrel{def}{=} \|\sum_{i=1}^{p}\widetilde{\mathbf{Y}}_i\|_{T,S}$ using (18) and Lemma 5 since (i) $\mathbb{E}\big[\sum_{i=1}^{p}\mathbf{Y}_i\big] = 0$ (Lemma 1), and (ii) $\mathbf{Y} \stackrel{def}{=} \|\sum_{i=1}^{p}\mathbf{Y}_i\|_{T,S} = \|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S}$. Second, we can obtain a large-deviation bound for $\mathbf{Y}$ using (19) and Theorem 2 since—by construction—$\{\widetilde{\mathbf{Y}}_i\}_{i=1}^{p}$ are independent, symmetric random variables in $\mathcal{B}$. Before can use Thereom 2 to characterize the tail behavior of $\widetilde{\mathbf{Y}}$, however, we need to establish that $\max_i \|\widetilde{\mathbf{Y}}_i\|_{T,S} \leq B$ for some B.

Towards this end, we first (rather trivially) establish that $\max_i \{\sqrt{\frac{m}{n}}\|\boldsymbol{\phi}_i^{\mathsf{H}}\|_\infty, \sqrt{\frac{m}{n}}\|\boldsymbol{\phi}_i'^{\mathsf{H}}\|_\infty\}$ cannot be too large with high probability. Specifically, note from Lemma 4 that we have for $r = 2\log p$

$$\Pr\left(\sqrt{\frac{m}{n}}\|\boldsymbol{\Phi}\|_{\max} > \sqrt{\frac{16\,e\mu_{\mathbf{U}}^2\log p}{n}}\right) \stackrel{(a)}{\leq}$$
$$\frac{\mathbb{E}\big[\|\boldsymbol{\Phi}\|_{\max}^r\big]}{e^{r/2}\cdot\mathbb{E}\big[\|\boldsymbol{\Phi}\|_{\max}^r\big]} = p^{-1} \quad (20)$$

where (a) follows from an application of Markov's inequality (see also [11, Lemma 5]). Next, define $B_1 \stackrel{def}{=} \frac{16\,e\mu_{\mathbf{U}}^2\log p}{n}$. Then we have from (20) that

$$\Pr\left(\Big\{\sqrt{\frac{m}{n}}\|\boldsymbol{\Phi}\|_{\max} > \sqrt{B_1}\Big\}\bigcup\right.$$
$$\left.\Big\{\sqrt{\frac{m}{n}}\|\boldsymbol{\Phi}'\|_{\max} > \sqrt{B_1}\Big\}\right) \stackrel{(b)}{\leq} 2p^{-1} \quad (21)$$

where $\boldsymbol{\Phi}'$ is comprised of $\{\boldsymbol{\phi}_i'^{\mathsf{H}}\}$ as its rows (in other words, $\boldsymbol{\Phi}'$ is an independent copy of $\boldsymbol{\Phi}$), and (b) follows from a simple union bounding argument. Further, we also have

$$\max_i \big\|\widetilde{\mathbf{Y}}_i\big\|_{T,S} \stackrel{(c)}{\leq} \max_i \left\{\Big\|\frac{m}{n}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathsf{H}}\Big\|_{T,S} + \Big\|\frac{m}{n}\boldsymbol{\phi}_i'\boldsymbol{\phi}_i'^{\mathsf{H}}\Big\|_{T,S}\right\}$$
$$\stackrel{(d)}{\leq} S\Big(\frac{m}{n}\|\boldsymbol{\Phi}\|_{\max}^2 + \frac{m}{n}\|\boldsymbol{\Phi}'\|_{\max}^2\Big) \quad (22)$$

where (c) mainly follows from triangle inequality, and (d) is a simple consequence of the definition of (T, S)-norm and the fact that $\|\boldsymbol{\Phi}\|_{\max} \stackrel{def}{=} \max_i \|\boldsymbol{\phi}_i^{\mathsf{H}}\|_\infty$ (and in the same way, $\|\boldsymbol{\Phi}'\|_{\max} \stackrel{def}{=} \max_i \|\boldsymbol{\phi}_i'^{\mathsf{H}}\|_\infty$). It is then easy to see from (20) and (22) that we have $\max_i \|\widetilde{\mathbf{Y}}_i\|_{T,S} \leq 2SB_1$ with probability exceeding $1 - 2p^{-1}$.

Finally, define the event $E \stackrel{def}{=} \{\max_i \|\widetilde{\mathbf{Y}}_i\|_{T,S} \leq 2SB_1\}$. Then, conditioned on this event, we have from (18), Lemma 5 and Theorem 2 that whenever $n \geq c_4\,\epsilon^{-2}\mu_{\mathbf{U}}^2\,S\log^3 p\log^2 S$

$$\Pr\left(\widetilde{\mathbf{Y}} \geq 16q\epsilon + 4rSB_1 + t\big|E\right) < \left(\frac{c_5}{q}\right)^r +$$
$$+ 2\exp\left(-\frac{t^2}{1024q\epsilon^2}\right) \quad (23)$$

for any integer $r \geq q$, any $t > 0$, and any $\epsilon \in (0, 1)$. Next, choose $q = \lceil ec_5\rceil$, $t = 32\sqrt{q}\eta\epsilon$, and $r = \lceil\frac{t}{2SB_1}\rceil$ for some $\eta > 1$. Further, define a new constant $c_1 \stackrel{def}{=} \max\{e\sqrt{q}, c_4\}$ and let $n \geq c_1\,\epsilon^{-2}\mu_{\mathbf{U}}^2\,S\log^3 p\log^2 S$. Note that this choice of n ensures $r \geq q$, resulting in

$$\Pr\left(\widetilde{\mathbf{Y}} \geq (16q + 96\sqrt{q}\eta)\epsilon\big|E\right) < \exp\left(-\frac{\sqrt{q}\eta\epsilon n}{3\mu_{\mathbf{U}}^2 S\log p}\right) +$$
$$+ 2\exp\left(-\eta^2\right). \quad (24)$$

We can now get rid of the conditioning in the above expression by noting that $\Pr(E^c) \leq 2p^{-1}$, which in turn implies

$$\Pr\left(\widetilde{\mathbf{Y}} \geq (16q + 96\sqrt{q}\eta)\epsilon\right) < \exp\left(-\frac{\sqrt{q}\eta\epsilon n}{3\mu_{\mathbf{U}}^2 S\log p}\right) +$$
$$+ 2\exp\left(-\eta^2\right) + 2p^{-1}. \quad (25)$$

In the end, what remains to be shown is that $\mathbf{Y} = \|\sum_{i=1}^{p}\mathbf{Y}_i\|_{T,S} = \|\mathbf{A}^{\mathsf{H}}\mathbf{A} - \mathbf{I}_p\|_{T,S} \leq \delta_S$ with high probability. To this end, note that if $n \geq c_1\,\epsilon^{-2}\mu_{\mathbf{U}}^2\,S\log^3 p\log^2 S$ then $\mathbb{E}[\mathbf{Y}] \leq \epsilon$ from Lemma 5. Consequently, we get from (19) and (25) that

$$\Pr\left(\mathbf{Y} \geq (2 + 16q + 96\sqrt{q}\eta)\epsilon\right) < 2\exp\left(-\frac{\sqrt{q}\eta\epsilon n}{3\mu_{\mathbf{U}}^2 S\log p}\right)$$
$$+ 4\exp\left(-\eta^2\right) + 4p^{-1}. \quad (26)$$

Finally, define $c_6 \stackrel{def}{=} (2 + 16q + 96\sqrt{q})$ and note that $c_6\eta\,\epsilon > (2 + 16q + 96\sqrt{q}\eta)\epsilon$ since $\eta > 1$. If we now choose $\eta = \frac{\delta_S}{c_6\epsilon}$ then $\frac{\sqrt{q}\eta\epsilon n}{3\mu_{\mathbf{U}}^2 S\log p} > \eta^2$ and, therefore, (26) can be simplified as

$$\Pr\left(\mathbf{Y} \geq \delta_S\right) < 10\max\left\{\exp\left(-c_2\delta_S^2 z\right), p^{-1}\right\} \quad (27)$$

where $c_2 \stackrel{def}{=} 1/c_6$ and $z \stackrel{def}{=} 1/\epsilon^2$. The theorem now trivially follows from the discussion at the start of the proof. ∎

III. Application: Estimation of Sparse Multiple-Antenna Channels

In this section, we discuss an application of structurally-subsampled unitary matrices in the area of estimation of multiple-antenna (MIMO) channels that have sparse impulse responses. For the sake of this exposition, we limit ourselves to sparse MIMO orthogonal frequency division multiplexing (OFDM) channels and devise quantitative error bounds for CS-based channel estimation schemes by leveraging the results of Theorem 1 for structurally-subsampled unitary matrices.

A. Problem Setup

Consider a MIMO OFDM channel $\mathcal{H}$ corresponding to a transmitter with N_T antennas, a receiver with N_R antennas, and an L-tap (discrete) impulse response. For simplicity, we assume uniform linear arrays of antennas and consider signaling over this channel using OFDM symbols of duration T and (two-sided) bandwidth W, thereby giving rise to a *temporal signal space* of dimension $N_o \stackrel{def}{=} TW$. Finally, as is customary in the wireless literature [16], [17], we assume that the number of taps in the channel impulse response, L, is much smaller than the number of OFDM subcarriers, N_o.

One of the most popular and widely used approaches to estimating a MIMO channel is to probe it with known signaling waveforms (referred to as training signals) and process the corresponding channel output to estimate the channel parameters. In the case of a MIMO OFDM channel, the N_T-dimensional (baseband) training signal can be expressed as

$$\mathbf{x}_{tr}(t) = \sqrt{\frac{\mathcal{E}}{N_T}} \sum_{n \in \mathcal{S}_{tr}} \tilde{\mathbf{x}}_n \, g(t) e^{j2\pi \frac{n}{T} t}, \ 0 \leq t \leq T \quad (28)$$

where $\mathcal{E}$ denotes the total transmit energy budget for training purposes, $g(t)$ is a prototype pulse having unit energy, $\mathcal{S}_{tr} \subset \mathcal{S} \stackrel{def}{=} \{0, \ldots, N_o - 1\}$ is the set of indices of *pilot subcarriers* used for training, and $\{\tilde{\mathbf{x}}_n \in \mathbb{C}^{N_T}\}$ is the (vector-valued) training sequence having energy $\sum_{\mathcal{S}_{tr}} \|\tilde{\mathbf{x}}_n\|_2^2 = N_T$.

At the receiver, the noisy received training signal $\mathbf{y}_{tr}(t) = \mathcal{H}(\mathbf{x}_{tr}(t)) + \mathbf{z}_{tr}(t)$ is matched filtered with the OFDM basis waveforms $\{g(t)e^{j2\pi \frac{n}{T} t}\}_{\mathcal{S}_{tr}}$ to yield [16], [17]

$$\tilde{\mathbf{y}}_n = \sqrt{\frac{\mathcal{E}}{N_T}} \mathbf{H}_n \tilde{\mathbf{x}}_n + \tilde{\mathbf{z}}_n, \ n \in \mathcal{S}_{tr} \quad (29)$$

where $\{\tilde{\mathbf{z}}_n\}$ are N_R-dimensional complex additive noise vectors that are independently distributed as $\mathcal{CN}(\mathbf{0}_{N_R}, \mathbf{I}_{N_R})$, while $\{\mathbf{H}_n\}$ are $N_R \times N_T$ matrices that are termed as *OFDM channel coefficients*. Finally, the OFDM channel coefficients $\{\mathbf{H}_n\}$ are related to the impulse response of $\mathcal{H}$ by [18]

$$\mathbf{H}_n \approx \sum_{\ell=0}^{L-1} \mathbf{A}_R \mathbf{H}_v^\mathsf{T}(\ell) \mathbf{A}_T^\mathsf{H} e^{-j2\pi \frac{\ell}{N_o} n}, \ n \in \mathcal{S}_{tr} \quad (30)$$

where $\mathbf{H}_v(\ell) \stackrel{def}{=} [\mathbf{h}_{v,1}(\ell) \ \ldots \ \mathbf{h}_{v,N_R}(\ell)]$ is an $N_T \times N_R$ matrix in which the i-th column, $\mathbf{h}_{v,i}(\ell)$, corresponds to the ℓ-th tap of the (vector) impulse response from the transmit array to the i-th receive antenna, while $\mathbf{A}_R$ and $\mathbf{A}_T$ are $N_R \times$

N_R and $N_T \times N_T$ (unitary) Fourier matrices, respectively. The goal then is to reliably estimate the impulse response of $\mathcal{H}$ using $\{\tilde{\mathbf{y}}_n, \tilde{\mathbf{x}}_n\}_{\mathcal{S}_{tr}}$ and a small number of pilot subcarriers.

B. Sparse MIMO OFDM Channel Estimation

Physical arguments and growing experimental evidence suggest that MIMO OFDM channels encountered in practice tend to exhibit impulse responses dominated by a relatively small number of dominant taps [19], [20]. Traditional MIMO OFDM channel estimation methods—typically comprising of linear reconstruction techniques (such as the maximum likelihood or the minimum mean squared error estimators), however, lead to overutilization of the key resources of energy and bandwidth in such *sparse channels*. To see this, define row vectors $\{\mathbf{y}_n^\mathsf{T} \stackrel{def}{=} \tilde{\mathbf{y}}_n^\mathsf{T} \mathbf{A}_R^*\}_{\mathcal{S}_{tr}}$ and note from (29) and (30) that

$$\mathbf{y}_n^\mathsf{T} = \sqrt{\frac{\mathcal{E}}{N_T}} \tilde{\mathbf{x}}_n^\mathsf{T} \mathbf{A}_T^* \sum_{\ell=0}^{L-1} \mathbf{H}_v(\ell) e^{-j2\pi \frac{\ell}{N_o} n} + \mathbf{z}_n^\mathsf{T} \quad (31)$$

where entries of the noise vectors $\{\mathbf{z}_n^\mathsf{T} \stackrel{def}{=} \tilde{\mathbf{z}}_n^\mathsf{T} \mathbf{A}_R^*\}_{\mathcal{S}_{tr}}$ are still (mutually) independently distributed as $\mathcal{CN}(0,1)$ due to the unitary nature of $\mathbf{A}_R^*$. Next, let $y_n(i), i = 1, \ldots, N_R$, denote the i-th entry of $\mathbf{y}_n^\mathsf{T}$. Then it can be shown using (31) and basic matrix identities involving Kronecker products that [15]

$$y_n(i) = \sqrt{\frac{\mathcal{E}}{N_T}} \tilde{\mathbf{x}}_n^\mathsf{T} (\mathbf{u}_n^\mathsf{T} \otimes \mathbf{A}_T^*) \mathbf{h}_{v,i} + z_n(i) \quad (32)$$

where $\otimes$ denotes the Kronecker product, $\mathbf{h}_{v,i}$ is an $N_T L$-dimensional column vector obtained by concatenating the vectors $\{\mathbf{h}_{v,i}(\ell)\}$, $\mathbf{u}_n^\mathsf{T} \stackrel{def}{=} [e^{-j0\omega_{n,N_o}} \ \ldots \ e^{-j(L-1)\omega_{n,N_o}}]$ is the collection of L samples of a discrete sinusoid with frequency $\omega_{n,N_o} \stackrel{def}{=} 2\pi \frac{n}{N_o}$, and $z_n(i)$ denotes the i-th entry of the noise vector $\mathbf{z}_n^\mathsf{T}$.

It is now easy to see from (31) and (32) that stacking the rows vectors $\{\mathbf{y}_n^\mathsf{T}\}_{\mathcal{S}_{tr}}$ into an $|\mathcal{S}_{tr}| \times N_R$ matrix $\mathbf{Y}$ yields the standard linear observation model

$$\mathbf{Y} = \sqrt{\frac{\mathcal{E}}{N_T}} \mathbf{X} \mathbf{H}_v + \mathbf{Z} \quad (33)$$

where $\mathbf{H}_v \stackrel{def}{=} [\mathbf{h}_{v,1} \ \ldots \ \mathbf{h}_{v,N_R}]$ is the unknown $N_T L \times N_R$ channel matrix, while $\mathbf{X}$ is an $|\mathcal{S}_{tr}| \times N_T L$ matrix comprising of $\{\tilde{\mathbf{x}}_n^\mathsf{T} (\mathbf{u}_n^\mathsf{T} \otimes \mathbf{A}_T^*) : n \in \mathcal{S}_{tr}\}$ as its rows. In order to estimate MIMO OFDM channels, traditional methods relying on linear reconstruction techniques (such as those in [21], [22]) therefore (i) require that the number of pilot subcarriers $|\mathcal{S}_{tr}| = \Omega(N_T L)$ so as to ensure that $\mathbf{X}$ has full column rank, and (ii) produce an estimate $\widehat{\mathbf{H}}_v$ of the channel matrix $\mathbf{H}_v$ that satisfies $\mathbb{E}[\|\widehat{\mathbf{H}}_v - \mathbf{H}_v\|_F^2] = \Omega(N_R N_T^2 / \mathcal{E})$.

In contrast, we now propose a CS-based approach to estimation of sparse MIMO OFDM channels that is based on the results of Theorem 1 for structurally-subsampled unitary matrices. The proposed approach uses a nonlinear reconstruction algorithm, known as the Dantzig selector (DS) [23], at the receiver and achieves a target reconstruction error using far less energy and bandwidth than that dictated by the

traditional methods based on linear reconstruction techniques. Before proceeding further, however, it is instructive to state the reconstruction error performance of the DS. The following theorem is a slight variation on [23, Theorem 1.1].

Theorem 3 (The Dantzig Selector [23]): Let $\boldsymbol{\nu} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\eta}$ be an $n \times 1$ vector of observations of any deterministic but unknown signal $\boldsymbol{\beta} \in \mathbb{C}^p$, where the entries of $\boldsymbol{\eta}$ are independently distributed as $\mathcal{CN}(0, \sigma^2)$. Assume that the columns of $\mathbf{A}$ have unit ℓ_2-norms and further let $\mathbf{A} \in RIP(2S, 0.3)$ for some integer $S \geq 1$. Choose $\lambda = \sqrt{2\sigma^2(1+a)\log p}$ for any $a \geq 0$. Then the vector $\boldsymbol{\beta}^{\mathsf{DS}}$ obtained as the solution of

$$\boldsymbol{\beta}^{\mathsf{DS}} = \underset{\tilde{\boldsymbol{\beta}} \in \mathbb{C}^p}{\arg\min} \|\tilde{\boldsymbol{\beta}}\|_1 \quad \text{subject to} \quad \|\mathbf{A}^{\mathsf{H}}(\boldsymbol{\nu} - \mathbf{A}\tilde{\boldsymbol{\beta}})\|_\infty \leq \lambda$$

satisfies

$$\|\boldsymbol{\beta}^{\mathsf{DS}} - \boldsymbol{\beta}\|_2^2 \leq c_0^2 \cdot S\,\sigma^2 \cdot \log p \tag{34}$$

with probability exceeding $1 - 2\left(\sqrt{\pi(1+a)\log p} \cdot p^a\right)^{-1}$. Here, the constant $c_0 \stackrel{def}{=} 4\sqrt{2(1+a)}/(1 - 3\delta_{2S})$.

We are now ready to state the training structure and the associated reconstruction algorithm of our proposed estimation scheme for d-sparse MIMO OFDM channels.

Training: Pick $\mathcal{S}_{tr}$—the set of indices of pilot subcarriers—to be a set of N_{tr} indices sampled uniformly at random (without replacement) from the set $\mathcal{S} = \{0, \ldots, N_o - 1\}$. Further, define the corresponding sequence of training vectors $\{\tilde{\mathbf{x}}_n, n \in \mathcal{S}_{tr}\}$ associated with $\mathbf{x}_{tr}(t)$ to be a sequence of i.i.d. Rademacher random vectors in which each entry independently takes the value $+1/\sqrt{N_{tr}}$ or $-1/\sqrt{N_{tr}}$ with probability $1/2$ each.

Reconstruction: Pick $\lambda = \sqrt{2\mathcal{E}(1+a)(\log N_R N_T L)/N_T}$ for some fixed $a \geq 0$. Next, define

$$\mathbf{h}_{v,i}^{\mathsf{DS}} = \underset{\mathbf{h} \in \mathbb{C}^{N_T L}}{\arg\min} \|\mathbf{h}\|_1 \quad \text{subject to}$$

$$\left\| \sqrt{\frac{\mathcal{E}}{N_T}} \mathbf{X}^{\mathsf{H}} \left(\mathbf{y}_i - \sqrt{\frac{\mathcal{E}}{N_T}} \mathbf{X} \mathbf{h} \right) \right\|_\infty \leq \lambda, \ i = 1, \ldots, N_R$$

where $\mathbf{y}_i \in \mathbb{C}^{N_{tr}}$ denotes the i-th column of the matrix $\mathbf{Y}$. The CS estimate of $\mathbf{H}_v$ is then simply given as follows

$$\mathbf{H}_v^{\mathsf{DS}} = \begin{bmatrix} \mathbf{h}_{v,1}^{\mathsf{DS}} & \cdots & \mathbf{h}_{v,N_R}^{\mathsf{DS}} \end{bmatrix}. \tag{35}$$

Theorem 4: Let $\mathcal{H}$ be a d-sparse MIMO OFDM channel in the sense that its impulse response satisfies

$$d \stackrel{def}{=} \sum_{i=1}^{N_R} \underbrace{\sum_{\ell=0}^{L-1} \|\mathbf{h}_{v,i}(\ell)\|_0}_{\stackrel{def}{=} d_i} \ll N_R N_T L. \tag{36}$$

Define $\bar{d} = \max_i d_i$ and suppose that $N_o, \bar{d} > 2$. Then for any $\delta_{2\bar{d}} \in (0, 0.3]$, the CS estimate of $\mathbf{H}_v$ satisfies

$$\left\| \mathbf{H}_v^{\mathsf{DS}} - \mathbf{H}_v \right\|_F^2 \leq c_0^2 \cdot \frac{d \cdot N_T}{\mathcal{E}} \cdot \log N_R N_T L \tag{37}$$

with probability exceeding $1 - 4\max\left\{ \left(\pi(1+a)\log N_R N_T L \cdot (N_R N_T L)^{2a}\right)^{-1/2}, 10 N_o^{-\delta_{2\bar{d}}^2} \right\}$, provided the number of pilot

subcarriers $N_{tr} \geq (2c_1/c_2)\bar{d}\log^6 N_o$. Here, the constants c_1, c_2 are the same as in Theorem 1, while the constant $c_0 = 4\sqrt{2(1+a)}/(1 - 3\delta_{2\bar{d}})$.

This theorem, which is proved in [15, Theorem 4.14] using the results of Theorem 1 and Theorem 3, essentially states that the proposed CS-based MIMO OFDM channel estimator can potentially reduce both the number of pilots subcarriers needed for channel estimation and the error in the resulting estimate by a factor of about $O(N_R N_T L/d)$ when used as an alternative to existing methods for estimating sparse MIMO OFDM channels.

IV. DISCUSSION

In this paper, we have introduced a new class of compressive sensing matrices—which we term as structurally-subsampled unitary matrices—that can be thought of as a generalization of subsampled unitary matrices. In particular, we have investigated the restricted isometry property of a specific form of structurally-subsampled unitary matrices in the paper that arise naturally in the estimation of multiple-antenna orthogonal frequency division multiplexing channels and successfully established in Theorem 1 that these matrices perform *nearly* as well as subsampled unitary matrices. Specifically, Theorem 1 for structurally-subsampled unitary matrices differs from [6, Theorem 3.3] for subsampled unitary matrices by only a factor of $\log p$. Note that this difference is primarily a consequence of the fact that the maximum magnitude of the entries in a subsampled unitary matrix is trivially given by $\mu_{\mathbf{U}}/\sqrt{n}$, whereas we could only bound the maximum magnitude of the entries in the structurally-subsampled unitary matrices of Theorem 1 by $\mu_{\mathbf{U}}\sqrt{\log p/n}$. However, it remains to be seen whether this is a fundamental characteristic of structurally-subsampled unitary matrices or just an artifact of the proof technique employed in Lemma 4. It is also instructive to note at this point that since the results for structurally-subsampled unitary matrices should coincide with those for subsampled unitary matrices for the case of a diagonal row-mixing matrix $\mathbf{R}$, it is heuristically plausible to conjecture that the performance of the structurally-subsampled unitary matrices of Theorem 1 should deviate from that of subsampled unitary matrices by a factor that is a function of k (instead of p). Such a conclusion, however, does not follow from the results established in this paper.

Finally, we conclude this paper with a brief discussion of the connections between the results of this paper and some existing works. As noted earlier, the work in Section II is closely related in terms of the general proof technique to the work of Romberg [13] and Tropp et al. [11] in general, and Rudelson and Vershynin [6] in particular. This is primarily a consequence of the fact that the arguments used by Rudelson and Vershynin in [6] are substantially simpler (and tighter) than, for instance, the ones used in [5] to establish the RIP of subsampled matrices.

In terms of the actual problem, however, our work in this paper is most closely related to the recent work of Tropp et al. [11], where they propose a sub-Nyquist sampling

architecture—termed *random demodulator*—to acquire sparse bandlimited signals. In particular, it is shown in [11] that the overall action of the random demodulator on a sparse bandlimited signal can be accurately described in terms of a sensing matrix, which the authors term as a *random demodulator matrix*. However, it is easy to see from [11, Section IV-B] that a random demodulator matrix is just a structurally-subsampled unitary matrix of the form described in Theorem 1 with $\mathbf{U}$ being a Fourier matrix and $k = p/n$ (in other words, no subsampling). In this regard, our work in this paper can also be thought of a generalization of the RIP analysis of a random demodulator matrix carried out in [11]. Based on the preceding discussion, it is perhaps best to think of the structurally-subsampled unitary matrices of Theorem 1 as filling the void between the two extremes of subsampled unitary matrices (maximum subsampling) and random demodulator matrices (no subsampling) through the choice of the design parameter k (with k ranging from 1 to p/n).

APPENDIX

PROOF OF LEMMA 1

Let $\mathbf{a}_i^{\mathsf{H}} \in \mathbb{C}^p$ denote the i-th row of $\mathbf{A}$. Then $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ can be written as a sum of rank-one matrices as follows

$$\mathbf{A}^{\mathsf{H}}\mathbf{A} = \sum_{i=1}^{|\Omega|} \mathbf{a}_i \mathbf{a}_i^{\mathsf{H}} = \frac{m}{n} \sum_{i=1}^{p} \zeta_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathsf{H}}$$
$$\Rightarrow \quad \mathbb{E}[\mathbf{A}^{\mathsf{H}}\mathbf{A}] = \mathbb{E}[\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi}] \tag{38}$$

where $\boldsymbol{\phi}_i^{\mathsf{H}}$ denotes the i-th row of $\boldsymbol{\Phi}$. Next, from the definition of $\boldsymbol{\Phi}$, we can write an expression for $\boldsymbol{\phi}_i^{\mathsf{H}}$ in terms of elements of the generating sequence $\mathcal{A}_p$ and the rows of $\mathbf{U}$ as follows

$$\boldsymbol{\phi}_i^{\mathsf{H}} = \sum_{\ell=1}^{k} a_{(i-1)k+\ell} \, \mathbf{u}_{(i-1)k+\ell}^{\mathsf{H}}, \; i = 1, \dots, m \tag{39}$$

where $\mathbf{u}_i^{\mathsf{H}}$ denotes the i-th row of $\mathbf{U}$. With the help of (39), we can further write the (i, j)-th entry of $\boldsymbol{\Phi}$ as

$$\phi_{i,j} = \sum_{\ell=1}^{k} a_{(i-1)k+\ell} \, u_{(i-1)k+\ell,j} \tag{40}$$

where $u_{i,j}$ is the (i, j)-th entry of $\mathbf{U}$. We then have from (40)

$$\mathbb{E}[\phi_{i,j}^* \phi_{i,j'}] = \sum_{q=1}^{k} \sum_{r=1}^{k} \mathbb{E}\big[a_{(i-1)k+q} \, a_{(i-1)k+r}\big] \times$$

$$u_{(i-1)k+q,j}^* \, u_{(i-1)k+r,j'} = \sum_{q=1}^{k} u_{(i-1)k+q,j}^* \, u_{(i-1)k+q,j'}. \tag{41}$$

Finally, define the Gram matrix $\mathbf{G} \overset{def}{=} \boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi}$. Then we have from (41) that the expected value of the (i, j)-th entry of $\mathbf{G}$, $g_{i,j} = \sum_{\ell=1}^{m} \phi_{\ell,i}^* \phi_{\ell,j}$, is given by

$$\mathbb{E}[g_{i,j}] = \sum_{\ell=1}^{m} \mathbb{E}[\phi_{\ell,i}^* \phi_{\ell,j}] = \sum_{\ell=1}^{m} \sum_{q=1}^{k} u_{(\ell-1)k+q,i}^* \, u_{(\ell-1)k+q,j}$$

$$= \sum_{\ell=1}^{p} u_{\ell,i}^* u_{\ell,j} \overset{(a)}{=} \delta_{ij}, \; i, j = 1, \dots, p \tag{42}$$

where δ_{ij} is the Kronecker delta and (a) follows from the fact that $\mathbf{U}$ is a unitary matrix. This completes the proof since (42) implies that $\mathbb{E}[\mathbf{G}] = \mathbf{I}_p \Rightarrow \mathbb{E}[\mathbf{A}^{\mathsf{H}}\mathbf{A}] = \mathbf{I}_p$ from (38).

REFERENCES

[1] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. of Mathematicians*, vol. III, Madrid, Spain, Aug. 2006, pp. 1433–1452.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

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

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

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

[6] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, no. 8, pp. 1025–1045, Aug. 2008.

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

[8] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

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

[10] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," submitted. [Online]. Available: http://www.math.princeton.edu/~wbajwa/pubs/proc08_ccs-v1.pdf

[11] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," submitted. [Online]. Available: arXiv:0902.0026v1

[12] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*. New York, NY: Springer-Verlag, 1991.

[13] J. Romberg, "Compressive sensing by random convolution," submitted. [Online]. Available: http://users.ece.gatech.edu/~justin/Publications.html

[14] U. Haagerup, "The best constants in the Khintchine inequality," *Studia Math.*, vol. 70, no. 3, pp. 231–283, 1981.

[15] W. U. Bajwa, "New information processing theory and methods for exploiting sparsity in wireless systems," Ph.D. dissertation, University of Wisconsin-Madison, Madison, WI, 2009.

[16] A. Goldsmith, *Wireless Communications*. New York, NY: Cambridge University Press, 2005.

[17] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge University Press, 2005.

[18] A. M. Sayeed, "A virtual representation for time- and frequency-selective correlated MIMO channels," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 4, Hong Kong, Apr. 2003, pp. 648–651.

[19] Z. Yan, M. Herdin, A. M. Sayeed, and E. Bonek, "Experimental study of MIMO channel statistics and capacity via the virtual channel representation," University of Wisconsin-Madison, Tech. Rep., Feb. 2007.

[20] 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.*, vol. 6, no. 4, pp. 1465–1475, Apr. 2007.

[21] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Processing*, vol. 51, no. 6, pp. 1615–1624, Jun. 2003.

[22] H. Minn and N. Al-Dhahir, "Optimal training signals for MIMO OFDM channel estimation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 5, pp. 1158–1168, May 2006.

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