THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

The restricted isometry property for random block diagonal matrices

OPEN ACCESS

The Restricted Isometry Property for Random Block Diagonal Matrices

Armin Eftekhari, Han Lun Yap, Christopher J. Rozell, and Michael B. Wakin[1]

Abstract

In Compressive Sensing, the Restricted Isometry Property (RIP) ensures that robust recovery of sparse vectors is possible from noisy, undersampled measurements via computationally tractable algorithms. It is by now well-known that Gaussian (or, more generally, sub-Gaussian) random matrices satisfy the RIP under certain conditions on the number of measurements. Their use can be limited in practice, however, due to storage limitations, computational considerations, or the mismatch of such matrices with certain measurement architectures. These issues have recently motivated considerable effort towards studying the RIP for structured random matrices. In this paper, we study the RIP for block diagonal measurement matrices where each block on the main diagonal is itself a sub-Gaussian random matrix. Our main result states that such matrices can indeed satisfy the RIP but that the requisite number of measurements depends on certain properties of the basis in which the signals are sparse. In the best case, these matrices perform nearly as well as dense Gaussian random matrices, despite having many fewer nonzero entries.

Keywords— Compressive Sensing, Block Diagonal Matrices, Restricted Isometry Property

1. Introduction

Many interesting classes of signals have a low-dimensional geometric structure that can be exploited to design efficient signal acquisition and recovery methods. The emerging field of Compressive Sensing (CS) [7, 18, 19] deals with signals that can be parsimoniously expressed in a basis or a dictionary. A canonical result in CS states that sparse signals, i.e., signals with very few nonzero entries, can be accurately recovered from a small number of linear measurements by solving a tractable convex optimization problem if the measurement system satisfies the Restricted Isometry Property (RIP) [7].

The RIP requires the linear measurement system to approximately maintain the distance between any pair of sparse signals in the measurement space, implying that the geometry of the family of sparse signals is approximately preserved in the measurement space. Apart from playing a central role in the analysis of numerous signal recovery algorithms in CS [8, 30, 14, 21, 23], the RIP also provides a framework to analyze signal processing and inference algorithms in the compressed measurement domain [13]. Moreover, measurement systems that satisfy the RIP, after undergoing some minor modifications, can approximately preserve the geometry of an arbitrary point cloud (as confirmed by the Johnson-Lindenstrauss lemma) [27] or a low-dimensional compact manifold [44].

A measurement system represented by a matrix populated with i.i.d. sub-Gaussian[2] random variables is

---

[2]Roughly speaking, the tail of a sub-Gaussian random variable is similar to that of a Gaussian random variable. This term is defined precisely in Section 2.

known to satisfy the RIP with high probability whenever the number of rows scales linearly with the sparsity of the signal and logarithmically with the length of the signal [7, 3]. Such matrices are also universal in that, with the same number of random measurements, they satisfy the RIP with respect to any fixed sparsity basis with high probability. We refer to such matrices—densely populated with i.i.d. random entries—as *unstructured* measurement matrices. There has been significant recent interest in studying *structured* measurement systems because unstructured random measurements may be undesirable due to memory limitations, computational costs, or specific constraints in the data acquisition architecture. Many structured systems have been studied in the CS literature, including subsampled bounded orthonormal systems [35, 32], random convolution systems (described by partial Toeplitz [22] and circulant matrices [33, 26]) and deterministic matrix constructions [15]. Generally, structured random matrices require more measurements to satisfy the RIP and lack the universality of unstructured random matrices.

In this paper, we are concerned with establishing the RIP for block diagonal matrices populated with i.i.d. sub-Gaussian random variables. The advantages of such matrices are varied. First, these matrices require less memory and computational resources than their unstructured counterparts. Second, they are particularly useful for representing acquisition systems with architectural constraints that prevent global data aggregation. For example, this type of architecture arises in distributed sensing systems where communication and environmental constraints limit the dependence of each sensor to only a subset of the data and in streaming applications where signals have data rates that necessitate operating on local signal blocks rather than on the entire signal simultaneously. In these scenarios, the data may be divided naturally into discrete blocks, with each block acquired via a local measurement operator.

To make things concrete, for some positive integers J, N, and M, set $\widetilde{M} := JM$ and $\widetilde{N} := JN$. We model a signal $x \in \mathbb{C}^{\widetilde{N}}$ as being partitioned into J blocks of length N, i.e., $x = \begin{bmatrix} x_1^T, & \cdots, & x_J^T \end{bmatrix}^T$ where $x_j \in \mathbb{C}^N$, $j \in [J]$. Here, $[J]$ denotes the set $\{1, 2, \cdots, J\}$. As an example, x can be a video sequence and $\{x_j\}, j \in [J]$, can be the individual frames in the video. For each $j \in [J]$, we suppose that a linear operator $\Phi_j : \mathbb{C}^N \to \mathbb{C}^M$ collects the measurements $y_j = \Phi_j x_j$. In our example, this means that each video frame x_j is measured with an operator Φ_j. Concatenating all of the measurements into a vector $y \in \mathbb{C}^{\widetilde{M}}$, we then have

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{bmatrix}}_{y: \ \widetilde{M} \times 1} = \begin{bmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_J \end{bmatrix} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_J \end{bmatrix}}_{x: \ \widetilde{N} \times 1}, \tag{1}$$

Thus we see that the overall measurement operator relating y to x will have a block diagonal structure. In this paper we consider two scenarios. When the $\{\Phi_j\}$ are distinct, we call the resulting matrix a Distinct Block Diagonal (DBD) matrix. On the other hand, when the $\{\Phi_j\}$ are all identical, we call the resulting matrix a Repeated Block Diagonal (RBD) matrix. Our results show that whenever the total number of measurements $\widetilde{M}$ is sufficiently large, DBD and RBD matrices can both satisfy the RIP. As we summarize in Sections 1.2 and 1.3 below, the requisite number of measurements depends on the type of matrix (DBD or RBD) and on the basis in which x has a sparse expansion. We also show that certain sparse matrices and random convolution systems considered in the CS literature can be studied in the framework of block diagonal matrices.

In general, proving the RIP for structured measurement systems requires analytic tools beyond the elementary approaches that suffice for unstructured matrices. For example, in [35] the authors employed tools such as Dudley's inequality from the theory of probabilities in Banach spaces, and in [33] a variant of Dudley's inequality for chaos random processes was used to obtain a result that was out of reach for elementary approaches. While some of the ideas and techniques utilized in [35] and [33] can be used to establish the RIP for random block diagonal matrices (see [42] for our preliminary study), even these sophisticated tools result in measurement rates that are worse than what we report in this paper. Fortunately, recent work by Krahmer et al. has established an improved bound on the suprema of chaos random processes that enabled them to prove the RIP for Toeplitz matrices with an optimal number of measurements [26]. The bound

2

in [26] is very general, and we have leveraged this result for the main results of this paper. Specifically, the work in [26] has allowed us to develop a unified treatment of DBD and RBD matrices with bounds on the measurement rates that are significantly improved over our preliminary work.

1.1. Definition of the RIP

A linear measurement operator satisfies the RIP if it acts as an approximate isometry on all sufficiently sparse signals. More specifically, the Restricted Isometry Constant (RIC) of a matrix $A \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}$ is defined as the smallest positive number δ_S for which

$$(1 - \delta_S)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_S)\|x\|_2^2 \quad \text{for all } x \text{ with } \|x\|_0 \leq S, \tag{2}$$

where $\|\cdot\|_0$ merely counts the number of nonzero entries of a vector. In many applications, however, signals may be sparse in an orthobasis U other than the canonical basis, and so we will find the notion of the U-RIP more convenient.

Definition 1. *Let U denote an orthobasis for $\mathbb{C}^{\widetilde{N}}$. The RIC of a matrix $A \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}$ in the basis U, $\delta_S = \delta_S(A, U)$, is defined as the smallest positive number for which*

$$(1 - \delta_S)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_S)\|x\|_2^2 \quad \text{for all } x \text{ with } \|U^*x\|_0 \leq S, \tag{3}$$

where U^* denotes the conjugate transpose of U.

In general, whenever the sparsity basis is clear from the context, U is dropped from our notation (after its first appearance). More generally, the notion of the RIP could be extended to the class of signals that are sparse in an overcomplete dictionary [9]. Nonetheless, for the clarity of exposition, we restrict ourselves throughout this paper to considering only orthobases.

1.2. The RIP for Distinct Block Diagonal (DBD) Matrices

Suppose the matrices $\{\Phi_j\}$, $j \in [J]$, in (1) are distinct, and let Ψ denote the resulting block diagonal matrix. Following [31], we say that Ψ has a DBD structure. DBD matrices arise naturally when modeling the information captured by individual (and different) sensors in a sensor network. In this setting, $y_j = \Phi_j x_j$ represents the local measurements of the signal x made by jth sensor and thus $y = \Psi x$ represents the total measurements of x captured by the whole network. Or, as mentioned previously, DBD matrices can be used to represent the process of measuring a video sequence frame-by-frame, but where each frame is observed using a different measurement matrix. DBD matrices also arise in the study of observability matrices in certain linear dynamical systems [40], and as another example, DBD matrices can be used as a simplified model for the visual pathway (because the information captured by the photoreceptors in the retina is aggregated locally by horizontal and bipolar cells [25]). Due to their structure, DBD matrices can be transformed into sparse measurement matrices after permutation of their rows and columns [1]. Also see [16] for an application of block diagonal matrices in developing fast and efficient compressive measurement operators.

We have previously derived concentration inequalities for DBD matrices populated with sub-Gaussian random variables [39, 31]. Rather than ensuring the stable embedding of an entire family of sparse signals, these equalities concern the probability that a bound such as (2) will hold for a single, arbitrary (not necessarily sparse) signal x. We have shown that, unlike the case for unstructured random matrices, the probability of concentration with DBD matrices is actually *signal dependent*, and in particular the concentration probability depends on the allocation of the signal energy among the signal blocks. However, for signals whose energy is nearly uniformly spread across the J blocks (this happens, for example, with signals that are sparse in the Fourier domain [31]), the highly structured DBD matrices can provide concentration performance that is on par with the unstructured matrices often used in CS.

While concentration of measure inequalities are useful for applications concerning compressive signal processing [13], it is not evident how such a concentration result can be extended to give an RIP bound as strong as the one in this paper. Specifically, in Section 3.2 of this paper, we show that if the total number of measurements $\widetilde{M}$ scales linearly with the sparsity of the signal S and poly-logarithmically with the ambient

dimension $\widetilde{N}$, DBD matrices populated with sub-Gaussian random entries will satisfy the U-RIP with high probability. In addition to a dependence on S and $\widetilde{N}$, however, our measurement bounds also reveal a dependence on a property known as the *coherence* of the sparsifying basis U. In this sense, the signal-dependent nature of our concentration of measure inequalities carries over to our RIP analysis for DBD matrices. (The fine details of how this occurs, however, are different.) Our study does confirm that for the class of signals that are sparse in the frequency domain, DBD matrices satisfy the RIP with approximately the same number of rows required in an unstructured Gaussian random matrix (despite having many fewer nonzero entries).

1.3. The RIP for Repeated Block Diagonal (RBD) Matrices

Alternatively, suppose the matrices $\{\Phi_j\}$, $j \in [J]$, in (1) are all equal, and let Ξ denote the resulting block diagonal matrix. Following [31], we say that Ξ has an RBD structure. In the context of the sensor network, video processing, and observability applications discussed before, RBD matrices arise when the same measurement matrix is used for all the signal blocks. In the delay embedding of dynamical systems, as another example, a time series is obtained by repeatedly applying a scalar measurement function to the trajectory of a dynamical system. This time series can then be embedded in a low-dimensional space (hence the name), and this embedding can be expressed through an RBD measurement matrix provided that the scalar measurement function is linear [43]. Though not obvious at first glance, RBD matrices also have structural similarities with random convolution matrices found in the CS literature [33, 26]. We revisit this connection in order to re-derive the RIP for partially circulant random matrices in Section 3.3.2.

We have previously derived concentration inequalities for RBD matrices populated with Gaussian random variables [39, 34, 31]. Our bounds for these matrices again reveal that the probability of concentration is signal dependent. However, in this case, our bounds depend on both the allocation of the signal energy among the signal blocks as well as the mutual orthogonality of the signal blocks.

In Section 3.3 of this paper, we show that if the total number of measurements $\widetilde{M}$ scales linearly with S and poly-logarithmically with $\widetilde{N}$, RBD matrices populated with sub-Gaussian random variables will satisfy the U-RIP with high probability. Our measurement bounds also reveal a dependence on a property known as the *block-coherence* of the sparsifying basis U that quantifies the dependence between its row blocks. When the block-coherence of U is small, RBD matrices perform favorably compared to unstructured Gaussian random matrices. Most sparsifying bases are in fact favorable in this regard; we prove that the block-coherence of U is small when U is selected randomly. Once again, for RBD matrices, the signal dependent nature of concentration inequalities and the dependence of the RIP on the sparsifying basis emerge as two sides of the same coin.

1.4. Outline

This paper is organized as follows. Section 2 introduces the notation used throughout the rest of the paper. Section 3 summarizes our main results regarding the RIP for DBD and RBD matrices; these results are later proved in Section 5. Section 4 presents numerical simulations that illustrate the dependence of signal recovery performance on the sparsifying basis U. We conclude the paper with a short discussion in Section 6. We note that, for the reader's convenience, the Toolbox (Appendix A) gathers some general tools from linear algebra and probability theory used in our analysis.

2. Notation

We reserve the letters $C, C_1, C_2, \cdots$ to represent universal positive constants. We adopt the following (semi-)order: $a \lesssim b$ means that there is an absolute constant C_1 such that $a \leq C_1 b$. If, instead of being an absolute constant, C_1 depends on some parameter c, we write $a \lesssim_c b$. Also $a \gtrsim b$ and $a \gtrsim_c b$ are defined similarly.

For an integer S, a signal with no more than S nonzero entries is called S-sparse, and S is known as the sparsity level. In particular, $\|a\|_0$ denotes the number of nonzero entries of a vector a. More generally, a signal that is a linear combination of at most S columns of a basis is said to be S-sparse in that basis.

The conjugate transpose of a matrix A will be denoted by A^*. In this paper, $\text{Rank}\,(A)$ stands for the rank of matrix A. In addition to the regular ℓ_p-norms in the Euclidean spaces, $1 \leq p \leq \infty$, we use $\|A\|_2$ and $\|A\|_F$ to denote the spectral and Frobenious norms of a matrix A, respectively. We use $\|A\|_{\max}$ to denote the largest entry of the matrix A in magnitude. For $1 \leq p \leq \infty$, the Schatten norm of order p of a matrix A is denoted by $\|A\|_{S_p}$ and is defined as

$$\|A\|_{S_p} := \|\sigma_A\|_p,$$

where σ_A is the vector formed by the singular values of A. Observe that $\|A\|_{S_\infty} = \|A\|_2$ and $\|A\|_{S_2} = \|A\|_F$. Throughout this paper, for a matrix A, $\text{vec}(A)$ returns the vector formed by stacking the columns of A. Also, we will use the conventions $[N] := \{1, 2, \cdots, N\}$ (for an integer N) and $\#T$ for the cardinality of a set T.

When it appears, the subscript of an expectation operator $\mathbb{E}$ specifies the (group of) random variable(s) with respect to which the expectation is taken. For a random variable Z taking values in $\mathbb{C}$, we define $\mathbb{E}^p|Z| := (\mathbb{E}|Z|^p)^{1/p}$, $p \geq 1$. A random variable Z is sub-Gaussian if its sub-Gaussian norm, defined below, is finite [38]:

$$\|Z\|_{\psi_2} := \sup_{p \geq 1} \frac{1}{\sqrt{p}} \mathbb{E}^p|Z|. \tag{4}$$

Qualitatively speaking, the tail of (the distribution of) a sub-Gaussian random variable is similar to that of a Gaussian random variable, hence the name. A Rademacher sequence is a sequence of i.i.d. random variables that take the values ± 1 with equal probability. In this paper, every appearance of such a sequence is independent of other random variables in this paper and, in particular, independent of other Rademacher sequences elsewhere in the paper. Finally, $\underset{\text{i.d.}}{=}$ means that the random variables on both sides of the equality have the same distribution.

A set $\mathcal{C}\,(\mathcal{S}, \|\cdot\|, r)$ is called a cover for the set $\mathcal{S}$ at resolution r and with respect to the metric $\|\cdot\|$ if for every $x \in \mathcal{S}$, there exists $x' \in \mathcal{C}\,(\mathcal{S}, \|\cdot\|, r)$ such that $\|x - x'\| \leq r$. The minimum cardinality of all such covers is called the covering number of $\mathcal{S}$ at resolution r and with respect to the norm $\|\cdot\|$, and is denoted here by $\mathcal{N}(\mathcal{S}, \|\cdot\|, r)$.

3. Main Results

3.1. Measures of Coherence

Our results for random block diagonal matrices depend on certain properties of the sparsity basis, i.e., the basis in which the signals have a sparse expansion. These properties are defined and studied in this section; this sets the stage for a detailed statement of our main results in Sections 3.2 and 3.3.

3.1.1. Coherence Definitions

The *coherence* of an orthobasis $U \in \mathbb{C}^{\widetilde{N} \times \widetilde{N}}$ is defined as follows [10]:

$$\mu(U) := \sqrt{\widetilde{N}} \max_{p,q \in [\widetilde{N}]} |U(p,q)|, \tag{5}$$

where $U(p,q)$ is the (p,q)th entry of U. If $\{u_{\widetilde{n}}\}$ and $\{e_{\widetilde{n}}\}$, $\widetilde{n} \in [\widetilde{N}]$, denote the columns of U and of the canonical basis for $\mathbb{C}^{\widetilde{N}}$, respectively, one can easily verify that

$$\mu(U) = \sqrt{\widetilde{N}} \max_{p,q \in [\widetilde{N}]} |\langle u_p, e_q \rangle|. \tag{6}$$

This allows us to interpret $\mu(U)$ as the similarity between U and the canonical basis.

A few more definitions are in order before we can define the second important property of a basis used in this paper. For $\alpha \in \mathbb{C}^{\widetilde{N}}$, set $x(\alpha) = x(\alpha, U) := U\alpha$, and define $x_j(\alpha) = x_j(\alpha, U) \in \mathbb{C}^N$, $j \in [J]$, such that

$$x(\alpha) = [x_1(\alpha)^T, x_2(\alpha)^T, \cdots, x_J(\alpha)^T]^T. \tag{7}$$

If we also define $U_j \in \mathbb{C}^{N \times \widetilde{N}}$, $j \in [J]$, such that

$$U = [U_1^T, U_2^T, \cdots, U_J^T]^T, \tag{8}$$

we observe that $x_j(\alpha) = U_j\alpha$ for every j. Define $X_R(\alpha, U) \in \mathbb{C}^{N \times J}$ as

$$X_R(\alpha) = X_R(\alpha, U) := \begin{bmatrix} x_1(\alpha) & x_2(\alpha) & \cdots & x_J(\alpha) \end{bmatrix} = \begin{bmatrix} U_1\alpha & \cdots & U_J\alpha \end{bmatrix}.$$

Now the *block-coherence* of U, denoted by $\gamma(U)$, is defined as

$$\gamma(U) := \sqrt{J} \max_{\widetilde{n} \in [\widetilde{N}]} \|X_R(e_{\widetilde{n}}, U)\|_2. \tag{9}$$

In words, $\gamma(U)$ is proportional to the maximal spectral norm when any column of U is reshaped into an $N \times J$ matrix. In the special case where $J = 1$ or $N = 1$, $\gamma(U)$ simply scales with the maximum ℓ_2-norm of the columns of U. In analogy with (6), one can also think of (9) as a (non-commutative) coherence measure between U and $I_{\widetilde{N}}$.[3] Qualitatively speaking, $\gamma(U)$ measures the orthogonality and distribution of energy between $U_1, U_2, \cdots, U_J$, the consecutive $N \times \widetilde{N}$ row-submatrices of U. If for every column of U, the energy is evenly distributed between its row-blocks and they are nearly orthogonal, $\gamma(U)$ will be small and, as we will see later, better suited for our purposes. In the next subsection, we compute the coherence and block-coherence of a few widely-used orthobases.

3.1.2. Computing the Coherence for a Few Orthonormal Bases

It is easily verified that

$$1 \le \mu(U) \le \sqrt{\widetilde{N}}. \tag{10}$$

The upper bound is achieved, for example, by the canonical basis in $\mathbb{C}^{\widetilde{N}}$, i.e., $\mu(I_{\widetilde{N}}) = \sqrt{\widetilde{N}}$. The lower bound, on the other hand, is achieved by any basis that is maximally incoherent with the canonical basis. For example, $\mu(F_{\widetilde{N}}) = 1$, where $F_{\widetilde{N}}$ denotes the Fourier basis in $\mathbb{C}^{\widetilde{N}}$. The next lemma, proved in Appendix B, indicates that *most* orthobases are also highly incoherent with the canonical basis.

Lemma 1. *Let $R \in \mathbb{R}^{\widetilde{N} \times \widetilde{N}}$ denote a generic orthobasis in $\mathbb{C}^{\widetilde{N}}$ chosen randomly from the uniform distribution on the orthogonal group. Then the following holds for fixed $t \gtrsim 1$ and $\widetilde{N} \gtrsim t^2 \log \widetilde{N}$:*

$$\mathrm{P}\left\{\mu(R) > t\sqrt{\log \widetilde{N}}\right\} \lesssim \widetilde{N}^{-t}. \tag{11}$$

We now turn to computing the block-coherence of the same orthobases. Since every column of U has unit ℓ_2-norm, it is easily observed that

$$1 \le \gamma(U) \le \sqrt{J}. \tag{12}$$

Consider the canonical basis in $\mathbb{C}^{\widetilde{N}}$. For every $\widetilde{n} \in [\widetilde{N}]$, $X(e_{\widetilde{n}}, I_{\widetilde{N}})$ has a single non-zero entry, which equals 1, and thus $\|X(e_{\widetilde{n}}, I_{\widetilde{N}})\|_2 = 1$. Hence, $\gamma(I_{\widetilde{N}}) = \sqrt{J}$. Moving on to $F_{\widetilde{N}}$, we observe that the entries of the first column of $F_{\widetilde{N}}$ equal $\widetilde{N}^{-1/2}$. As a result, the entries of $X(e_1, F_{\widetilde{N}})$ all equal $\widetilde{N}^{-1/2}$. It follows that $\|X(e_1, F_{\widetilde{N}})\|_2 = 1$ and therefore $\gamma(F_{\widetilde{N}}) = \sqrt{J}$.

Because the canonical basis and the Fourier basis—which one might naturally consider to be opposite ends on some spectrum of orthobases—both have large block-coherence, one might wonder whether any orthobasis could have small block-coherence. As we will see, *most* possible orthobases actually do have small block-coherence. For example, consider the generic orthobasis R constructed in Lemma 1. The columns of $X(e_1, R)$ are J random vectors in $\mathbb{R}^N$. These vectors are weakly dependent because the first column of R has unit ℓ_2-norm. With high probability, the length of each vector is approximately $1/\sqrt{J}$ (so that the ℓ_2-norm

---

[3]It can be easily verified that, in general, $\max_{\widetilde{n}} \|X_R(e_{\widetilde{n}}, U)\|_2 \ne \max_{\widetilde{n}} \|X_R(u_{\widetilde{n}}, I_{\widetilde{N}})\|_2$, where $u_{\widetilde{n}}$ is the $\widetilde{n}$th column of U.

of the first column of R is one). If $J \leq N$, then with high probability these points are spread out in $\mathbb{R}^N$ so that $\|X(e_1, R)\|_2 \approx 1/\sqrt{J}$. Now, since the columns of R have the same distribution, $\|X(e_{\widetilde{n}}, R)\|_2 \approx 1/\sqrt{J}$ for every $\widetilde{n} \in [\widetilde{N}]$. Therefore, $\gamma(R) \approx 1$, which is much smaller than the block-coherence of the canonical and Fourier bases. The next result, proved in Appendix C, formalizes this discussion.

Lemma 2. *Consider the generic orthobasis R constructed in Lemma 1. For fixed $t \leq 1$, the following holds if $J \leq N$ and $N \gtrsim t^{-2} \log \widetilde{N}$:*

$$P\left\{ \gamma(R) \gtrsim 1 + \sqrt{\frac{J}{N}} + t \right\} \lesssim \widetilde{N}^{-t}. \tag{13}$$

We close this section by noting that Section 3.3.2 provides an example of a deterministic basis with small block-coherence, which is obtained by modifying the Kronecker product of the canonical and Fourier bases. (This basis is then used to prove the RIP for partial random circulant matrices.)

3.2. The RIP for DBD Matrices

Let $\Phi \in \mathbb{R}^{M \times N}$ denote a matrix populated with i.i.d. sub-Gaussian random variables having mean zero, standard deviation $1/\sqrt{M}$, and sub-Gaussian norm $\tau/\sqrt{M}$, for some $\tau > 0$. Take $\{\Phi_j\}$ in (1) to be J independent copies of Φ and let $\Psi \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}$ denote the resulting block diagonal matrix in (1). Our first main result, proved in Section 5, establishes the RIP for DBD matrices with this construction.

Theorem 1. *Let U denote an orthobasis for $\mathbb{C}^{\widetilde{N}}$ and set*

$$\widetilde{\mu}(U) := \min\left(\sqrt{J}, \ \mu(U) \right), \qquad \text{where} \qquad \mu(U) = \sqrt{\widetilde{N}} \max_{p,q \in [\widetilde{N}]} |U(p,q)|. \tag{14}$$

If $S \gtrsim 1$ and

$$\widetilde{M} \gtrsim_\tau \delta^{-2} \widetilde{\mu}^2(U) \cdot S \cdot \log^2 S \log^2 \widetilde{N}, \tag{15}$$

then $\delta_S(\Psi, U) \leq \delta < 1$, except with a probability of at most $O(\widetilde{N}^{-\log \widetilde{N} \log^2 S})$.

A few remarks are in order. The requisite number of measurements is linear in the sparsity level S and (poly-)logarithmic in the ambient dimension $\widetilde{N}$, on par with an unstructured random Gaussian matrix [7]. More importantly, the requisite number of measurements scales with $\widetilde{\mu}^2(U)$ which takes a value in the interval $[1, J]$. For the Fourier basis, we calculated that $\mu(F_{\widetilde{N}}) = 1$. Therefore, when measuring signals that are sparse in the frequency domain, we observe that a DBD matrix compares favorably to an unstructured Gaussian matrix of the same size. This is in the sense that they both require the same number of measurements to achieve the RIP (up to a poly-logarithmic factor).

On the other hand, when the orthobasis U is highly coherent with the canonical basis, the requisite number of measurements is proportional to SJ (instead of S). While possibly unfavorable, this is indeed necessary (to within a poly-logarithmic factor) to achieve the RIP in some cases. For example, recall that $\mu(I_{\widetilde{N}}) = \sqrt{\widetilde{N}}$ and so $\widetilde{\mu}(I_{\widetilde{N}}) = \sqrt{J}$. To see why the results are optimal in this case, consider the class of S-sparse signals in $I_{\widetilde{N}}$ whose nonzero entries are located within the first length-N block of the signal. Achieving a stable embedding of this class of signals requires Φ_1 itself to satisfy the RIP. This matrix, Φ_1, is an unstructured sub-Gaussian matrix, and ensuring that it satisfies the RIP requires $M \gtrsim_\tau \delta^{-2} S \log(N/S)$ [7]. Consequently, ensuring the $I_{\widetilde{N}}$-RIP for Ψ is only possible when $\widetilde{M} \gtrsim_\tau \delta^{-2} JS \log(N/S)$, as predicted by Theorem 1 (up to a poly-logarithmic term). The required number of measurements in this case can still be parsimonious ($\widetilde{M} \ll \widetilde{N}$), however, if the sparsity level S of the signal x is much less than N, the length of each signal block x_j.

As a final note, Theorem 1 implies the RIP for a certain class of sparse matrices which are of potential interest in their own right [4, 1].

Corollary 1. *Let U denote an orthobasis for $\mathbb{C}^{\widetilde{N}}$ and define $\widetilde{\mu}(U) := \min\left(\sqrt{J},\ \mu(U)\right)$. Let Ψ' denote the (sparse) matrix obtained by an arbitrary permutation of the rows and columns of Ψ. If $S \gtrsim 1$ and*

$$\widetilde{M} \gtrsim_\tau \delta^{-2}\widetilde{\mu}^2(U) \cdot S \cdot \log^2 S \log^2 \widetilde{N}, \tag{16}$$

then $\delta_S(\Psi', U) \leq \delta < 1$, except with a probability of at most $O(\widetilde{N}^{-\log \widetilde{N} \log^2 S})$.

Proof. Without loss of generality, consider no permutation in the rows and let $P_c \in \mathbb{R}^{\widetilde{N}}$ denote the permutation matrix for the columns of Ψ. Since P_cU has the same coherence as U, the claim follows by applying Theorem 1. $\square$

3.3. The RIP for RBD Matrices

3.3.1. Main Result for RBD Matrices

Let $\Phi \in \mathbb{R}^{M \times N}$ denote a matrix populated with i.i.d. sub-Gaussian random variables having mean zero, standard deviation $1/\sqrt{M}$, and sub-Gaussian norm $\tau/\sqrt{M}$, for some $\tau > 0$. Take $\Phi_j = \Phi$ for every $j \in [J]$ in (1) and let $\Xi \in \mathbb{R}^{\widetilde{M} \times \widetilde{N}}$ denote the resulting block diagonal matrix in (1). Our second main result, also proved in Section 5, establishes the RIP for RBD matrices with this construction.

Theorem 2. *Let U denote an orthobasis for $\mathbb{C}^{\widetilde{N}}$. If $S \gtrsim 1$ and*

$$\widetilde{M} \gtrsim_\tau \delta^{-2}\gamma^2(U) \cdot S \cdot \log^2 S \log^2 \widetilde{N},$$

then $\delta_S(\Xi, U) \leq \delta < 1$, except with a probability of at most $O(\widetilde{N}^{-\log \widetilde{N} \log^2 S})$.

A few remarks are in order. At one end of the spectrum, the block-coherence of an orthobasis could equal $\sqrt{J}$ and consequently the required number of measurements above would scale with JS. This happens for signals that are sparse, for example, in the time (canonical basis) or frequency domains. Our result is indeed optimal for both of these bases (up to a poly-logarithmic factor). The same argument for the canonical basis carries over from the DBD matrices. For the Fourier basis, we note that it is possible to construct certain classes of periodic signals in $\mathbb{C}^{\widetilde{N}}$ that would require the lower bound on $\widetilde{M}$ to scale with JS. Consider, for example, the class of signals consisting of all S-sparse combinations of columns $1, J+1, \ldots, (J-1)N+1$ from $F_{\widetilde{N}}$. If x belongs to this class, then, by construction, $x_1, x_2, \cdots, x_J$ (as defined in (7)) are all equal because x is periodic with a period of N. As a result, different blocks of Ξ take the same measurements from x. Therefore, as was the case with the DBD matrices, obtaining a stable embedding of this class of signals requires $M \gtrsim_\tau \delta^{-2}S \log(N/S)$, and equivalently, $\widetilde{M} \gtrsim_\tau \delta^{-2}JS \log(N/S)$.

At the other end of the spectrum, for a generic orthobasis R we computed that $\gamma(R) \lesssim 1$ with high probability. For signals that are sparse in this basis (and therefore for many possible orthobases in general), an RBD matrix performs nearly as well as an unstructured Gaussian random matrix. The RBD structure allows us to prove the RIP for certain classes of structured random matrices as a special case of Theorem 2. In particular, in the next subsection we re-derive an RIP bound for partial random circulant matrices that originally appeared in [26]. As a byproduct, we also construct a deterministic sparsity basis that achieves a performance similar to the generic orthobasis we have considered above.

3.3.2. The RIP for Partial Random Circulant Matrices

This section demonstrates that the RBD model, together with Theorem 2, can be used to derive the RIP for partial random circulant matrices.[4] More specifically, we focus on proving the RIP for $\Gamma \in \mathbb{R}^{J \times P}$, with

---

[4]The arguments in this section extend without much effort to the more general case of establishing the RIP for partial random Toeplitz matrices.

$J \leq P$, defined as

$$
\Gamma = \frac{1}{\sqrt{J}} \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_P \\ \epsilon_P & \epsilon_1 & \cdots & \epsilon_{P-1} \\ \vdots & \vdots & & \vdots \\ \epsilon_{P-J+2} & \epsilon_{P-J+3} & \cdots & \epsilon_{P-J+1} \end{bmatrix},
$$

where $\{\epsilon_p\}$, $p \in [P]$, is a sequence of i.i.d. zero-mean, unit-variance random variables with sub-Gaussian norm τ. We let ε denote the vector formed by $\{\epsilon_p\}$. In order to use Theorem 2 in this setting, we make the following argument: for any signal $x \in \mathbb{C}^P$, we can write Γx as the multiplication of an RBD matrix $\Xi \in \mathbb{R}^{J \times PJ}$ and an extended vector $\widehat{x} \in \mathbb{C}^{PJ}$:

$$
\Gamma x = \overbrace{\begin{bmatrix} \varepsilon^* & & & \\ & \varepsilon^* & & \\ & & \ddots & \\ & & & \varepsilon^* \end{bmatrix}}^{\Xi} \cdot \frac{1}{\sqrt{J}} \overbrace{\begin{bmatrix} \mathrm{S}^0 x \\ \mathrm{S}^1 x \\ \vdots \\ \mathrm{S}^{J-1} x \end{bmatrix}}^{\widehat{x}} =: \Xi \cdot \frac{1}{\sqrt{J}} \widehat{x}, \tag{17}
$$

where S is the cyclic shift-up operator on column vectors in $\mathbb{C}^P$. The next lemma (proved in Appendix D) states that if x is sparse, then $\widehat{x}$ has a sparse representation in a favorable orthobasis T (constructed in the proof).

Lemma 3. *There exists an orthobasis T with $\gamma(T) = 1$, such that every $\widehat{x}$ has an S-sparse representation in T if the corresponding x is S-sparse. That is, for every such $\widehat{x}$, there exists x_e with $\|x_e\|_0 \leq S$ such that $\widehat{x}/\sqrt{J} = T x_e$.*

Therefore, the T-RIP for Ξ implies the RIP for Γ. To be more specific, after setting $M = 1$, Theorem 2 implies that $\delta_S(\Gamma) \leq \delta_S(\Xi, T) \leq \delta$ except with a probability of at most $O(P^{-\log P \log^2 S})$, provided that

$$
J \gtrsim_\tau \delta^{-2} \cdot S \log^2 S \log^2 P,
$$

which is equivalent to Theorem 1.1 in [26].

4. Numerical Simulations

This section contains a series of simulations that are intended to reinforce our findings for the reader. An ideal scenario might involve generating several random block diagonal matrices while varying the sparsity level S and the number of measurements $\widetilde{M}$ and measuring the fraction of realizations in which the RIC falls below a fixed threshold. Unfortunately, checking the RIP for a matrix is known to be an NP hard problem [37], although certain algorithms [24, 20] have been proposed for computing lower bounds on the RIC for moderately sized random matrices. In particular, [20] contains a fast heuristic algorithm for this purpose.

As we have discussed in Section 1, one context where the RIP often arises is in the analysis of sparse signal recovery algorithms such as Basis Pursuit (BP) [8]. In fact, for dense random matrices, it is known that the RIP serves as a rather pessimistic criterion to ensure the success of sparse signal recovery algorithms (particularly BP); we refer the interested reader to [5, 6, 17, 29] for a more comprehensive discussion. In this section, we use empirical signal recovery performance—with a limited number of signals—to illustrate how various choices for the sparsity basis will interact with a block diagonal measurement matrix. To be clear, this is different from checking whether the RIP holds. What we see, however, is that the same bases which are most favorable for our RIP bounds tend to also be most favorable for empirical signal recovery performance.

We now detail the setup in this section. With $N = 100$ and $J = 10$ (and consequently, $\widetilde{N} = 1000$), we generate a single $\widetilde{N} \times \widetilde{N}$ DBD matrix whose entries are i.i.d. Gaussian random variables having zero mean and

9

unit standard deviation. Note that each diagonal block is of size $N \times N$. For each pair $(S, M) \in [\widetilde{M}] \times [N]$, the following procedure is executed. An $\widetilde{M} \times \widetilde{N}$ DBD matrix Ψ is formed by keeping (and appropriately normalizing) the first M rows of each diagonal block of the large $\widetilde{N} \times \widetilde{N}$ matrix. Then 20 random S-sparse signals in the canonical basis of $\mathbb{C}^{\widetilde{N}}$ are generated. These sparse signals are measured using Ψ and reconstructed back (from incomplete measurements) via BP.[5] We deem the recovery successful if the relative ℓ_2 reconstruction error is less than the fixed threshold 10^{-2}. The fraction of successful recovery is recorded and this procedure is repeated for other pairs (S, M). To be clear, we use a fixed measurement matrix to generate the data corresponding to each pair (S, M). As mentioned earlier, the $\widetilde{M} \times \widetilde{N}$ measurement matrix itself is generated from the $\widetilde{N} \times \widetilde{N}$ matrix fixed throughout the experiment. The resulting phase transition graph is depicted in Figure 1a, where the color of each pixel ranges from black for perfect recovery in every realization to white for failed recovery every time.

We repeat the above simulation for signals that are sparse in the Fourier and generic bases (see Lemma 1). A new generic basis is generated in each iteration. All of the above simulations are then repeated with an RBD matrix. The results of the simulations are displayed in Figures 1 and 2. In the simulations with DBD matrices, recovery of canonical and frequency sparse signals are, respectively, the least and most successful of the three instances. Signals that are sparse in the generic bases can be recovered nearly as well as frequency sparse signals.

In the simulations with RBD matrices, signals that are sparse in the generic bases are recovered best, while the results for the canonical and Fourier bases are less satisfactory. These observations mirror our findings in Section 3. We do point out, however, that for the case of RBD measurement matrices we are able to recover frequency sparse signals somewhat better than signals that are sparse in the canonical basis. We note that such difference is not reflected in our RIP bounds. While this performance difference could be due to a lack of explicit numerical constants in our results, it is more likely an artifact of our simulations: We are not directly confirming the RIP but rather testing the recovery of randomly generated test signals, and those few signals which make it difficult to satisfy the RIP may be more pathological in the Fourier basis than in the canonical basis.

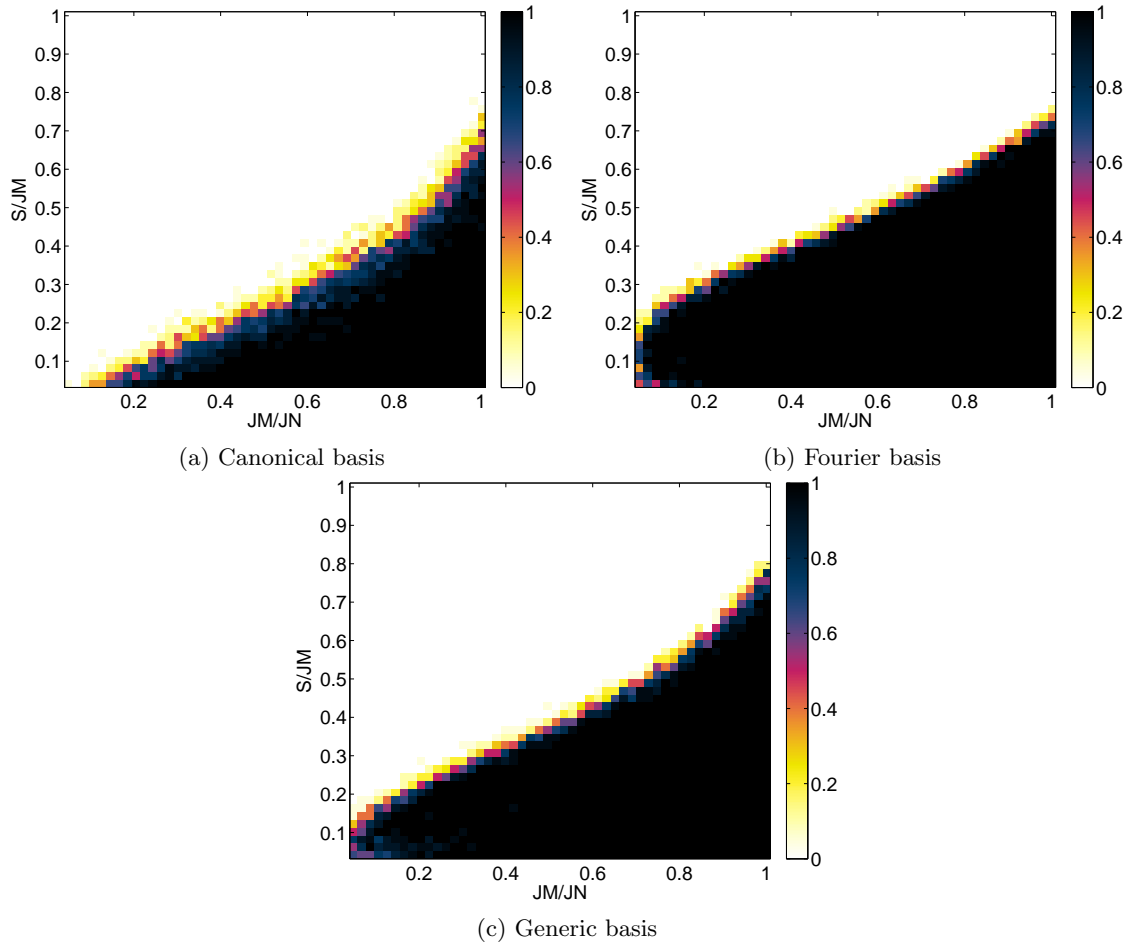[5]In order to implement BP, we used YALL1, a package for solving ℓ_1 problems [45, 41].

(a) Canonical basis

(b) Fourier basis

(c) Generic basis

Figure 1: Simulation results for DBD matrices. Refer to Section 4 for details.

(a) Canonical basis

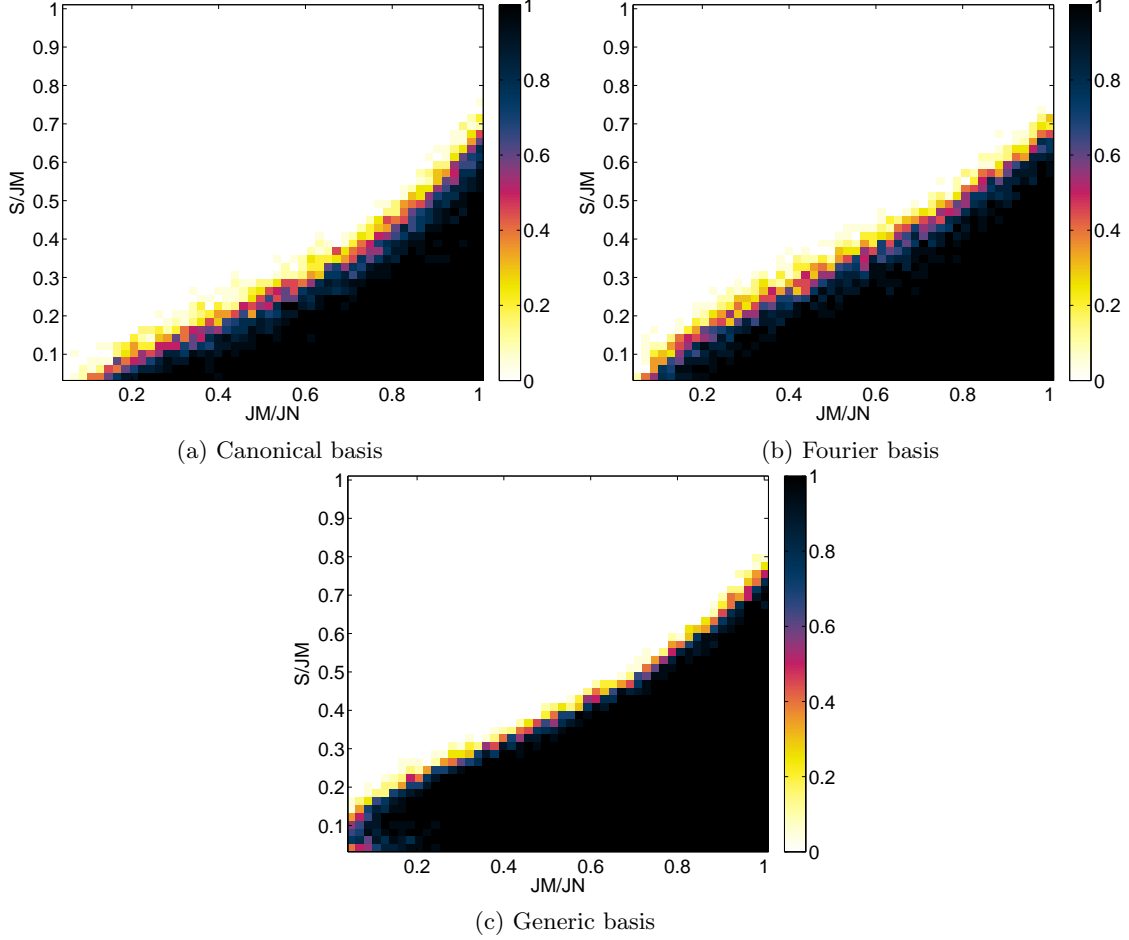(b) Fourier basis

(c) Generic basis

Figure 2: Simulation results for RBD matrices. Refer to Section 4 for details.

5. Proofs of Theorems 1 and 2

5.1. Preliminaries

First, define the set of all S-sparse signals with unit norm as

$$\Omega_S := \left\{ \alpha \in \mathbb{C}^{\widetilde{N}} \ : \ \|\alpha\|_0 \leq S, \ \|\alpha\|_2 = 1 \right\}. \tag{18}$$

With this definition, we observe that the RIC for DBD matrices can be written as

$$\delta_S \quad = \quad \sup_{\alpha \in \Omega_S} \left| \|\Psi \cdot x(\alpha)\|_2^2 - 1 \right|,$$

where we leveraged the fact that $\|\alpha\|_2 = 1$ implies

$$\mathbb{E}\left\{ \|\Psi \cdot x(\alpha)\|_2^2 \right\} = \alpha^* U^* \mathbb{E}\{\Psi^* \Psi\} U \alpha = \alpha^* \alpha = 1.$$

The RIC for RBD matrices can be similarly written as

$$\delta_S \quad = \quad \sup_{\alpha \in \Omega_S} \left| \|\Xi \cdot x(\alpha)\|_2^2 - 1 \right|.$$

Given $\delta < 1$ and under the conditions in Theorems 1 and 2, our objective is to show that $\delta_S \leq \delta$ for both DBD and RBD matrices. To achieve this goal, we require the following powerful result due to Krahmer et al.:

12

Theorem 3. [26, Theorem 3.1] *Let $\mathcal{A} \subset \mathbb{C}^{\overline{M} \times \overline{N}}$ be a set of matrices, and let ε be a random vector whose entries are i.i.d., zero-mean, unit-variance random variables with sub-Gaussian norm τ. Set*

$$d_F(\mathcal{A}) := \sup_{A \in \mathcal{A}} \|A\|_F,$$

$$d_2(\mathcal{A}) := \sup_{A \in \mathcal{A}} \|A\|_2,$$

and

$$
\begin{aligned}
E_1 &:= \gamma_2\left(\mathcal{A}, \|\cdot\|_2\right)\left(\gamma_2\left(\mathcal{A}, \|\cdot\|_2\right) + d_F(\mathcal{A})\right) + d_F(\mathcal{A}) d_2(\mathcal{A}), \\
E_2 &:= d_2(\mathcal{A})\left(\gamma_2\left(\mathcal{A}, \|\cdot\|_2\right) + d_F(\mathcal{A})\right), \\
E_3 &:= d_2^2(\mathcal{A}).
\end{aligned}
$$

Then, for $t > 0$, it holds that

$$\log \mathrm{P}\left\{\sup_{A \in \mathcal{A}} \left| \|A\varepsilon\|_2^2 - \mathbb{E}\|A\varepsilon\|_2^2 \right| \gtrsim_\tau E_1 + t\right\} \lesssim_\tau -\min\left(\frac{t^2}{E_2^2}, \frac{t}{E_3}\right).$$

Without going into the details, we note that the γ_2-functional of $\mathcal{A}$, $\gamma_2\left(\mathcal{A}, \|\cdot\|_2\right)$, is a geometrical property of $\mathcal{A}$, i.e., the index set of the random process, and is widely used in the theory of probability in Banach spaces [36, 28].[6] In particular, the following lemma gives an estimate of this quantity.

Lemma 4. [36] *With $\mathcal{A}$ as defined above, it holds that*

$$\gamma_2\left(\mathcal{A}, \|\cdot\|_2\right) \lesssim \int_0^\infty \log^{\frac{1}{2}}\left(\mathcal{N}\left(\mathcal{A}, \|\cdot\|_2, \nu\right)\right) \, d\nu. \tag{19}$$

Clearly, we need to express the problem of bounding the RIC of DBD and RBD matrices in a form that is amenable to the setting of Theorem 3. First, for DBD matrices, let us define $X_{D,j} \in \mathbb{C}^{M \times MN}$, $j \in [J]$, as

$$
X_{D,j}(\alpha) = X_{D,j}(\alpha, U) := \begin{bmatrix} x_j^*(\alpha) & & & \\ & x_j^*(\alpha) & & \\ & & \ddots & \\ & & & x_j^*(\alpha) \end{bmatrix}. \tag{20}
$$

It can then be easily verified that

$$
\begin{aligned}
\|\Psi \cdot x(\alpha)\|_2^2 &= \sum_{j \in [J]} \|\Phi_j \cdot x_j(\alpha)\|_2^2 \\
&= \sum_{j \in [J]} \|X_{D,j}(\alpha) \cdot \mathrm{vec}(\Phi_j^*)\|_2^2 \\
&\overset{\text{i.d.}}{=} \sum_{j \in [J]} \left\|\frac{1}{\sqrt{M}} X_{D,j}(\alpha) \cdot \varepsilon_j\right\|_2^2 \\
&\overset{\text{i.d.}}{=} \|A_D(\alpha) \cdot \varepsilon\|_2^2,
\end{aligned}
$$

where the linear map $A_D : \Omega_S \to \mathbb{C}^{\widetilde{M} \times JMN}$ is defined as

$$
A_D(\alpha) = A_D(\alpha, U) := \frac{1}{\sqrt{M}} \begin{bmatrix} X_{D,1}(\alpha) & & & \\ & X_{D,2}(\alpha) & & \\ & & \ddots & \\ & & & X_{D,J}(\alpha) \end{bmatrix}, \tag{21}
$$

---

[6]The γ_2-functional of $\mathcal{A}$, $\gamma_2(\mathcal{A}, \|\cdot\|_2)$, is not to be confused with the block-coherence of an orthobasis U, $\gamma(U)$.

and entries of $\varepsilon_j \in \mathbb{R}^{MN}$, $j \in [J]$, and $\varepsilon \in \mathbb{R}^{JMN}$ are i.i.d. zero-mean, unit-variance random variables with sub-Gaussian norm τ. The index set of the random process is

$$\mathcal{A}_D := \{A_D(\alpha) \, : \, \alpha \in \Omega_S\}. \tag{22}$$

We have therefore completely expressed the DBD problem in the setting of Theorem 3, where A and $\mathcal{A}$ are replaced with A_D and $\mathcal{A}_D$, respectively.

Next, for RBD matrices, we observe that

$$
\begin{aligned}
\|\Xi \cdot x(\alpha)\|_2^2 &= \sum_{j \in [J]} \|\Phi \cdot x_j(\alpha)\|_2^2 \\
&= \|\Phi \cdot X_R(\alpha)\|_F^2 \\
&= \|X_R^*(\alpha) \cdot \Phi^*\|_F^2 \\
&\underset{\text{i.d.}}{=} \sum_{m \in [M]} \left\| \frac{1}{\sqrt{M}} X_R^*(\alpha) \cdot \varepsilon_m' \right\|_2^2 \\
&\underset{\text{i.d.}}{=} \|A_R(\alpha) \cdot \varepsilon'\|_2^2,
\end{aligned}
\tag{23}
$$

where in the last line we defined the linear map $A_R : \Omega_S \to \mathbb{C}^{\widetilde{M} \times MN}$ as

$$
A_R(\alpha) = A_R(\alpha, U) := \frac{1}{\sqrt{M}}
\begin{bmatrix}
X_R^*(\alpha) & & & \\
& X_R^*(\alpha) & & \\
& & \ddots & \\
& & & X_R^*(\alpha)
\end{bmatrix},
\tag{24}
$$

and the entries of $\varepsilon_m' \in \mathbb{R}^N$, $m \in [M]$, and $\varepsilon' \in \mathbb{R}^{MN}$ are i.i.d. zero-mean, unit-variance random variables with sub-Gaussian norm of τ. Here, the index set of random process is

$$\mathcal{A}_R := \{A_R(\alpha) \, : \, \alpha \in \Omega_S\}. \tag{25}$$

So, we have also managed to express the RBD problem in the setting of Theorem 3, where A and $\mathcal{A}$ are replaced with A_R and $\mathcal{A}_R$, respectively.

The next two subsections are concerned with estimating the quantities involved in Theorem 3 for both the DBD and RBD problems.

5.2. Calculating $d_2(\mathcal{A}_D)$, $d_F(\mathcal{A}_D)$, and $\gamma_2(\mathcal{A}_D, \|\cdot\|_2)$

Recall the definitions of A_D and $\mathcal{A}_\mathcal{D}$ in (21) and (22). We begin with defining the following norm on $\mathbb{C}^{\widetilde{N}}$, which will find extensive use in the analysis of the DBD problem:

$$\|\alpha\|_{A_D} := \|A_D(\alpha)\|_2 \tag{26}$$

for $\alpha \in \mathbb{C}^{\widetilde{N}}$. We record a useful property of this norm below.

Lemma 5. *For every $\alpha \in \mathbb{C}^{\widetilde{N}}$, it holds that*

$$\|\alpha\|_{A_D} \leq \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1. \tag{27}$$

Proof. Let $u_{j,n}$, $j \in [J]$ and $n \in [N]$, denote the $((j-1)N + n)$th row of U. We then have that

$$
\begin{aligned}
\|\alpha\|_{A_D} &= \|A_D(\alpha)\|_2 \\
&= \|A_D(\alpha)A_D^*(\alpha)\|_2^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{M}} \max_{j \in [J]} \|x_j\|_2 \\
&= \frac{1}{\sqrt{M}} \max_{j \in [J]} \|U_j \alpha\|_2 \\
&\leq \sqrt{\frac{N}{M}} \max_{j \in [J], n \in [N]} |\langle u_{j,n}, \alpha \rangle| \\
&\leq \sqrt{\frac{N}{M}} \max_{j \in [J], n \in [N]} \|u_{j,n}\|_\infty \|\alpha\|_1 \\
&= \frac{\mu}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1,
\end{aligned}
$$

where the second to last line uses the Hölder inequality and the last line follows from the definition of μ. On the other hand, one may also write that

$$
\|\alpha\|_{A_D} = \frac{1}{\sqrt{M}} \max_{j \in [J]} \|x_j\|_2 \leq \frac{1}{\sqrt{M}} \|x\|_2 = \frac{1}{\sqrt{M}} \|U\alpha\|_2 = \frac{1}{\sqrt{M}} \|\alpha\|_2 \leq \frac{1}{\sqrt{M}} \|\alpha\|_1,
$$

where we used the fact that U is an orthobasis. Overall, we arrive at

$$
\|\alpha\|_{A_D} \leq \frac{1}{\sqrt{\widetilde{M}}} \min\left(\mu, \sqrt{J}\right) \|\alpha\|_1 = \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1, \tag{28}
$$

as claimed. The equality above follows from the definition of $\widetilde{\mu}$ in (14). $\qquad \square$

We continue with computing the quantities involved in Theorem 3 in the case of the DBD problem. First, we have that

$$
d_F(\mathcal{A}_D) = \sup_{A_D(\alpha) \in \mathcal{A}_D} \|A_D(\alpha)\|_F = \sup_{\alpha \in \Omega_S} \|x(\alpha)\|_2 = \sup_{\alpha \in \Omega_S} \|U\alpha\|_2 = \sup_{\alpha \in \Omega_S} \|\alpha\|_2 = 1. \tag{29}
$$

The second to last equality holds because U is an orthonormal basis. Second, we have that

$$
d_2(\mathcal{A}_D) = \sup_{A_D(\alpha) \in \mathcal{A}_D} \|A_D(\alpha)\|_2 = \sup_{\alpha \in \Omega_S} \|\alpha\|_{A_D} \leq \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}} \sup_{\alpha \in \Omega_S} \|\alpha\|_1 \leq \widetilde{\mu} \sqrt{\frac{S}{\widetilde{M}}}. \tag{30}
$$

The first inequality above holds on account of Lemma 5. The second inequality above follows because $\|\alpha\|_2 = 1$ and $\|\alpha\|_0 \leq S$ when $\alpha \in \Omega_S$. It is only left to bound $\gamma_2(\mathcal{A}_D, \|\cdot\|_2)$. According to Lemma 4, we have that

$$
\begin{aligned}
\gamma_2(\mathcal{A}_D, \|\cdot\|_2) &\leq \int_0^\infty \log^{\frac{1}{2}} \left(\mathcal{N}\left(\mathcal{A}_D, \|\cdot\|_2, \nu\right)\right) \, d\nu \\
&= \int_0^\infty \log^{\frac{1}{2}} \left(\mathcal{N}\left(\Omega_S, \|\cdot\|_{A_D}, \nu\right)\right) \, d\nu,
\end{aligned}
$$

where the isometry between $\mathcal{A}_D$ (with metric $\|\cdot\|_2$) and Ω_S (with metric $\|\cdot\|_{A_D}$) implies the second line. This isometry, in turn, follows from (26) and the linearity of $A_D(\cdot)$. Consequently,

$$
\begin{aligned}
\gamma_2(\mathcal{A}_D, \|\cdot\|_2) &\leq \int_0^\infty \log^{\frac{1}{2}} \left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \frac{\nu}{\sqrt{S}}\right)\right) d\nu \\
&\leq \sqrt{S} \int_0^\infty \log^{\frac{1}{2}} \left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \nu\right)\right) d\nu, \tag{31}
\end{aligned}
$$

15

where the first line uses the second inequality in (A.9) and the last line follows from a change of variables in the integral. An estimate of the covering number involved in (31) can be found through the next result, which is proved in Appendix E.

Lemma 6. *Consider a norm* $\|\cdot\|_A$ *on* $\mathbb{C}^{\widetilde{N}}$ *that, for every* $\alpha \in \mathbb{C}^{\widetilde{N}}$, *satisfies*

$$\|\alpha\|_A = \|A(\alpha)\|_2 \leq \frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1,$$

for some linear map $A(\cdot) : \mathbb{C}^{\widetilde{N}} \to \mathbb{C}^{N'}$ *with rank of at most* $\widetilde{M}$ *and some* $\kappa > 0$ *and integer* N'. *Then, for* $0 < \nu < \kappa/\sqrt{\widetilde{M}}$ *and* $\widetilde{M} \gtrsim 1$, *we have that*

$$\log\left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right)\right) \lesssim \min\left(S\log\widetilde{N} + S\log\left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right), \frac{\kappa^2}{\nu^2\widetilde{M}} \cdot \log^2\widetilde{N}\right). \tag{32}$$

When $\nu \geq \kappa/\sqrt{\widetilde{M}}$, *we have* $\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) = 1$.

Qualitatively speaking, of the two bounds on the right hand of (32), the first is tighter when ν is small while the second is more effective for larger values of ν. Of course, $\|\cdot\|_{A_D}$ satisfies the hypothesis of Lemma 6 with $\kappa = \widetilde{\mu}$ and the map $A_D(\cdot)$. Consequently, for $0 < \nu_0 \leq \widetilde{\mu}/\sqrt{\widetilde{M}}$ to be set later, we have that

$$\int_0^\infty \log^{\frac{1}{2}}\left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \nu\right)\right) d\nu$$

$$= \int_0^{\nu_0} \log^{\frac{1}{2}}\left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \nu\right)\right) d\nu + \int_{\nu_0}^{\frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}} \log^{\frac{1}{2}}\left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \nu\right)\right) d\nu$$

$$\lesssim \int_0^{\nu_0}\left(\sqrt{S\log\widetilde{N}} + \sqrt{S\log\left(1 + \frac{2\widetilde{\mu}}{\nu\sqrt{\widetilde{M}}}\right)}\right) d\nu + \log\widetilde{N}\int_{\nu_0}^{\frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}} \frac{\widetilde{\mu}}{\nu\sqrt{\widetilde{M}}} d\nu$$

$$\lesssim \nu_0\sqrt{S\log\widetilde{N}} + \nu_0\sqrt{S\log\left(1 + \frac{2\widetilde{\mu}}{\nu_0\sqrt{\widetilde{M}}}\right)} + \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}\log\widetilde{N}\log\left(\frac{\widetilde{\mu}}{\nu_0\sqrt{\widetilde{M}}}\right). \tag{33}$$

The second line above follows from the second statement in Lemma 6. In the third line, different upper bounds from (32) are used to bound each summand. We benefited from (A.11) in the Toolbox to compute the logarithmic integral in the third line. With the choice of $\nu_0 = \widetilde{\mu}/\sqrt{S\widetilde{M}}$, we obtain that

$$\int_0^\infty \log^{\frac{1}{2}}\left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_D}, \nu\right)\right) d\nu$$

$$\lesssim \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}\sqrt{\log\widetilde{N}} + \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}\sqrt{\log(1 + 2\sqrt{S})} + \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}\log S\log\widetilde{N}$$

$$\lesssim \frac{\widetilde{\mu}}{\sqrt{\widetilde{M}}}\log S\log\widetilde{N} \tag{34}$$

for $S \gtrsim 1$. Now plugging back (34) into (31), we arrive at

$$\gamma_2(\mathcal{A}_D, \|\cdot\|_2) \lesssim \widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}}\log S\log\widetilde{N}. \tag{35}$$

Before completing the analysis of the DBD problem, let us calculate the same quantities for the RBD case.

5.3. Calculating $d_2(\mathcal{A}_R)$, $d_F(\mathcal{A}_R)$, and $\gamma_2(\mathcal{A}_R, \|\cdot\|_2)$

Recall the definitions of A_R and $\mathcal{A}_\mathcal{R}$ in (24) and (25). Again, we first introduce the following norm on $\mathbb{C}^{\widetilde{N}}$, which will be useful in the analysis of the RBD problem:

$$\|\alpha\|_{A_R} := \|A_R(\alpha)\|_2 \tag{36}$$

for $\alpha \in \mathbb{C}^{\widetilde{N}}$. This norm has the following property.

Lemma 7. *For every $\alpha \in \mathbb{C}^{\widetilde{N}}$, it holds that*

$$\|\alpha\|_{A_R} \leq \frac{\gamma}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1. \tag{37}$$

Proof. Note that

$$
\begin{aligned}
\|\alpha\|_{A_R} &= \|A_R(\alpha)\|_2 \\
&= \frac{1}{\sqrt{M}} \|X_R(\alpha)\|_2 \\
&= \frac{1}{\sqrt{M}} \| \sum_{\widetilde{n} \in [\widetilde{N}]} \alpha(\widetilde{n}) X_R(e_{\widetilde{n}})\|_2 \\
&\leq \frac{1}{\sqrt{M}} \sum_{\widetilde{n} \in [\widetilde{N}]} |\alpha(\widetilde{n})| \cdot \|X_R(e_{\widetilde{n}})\|_2 \\
&\leq \frac{1}{\sqrt{M}} \max_{\widetilde{n} \in [\widetilde{N}]} \|X_R(e_{\widetilde{n}})\|_2 \cdot \|\alpha\|_1 \\
&= \frac{\gamma}{\sqrt{\widetilde{M}}} \cdot \|\alpha\|_1.
\end{aligned}
$$

The third line above uses the linearity of $X_R(\cdot)$. The fourth line follows from the triangle inequality and the fifth line is implied by the Hölder inequality. We made use of the definition of γ in (9) to get the last line. $\qquad\square$

We now continue with computing the quantities involved in Theorem 3 in the case of the RBD problem. First, we have that

$$d_F(\mathcal{A}_R) = \sup_{A_R(\alpha) \in \mathcal{A}_R} \|A_R(\alpha)\|_F = \sup_{\alpha \in \Omega_S} \|X_R(\alpha)\|_F = \sup_{\alpha \in \Omega_S} \|x(\alpha)\|_2 = \sup_{\alpha \in \Omega_S} \|\alpha\|_2 = 1. \tag{38}$$

Second, we have that

$$d_2(\mathcal{A}_R) = \sup_{A_R(\alpha) \in \mathcal{A}_R} \|A_R(\alpha)\|_2 = \sup_{\alpha \in \Omega_S} \|\alpha\|_{A_R} \leq \frac{\gamma}{\sqrt{\widetilde{M}}} \sup_{\alpha \in \Omega_S} \|\alpha\|_1 \leq \gamma \sqrt{\frac{S}{\widetilde{M}}}, \tag{39}$$

where the first inequality follows from Lemma 7. The last quantity to estimate is $\gamma_2(\mathcal{A}_R, \|\cdot\|_2)$. As was the case in the previous subsection, we may write that

$$\gamma_2(\mathcal{A}_R, \|\cdot\|_2) \leq \sqrt{S} \int_0^\infty \log^{\frac{1}{2}} \left(\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_{A_R}, \nu\right)\right) d\nu. \tag{40}$$

Of course, $\|\cdot\|_{A_R}$ satisfies the hypothesis of Lemma 6 with $\kappa = \gamma$ and the map $A_R(\cdot)$. Therefore, following the same steps as in the previous subsection, we arrive at

$$\gamma_2(\mathcal{A}_R, \|\cdot\|_2) \lesssim \gamma \sqrt{\frac{S}{\widetilde{M}}} \log S \log \widetilde{N}. \tag{41}$$

5.4. Denouement

We notice that the quantities $d_F(\mathcal{A}_D), d_2(\mathcal{A}_D)$, and $\gamma_2(\mathcal{A}_D, \|\cdot\|_2)$ have the same bounds as their counterparts $d_F(\mathcal{A}_R), d_2(\mathcal{A}_R)$, and $\gamma_2(\mathcal{A}_R, \|\cdot\|_2)$ except for the type of the coherence factor involved. Therefore, it suffices to focus on one—the same result holds for the other with its corresponding coherence factor. Given $\delta < 1$, assume that

$$\widetilde{M} \gtrsim_\tau \delta^{-2}\widetilde{\mu}^2 \cdot S \log^2 S \log^2 \widetilde{N}. \tag{42}$$

Equipped with the estimates in Section 5.2, i.e., (29), (30), and (35), we now compute E_1 in Theorem 3:

$$
\begin{aligned}
E_1 &:= \gamma_2\left(\mathcal{A}_D, \|\cdot\|_2\right)\left(\gamma_2\left(\mathcal{A}_D, \|\cdot\|_2\right) + d_F(\mathcal{A}_D)\right) + d_F(\mathcal{A}_D)d_2(\mathcal{A}_D) \\
&\lesssim \widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}}\log S \log \widetilde{N}\left(\widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}}\log S \log \widetilde{N} + 1\right) + \widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}} \\
&\lesssim_\tau \delta\left(\delta + 1\right) + \frac{\delta}{\log S \log \widetilde{N}} \\
&\leq 2\delta + \frac{\delta}{\log S \log \widetilde{N}} \\
&\lesssim \delta,
\end{aligned}
$$

where we assumed that $S \gtrsim 1$ and used the hypothesis that $\delta < 1$ in the second to last line. Doing the same to E_2, we obtain that

$$
\begin{aligned}
E_2 &:= d_2(\mathcal{A}_D)\left(\gamma_2\left(\mathcal{A}_D, \|\cdot\|_2\right) + d_F(\mathcal{A}_D)\right) \\
&\lesssim \widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}}\left(\widetilde{\mu}\sqrt{\frac{S}{\widetilde{M}}}\log S \log \widetilde{N} + 1\right) \\
&\lesssim_\tau \frac{\delta}{\log S \log \widetilde{N}}\left(\delta + 1\right) \\
&\lesssim \frac{\delta}{\log S \log \widetilde{N}}.
\end{aligned}
$$

Similarly for E_3, we write that

$$E_3 := d_2^2(\mathcal{A}_D) \leq \frac{\widetilde{\mu}^2 S}{\widetilde{M}} \lesssim_\tau \frac{\delta^2}{\log^2 S \log^2 \widetilde{N}}.$$

Plugging the above estimates of E_1, E_2, and E_3 into the tail bound in Theorem 3, we obtain that

$$\log \mathrm{P}\left\{\sup_{\alpha \in \Omega_S}\left|\|\Psi \cdot x(\alpha)\|_2^2 - 1\right| \gtrsim_\tau \delta + t\right\} \lesssim_\tau -\min\left(\delta^{-2}t^2 \log^2 S \log^2 \widetilde{N}, \delta^{-1}t \log^2 S \log^2 \widetilde{N}\right).$$

Substituting $t = \delta$, we arrive at

$$\log \mathrm{P}\left\{\sup_{\alpha \in \Omega_S}\left|\|\Psi \cdot x(\alpha)\|_2^2 - 1\right| \gtrsim_\tau \delta\right\} \lesssim_\tau -\log^2 S \log^2 \widetilde{N},$$

assuming that $S \gtrsim 1$. After absorbing the factor depending on τ into (a redefined) δ, we finally arrive at

$$\log \mathrm{P}\left\{\sup_{\alpha \in \Omega_S}\left|\|\Psi \cdot x(\alpha)\|_2^2 - 1\right| > \delta\right\} \lesssim_\tau -\log^2 S \log^2 \widetilde{N},$$

which completes the proof of Theorem 1. Replacing $\widetilde{\mu}$ with γ and repeating this argument concludes the proof of Theorem 2.

6. Conclusion

In this paper, we studied two classes of structured random matrices, namely DBD and RBD matrices. Our main results state that matrices with block diagonal constructions can indeed satisfy the RIP but that the requisite number of measurements depends on certain properties of the sparsifying basis. These properties were detailed and interpreted in the paper. In the best case, DBD and RBD matrices perform nearly as well as the dense i.i.d. random matrices generally used in CS despite having many fewer nonzero entries. Moreover, we have shown that random block diagonal matrices are intimately related to the random convolution and random Toeplitz matrices considered in the literature.

Our findings lead us to conclude that structured random matrices can be useful in sensing architectures as long as the statistics of the data are well-matched with the structure of the measurement matrix. While this intuition is similar to other results on structured measurement matrices, our results on block diagonal matrix constructions are novel in extending this intuition to matrices with (potentially) many entries that are zero. The approach required to reach our results also leads us to conclude that future progress in the field of probability in Banach spaces is likely to play a significant role in establishing optimal performance guarantees for other structured measurement systems. This may be especially true given the improved performance we were able to achieve even over other bounding techniques that require sophisticated mathematical machinery. We do however remain uncertain about the necessity of the poly-logarithmic factors in the final measurement rates (which may be a proof artifact). Finally, as a means of visualizing how various choices for the sparsity basis will interact with a block diagonal measurement matrix, we study the empirical performance of signal recovery with random block diagonal measurement matrices. The simulation results show that the same bases which are most favorable for our RIP bounds tend to also be most favorable for empirical signal recovery performance.

There are several directions that can be explored in the future. First, as we have discussed in the introduction, block diagonal matrices are useful for modeling distributed measurement systems. It may therefore be of interest to specialize our results to some particular distributed systems. Take for example MIMO radar where multiple independent transmitters and receivers are arbitrarily distributed over an area of interest to sense targets in a scene. Data from the receivers with potentially high data rates needs to be sent to a central processor and to be coherently processed to achieve a maximum processing gain. The block diagonal structure studied here is potentially useful to analyze the possibility of compressing the data at the individual receivers before sending it to the central processor. For another example, block diagonal structure has also been exploited in observability studies of dynamical systems [40]. Our understanding of this and similar problems [2, 12, 11] may be enhanced using the results in this paper.

Acknowledgements

References

[1] H. Arguello and G. Arce. Restricted isometry property in coded aperture compressive spectral imaging. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2012.

[2] S.M. Asif and A.S. Charles. Estimation and dynamic updating of time-varying signals with sparse variations. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[3] B. Bah and J. Tanner. Improved bounds on restricted isometry constants for gaussian matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2882–2898, 2010.

[4] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *46th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2008.

[5] J. D. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: How sharp is the restricted isometry property? *SIAM Review*, 53(1):105–125, 2011.

[6] J. D. Blanchard, C. Cartis, J. Tanner, and A. Thompson. Phase transitions for greedy sparse approximation algorithms. *Applied and Computational Harmonic Analysis*, 30(2):188–203, 2011.

[7] E.J. Candès. Compressive sampling. In *Proceedings of International Congress of Mathematics*, 2006.

[8] E.J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

[9] E.J. Candès, Y.C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[10] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[11] A. Charles and C. Rozell. Dynamic filtering of sparse signals using reweighted ℓ_1. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[12] A.S. Charles, S.M. Asif, J.K. Romberg, and C.J. Rozell. Sparsity penalties in dynamical system estimation. In *Proceedings of IEEE Conference on Information Sciences and Systems (CISS)*, 2011.

[13] M.A. Davenport, P.T. Boufounos, M.B. Wakin, and R.G. Baraniuk. Signal processing with compressive measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, 2010.

[14] M.A. Davenport and M.B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401, 2010.

[15] R.A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4-6):918–925, 2007.

[16] T.T. Do, L. Gan, N.H. Nguyen, and T.D. Tran. Fast and efficient compressive sensing using structurally random matrices. *IEEE Transactions on Signal Processing*, 60(1):139–154, 2012.

[17] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[18] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

[19] D. L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

[20] C. Dossal, G. Peyré, and J. Fadili. A numerical exploration of compressed sampling recovery. *Linear Algebra and its Applications*, 432(7):1663–1679, 2010.

[21] S. Foucart. Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants. In M. Neamtu and L. Schumaker, editors, *Approximation Theory XIII: San Antonio 2010*, volume 13 of *Springer Proceedings in Mathematics*, pages 65–77. Springer New York, 2012.

[22] J. Haupt, W.U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862 –5875, 2010.

[23] S. Huang and J. Zhu. Recovery of sparse signals using omp and its variants: convergence analysis based on rip. *Inverse Problems*, 27(3):035003, 2011.

[24] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

[25] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

[26] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*. To appear.

[27] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[28] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

[29] H. Monajemi, S. Jafarpour, M. Gavish, D. L. Donoho, S. Ambikasaran, S. Bacallado, D. Bharadia, Y. Chen, Y. Choi, M. Chowdhury, et al. Deterministic matrices matching the compressed sensing phase transitions of gaussian random matrices. *Proceedings of the National Academy of Sciences*, 110(4):1181–1186, 2013.

[30] D. Needell and J.A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[31] J.Y. Park, H.L. Yap, C.J. Rozell, and M.B. Wakin. Concentration of measure for block diagonal matrices with applications to compressive signal processing. *IEEE Transactions on Signal Processing*, 59(12):5859 –5875, 2011.

[32] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series Comp. Appl. Math.*, pages 1–92. deGruyter, 2010.

[33] H. Rauhut, J. Romberg, and J.A. Tropp. Restricted isometries for partial random circulant matrices. *Applied and Computational Harmonic Analysis*, 32(2):242 – 254, 2012.

[34] C.J. Rozell, H.L. Yap, J.Y. Park, and M.B. Wakin. Concentration of measure for block diagonal matrices with repeated blocks. In *Proceedings of IEEE Conference on Information Sciences and Systems (CISS)*, 2010.

[35] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[36] M. Talagrand. *The generic chaining*. Springer, 2005.

[37] A.M. Tillmann and M.E. Pfetsch. The computational complexity of RIP, NSP, and related concepts in compressed sensing. *Arxiv preprint arXiv:1205.2081*, 2012.

[38] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y.C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[39] M.B. Wakin, J.Y. Park, H.L. Yap, and C.J. Rozell. Concentration of measure for block diagonal measurement matrices. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[40] M.B. Wakin, B.M. Sanandaji, and T.L. Vincent. On the observability of linear systems from random, compressive measurements. *IEEE Conf. Decision and Control (CDC)*, 2010.

[41] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM Journal on Scientific Computing*, 33(1):250–278, 2011.

[42] H.L. Yap, A. Eftekhari, M.B. Wakin, and C.J. Rozell. The restricted isometry property for block diagonal matrices. In *Proceedings of IEEE Conference on Information Sciences and Systems (CISS)*, 2011.

[43] H.L. Yap and C.J. Rozell. Stable Takens' embeddings for linear dynamical systems. *IEEE Transactions on Signal Processing*, 59(10):4781 –4794, 2011.

[44] H.L. Yap, M.B. Wakin, and C.J. Rozell. Stable manifold embeddings with operators satisfying the restricted isometry property. In *Proceedings of IEEE Conference on Information Sciences and Systems (CISS)*, 2011.

[45] Y. Zhang, J. Yang, and W. Yin. Yall1: Your algorithms for l1. yall1.blogs.rice.edu, 2011.

Appendix A. Toolbox

This section collects a few general results that are used throughout the paper (mainly without proofs, for the benefit of the space).

Schatten norms possess the following useful property that mirrors Euclidean norms:

$$\operatorname{Rank}(A)^{\frac{1}{p}-\frac{1}{q}} \|A\|_{S_q} \leq \|A\|_{S_p} \leq \|A\|_{S_q}, \tag{A.1}$$

for a matrix A, when $1 \leq q \leq p$. The following version of the Hölder inequality for matrices is used in this paper. For any pair of matrices A, B (such that AB exists), the following holds:[7]

$$\|AB\|_F \leq \|A\|_2 \|B\|_F. \tag{A.2}$$

For a random variable Z and $1 \leq p \leq q$, the following holds [32, Page 30]:

$$\mathbb{E}^p|Z| \leq \mathbb{E}^q|Z|. \tag{A.3}$$

Also, $\|C \cdot Z\|_{\psi_2} = |C|\|Z\|_{\psi_2}$ and, according to [38, Lemma 5.9], the following holds for a finite sequence of zero-mean independent random variables $\{Z_j\}$:

$$\|\sum_j Z_j\|_{\psi_2}^2 \lesssim \sum_j \|Z_j\|_{\psi_2}^2. \tag{A.4}$$

Throughout this section, let $g \in \mathbb{R}^N$ denote a vector whose entries are i.i.d. zero-mean unit variance sub-Gaussian random variables. For convenience, set $K := \max_{n \in [N]} \|g(n)\|_{\psi_2}$. For $t > 0$ and $n \in [N]$, the following holds by the definition of the sub-Gaussian norm [38]:

$$\log P\{|g(n)| > t\} \lesssim -\frac{t^2}{\|g(n)\|_{\psi_2}^2}. \tag{A.5}$$

Also, the following holds when $N \gtrsim 1$, [38, Lemma 6.6]:

$$\sqrt{\mathbb{E}\|g\|_\infty^2} = (\mathbb{E} \max_{n \in [N]} g^2(n))^{1/2} \lesssim K\sqrt{\log N}. \tag{A.6}$$

---

[7] Let $\{b_j\}$ denote the columns of B. Then $\|AB\|_F^2 = \sum_j \|Ab_j\|_2^2 \leq \sum_j \|A\|_2^2\|b_j\|_2^2 = \|A\|_2^2\|B\|_F^2$.

Furthermore, for $t \leq 1$, the following inequality is from [38, Corollary 5.17] and provides a lower bound on $\|g\|_2$:

$$\log P\Big\{ \|g\|_2 < (1-t)\sqrt{N} \Big\} \lesssim -\min(\frac{t^2}{K^4}, \frac{t}{K^2})N. \tag{A.7}$$

Suppose that the entries of $G \in \mathbb{R}^{N \times J}$ are Gaussian random variables with zero-mean and unit variance. The next inequality provides an upper bound on the spectral norm of G, which directly follows from Corollary 5.35 in [38]. When $J \leq N$, and for $t > 0$, the following holds:

$$\log P\Big\{ \|G\|_2 > (1+t)\sqrt{N} + \sqrt{J} \Big\} \leq -t^2 N/2. \tag{A.8}$$

The next result essentially bounds the moments of a sum of independent random variables with those of a Rademacher sequence. The proof of this result (and the next one) uses the symmetrization technique, which has the following argument at its heart. If Z is a random variable taking values in a Banach space χ, then we can define its symmetrized version $Y = Z - Z'$, where Z' is an independent copy of Z. As suggested by the name, the distribution of Y is symmetric about the origin. In addition, the distributions of Y and ξY are the same, where ξ is a standard Bernoulli random variable that takes ± 1 with equal probability.

Lemma 8. [32, Lemma 6.7] *Let $\{Z_l\}$ be a finite sequence of independent random variables in a Banach space $(\chi, \| \cdot \|_\chi)$. Then, the following holds:*

$$\mathbb{E}\| \sum_l Z_l - \mathbb{E}Z_l \|_\chi \leq 2\mathbb{E}\| \sum_l \xi_l Z_l \|_\chi,$$

where $\{\xi_l\}$ is a Rademacher sequence.

In our proofs, we also require a (weak version of) the Khintchine inequality for operator norms that we state next.

Lemma 9. *If $\{A_l\}$, $l \in [L]$, is a sequence of matrices of the same dimension and rank of at most $J \gtrsim 1$, then the following holds:*

$$\mathbb{E} \left\| \sum_{l \in [L]} \xi_l A_l \right\|_2 \lesssim \sqrt{\log J} \left(\sum_{l \in [L]} \|A_l\|_2^2 \right)^{1/2},$$

where $\{\xi_l\}$ is a Rademacher sequence.

Proof. From [32, Theorem 6.14] and for every $2 \leq p < \infty$, we recall that

$$\mathbb{E}^p \left\| \sum_{l \in [L]} \xi_l A_l \right\|_{S_p} \lesssim \sqrt{p} \cdot \max \left(\left\| (\sum_{l \in [L]} A_l A_l^*)^{\frac{1}{2}} \right\|_{S_p}, \left\| (\sum_{l \in [L]} A_l^* A_l)^{\frac{1}{2}} \right\|_{S_p} \right).$$

The spectral norm is a special case of the Schatten norm with $p = \infty$. Therefore, the inequality above does not directly apply to our problem. As such, we need a more detailed argument here, which follows the approach of [38]. From (A.1), with $p = \infty$ and $q = \log J$, recall that

$$e^{-1}\|A\|_{S_{\log J}} \leq \|A\|_2 \leq \|A\|_{S_{\log J}}$$

for $J \geq e$. This equivalence, in combination with the fact that $\mathbb{E}^p$ is increasing in p, i.e., (A.3), allows us to

23

write

$$\mathbb{E}\left\|\sum_{l\in[L]}\xi_l A_l\right\|_2 \leq \mathbb{E}\left\|\sum_{l\in[L]}\xi_l A_l\right\|_{S_{\log J}}$$

$$\leq \mathbb{E}^{\log J}\left\|\sum_{l\in[L]}\xi_l A_l\right\|_{S_{\log J}}$$

$$\lesssim \sqrt{\log J}\max\left(\left\|(\sum_{l\in[L]}A_l A_l^*)^{\frac{1}{2}}\right\|_{S_{\log J}}, \left\|(\sum_{l\in[L]}A_l^* A_l)^{\frac{1}{2}}\right\|_{S_{\log J}}\right)$$

$$\leq e\sqrt{\log J}\max\left(\left\|(\sum_{l\in[L]}A_l A_l^*)^{\frac{1}{2}}\right\|_2, \left\|(\sum_{l\in[L]}A_l^* A_l)^{\frac{1}{2}}\right\|_2\right)$$

$$= e\sqrt{\log J}\max\left(\left\|\sum_{l\in[L]}A_l A_l^*\right\|_2^{\frac{1}{2}}, \left\|\sum_{l\in[L]}A_l^* A_l\right\|_2^{1/2}\right)$$

$$\leq e\sqrt{\log J}\max\left((\sum_{l\in[L]}\|A_l\|_2^2)^{\frac{1}{2}}, (\sum_{l\in[L]}\|A_l\|_2^2)^{\frac{1}{2}}\right)$$

$$= e\sqrt{\log J}\cdot(\sum_{l\in[L]}\|A_l\|_2^2)^{\frac{1}{2}},$$

as claimed. We assumed above that $J \geq e^2$ to produce the first line, and the second to the last line above uses the triangle inequality and the fact that $\|AA^*\|_2 = \|A\|_2^2$ for any matrix A. We remark that had we stopped at the fifth line above, we would have ended with the stronger original non-commutative Khintchine inequality for the spectral norm. However, the weaker bound given in the last line suffices for our purposes in this paper and completes the proof of Lemma 9.

$\square$

We also list two trivial identities regarding covering numbers, which hold for every set $\mathcal{S}$, norm $\|\cdot\|$, and $r, a > 0$:

$$\mathcal{N}(\mathcal{S}, a\|\cdot\|, r) = \mathcal{N}(\mathcal{S}, \|\cdot\|, r/a)$$
$$\mathcal{N}(a\mathcal{S}, \|\cdot\|, r) = \mathcal{N}(\mathcal{S}, \|\cdot\|, r/a). \tag{A.9}$$

For reference, we also provide an estimation for two integrals we encounter in the analysis:

$$\int_0^a \log\left(1 + \frac{b}{\nu}\right) d\nu \lesssim a\log\left(1 + \frac{b}{a}\right). \tag{A.10}$$

$$\int_0^a \sqrt{\log\left(1 + \frac{b}{\nu}\right)} d\nu \lesssim a\sqrt{\log\left(1 + \frac{b}{a}\right)}, \tag{A.11}$$

Both inequalities hold when $a \leq b$.

Appendix B. Proof of Lemma 1

Equivalently, R can be constructed as follows. Let $\{r_{\widetilde{n}}\}$, $\widetilde{n} \in [\widetilde{N}]$, denote the columns of R. The first column, r_1, is chosen from the uniform distribution on the unit sphere in $\mathbb{R}^{\widetilde{N}}$. For every $\widetilde{n} \in [\widetilde{N}]\backslash\{1\}$, $r_{\widetilde{n}}$

is chosen from the uniform distribution on the unit sphere in the orthogonal complement of the span of the first $\widetilde{n} - 1$ columns.

Let $g \in \mathbb{R}^{\widetilde{N}}$ denote a standard Gaussian vector, that is a vector whose entries are i.i.d. zero-mean Gaussian random variables with unit variance. Since r_1 is drawn from the uniform distribution on the unit sphere in $\mathbb{R}^{\widetilde{N}}$, the entries of r_1 have the same distribution as those in $g/\|g\|_2$. Since the distribution of R remains unchanged under permutation of its rows, every column of R has the same (marginal) distribution as $g/\|g\|_2$. This, alongside with the union bound, allows us to write the following for any $t > 0$:

$$
\begin{aligned}
\mathrm{P}\Big\{\mu(R) > t\sqrt{\log\widetilde{N}}\Big\} &= \mathrm{P}\Big\{\max_{\widetilde{n}\in[\widetilde{N}]}\|r_{\widetilde{n}}\|_{\max} > t\sqrt{\log\widetilde{N}}/\sqrt{\widetilde{N}}\Big\} \\
&\leq \widetilde{N}\cdot\max_{\widetilde{n}\in[\widetilde{N}]}\mathrm{P}\Big\{\|r_{\widetilde{n}}\|_{\max} > t\sqrt{\log\widetilde{N}}/\sqrt{\widetilde{N}}\Big\} \\
&= \widetilde{N}\cdot\mathrm{P}\Big\{\max_{\widetilde{n}\in[\widetilde{N}]}\frac{|g(\widetilde{n})|}{\|g\|_2} > t\sqrt{\frac{\log\widetilde{N}}{\widetilde{N}}}\Big\} \\
&\leq \widetilde{N}^2\cdot\mathrm{P}\Big\{\frac{|g(1)|}{\|g\|_2} > t\sqrt{\frac{\log\widetilde{N}}{\widetilde{N}}}\Big\} \\
&= \widetilde{N}^2\cdot\mathrm{P}\Big\{\frac{|g(1)|}{\|g\|_2} > \frac{t/2\cdot\sqrt{\log\widetilde{N}}}{(1-1/2)\sqrt{\widetilde{N}}}\Big\} \\
&\leq \widetilde{N}^2\cdot\mathrm{P}\Big\{|g(1)| > t/2\cdot\sqrt{\log\widetilde{N}}\Big\} + \widetilde{N}^2\cdot\mathrm{P}\Big\{\|g\|_2 < (1-1/2)\sqrt{\widetilde{N}}\Big\} \\
&\lesssim \widetilde{N}^2 e^{-\frac{C_2}{4\|g(1)\|_{\psi_2}^2}t^2\log\widetilde{N}} + \widetilde{N}^2 e^{-C_3\min\left(\frac{1}{4\|g(1)\|_{\psi_2}^4},\frac{1}{2\|g(1)\|_{\psi_2}^2}\right)\widetilde{N}} \\
&\leq \widetilde{N}^2 e^{-\frac{C_2}{4\|g(1)\|_{\psi_2}^2}t^2\log\widetilde{N}} + \widetilde{N}^2 e^{-\frac{C_3}{4\|g(1)\|_{\psi_2}^4}\widetilde{N}},
\end{aligned}
$$

where we used (A.5) and (A.7) to bound the failure probability. The last line holds because $\|g(1)\|_{\psi_2} = \sqrt{2/\pi}$.[8] If we take $\widetilde{N} \geq t^2\log\widetilde{N}$, we obtain that

$$
\mathrm{P}\Big\{\mu(R) > t\sqrt{\log\widetilde{N}}\Big\} \lesssim \widetilde{N}^{2-\frac{C_2}{4\|g(1)\|_{\psi_2}^2}t^2} + \widetilde{N}^{2-\frac{C_3}{4\|g(1)\|_{\psi_2}^4}t^2}. \tag{B.1}
$$

We arrive at the advocated result when $t \gtrsim 1$:

$$
\mathrm{P}\Big\{\mu(R) > t\sqrt{\log\widetilde{N}}\Big\} \lesssim \widetilde{N}^{-t} + \widetilde{N}^{-t} = 2\widetilde{N}^{-t}. \tag{B.2}
$$

Appendix C. Proof of Lemma 2

We use here the construction of R laid down in the beginning of Appendix B. As pointed out in the proof of Lemma 1, the columns of R are dependent but identically distributed as $g/\|g\|_2$, where g was defined

---

[8] This is easily verified using the moments of the Gaussian distribution.

there. Now, for every t, we can write

$$P\left\{\gamma(R) \gtrsim 1 + \sqrt{\frac{J}{N}} + t\right\} = P\left\{\max_{\widetilde{n} \in [\widetilde{N}]} \|X(R, e_{\widetilde{n}})\|_2 \gtrsim \frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}}\right\}$$

$$\leq \widetilde{N} \cdot \max_{\widetilde{n} \in [\widetilde{N}]} P\left\{\|X(R, e_{\widetilde{n}})\|_2 \gtrsim \frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}}\right\}$$

$$= \widetilde{N} \cdot P\left\{\|X(R, e_1)\|_2 \gtrsim \frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}}\right\}. \tag{C.1}$$

The second line uses the union bound and the last line holds due to the identical distribution of the columns of R. It remains to find an upper bound for the probability in the last line above. Recall that r_1 has the same distribution as $g/\|g\|_2$, and thus $X(R, e_1)$ has the same distribution as $G/\|G\|_F$, where $G \in \mathbb{C}^{N \times J}$ is formed by reshaping g. Therefore, $\|X(R, e_1)\|_2$ has the same distribution as $\|G\|_2 / \|G\|_F$. For fixed $t \leq 1$, the following convenient inequality holds:[9]

$$\frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}} \gtrsim \frac{(1+t/3)\sqrt{N} + \sqrt{J}}{(1-t/3)\sqrt{\widetilde{N}}}. \tag{C.2}$$

Now, we can write that

$$P\left\{\|X(R, e_1)\|_2 \gtrsim \frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}}\right\}$$

$$= P\left\{\frac{\|G\|_2}{\|G\|_F} \gtrsim \frac{1}{\sqrt{N}} + \frac{1+t}{\sqrt{J}}\right\}$$

$$\leq P\left\{\frac{\|G\|_2}{\|G\|_F} \gtrsim \frac{\left(1 + \frac{t}{3}\right)\sqrt{N} + \sqrt{J}}{\left(1 - \frac{t}{3}\right)\sqrt{\widetilde{N}}}\right\}$$

$$\leq P\left\{\|G\|_2 \gtrsim \left(1 + \frac{t}{3}\right)\sqrt{N} + \sqrt{J}\right\} + P\left\{\|G\|_F \lesssim \left(1 - \frac{t}{3}\right)\sqrt{\widetilde{N}}\right\}$$

$$\lesssim e^{-\frac{1}{18}t^2 N} + e^{-C_3 \min\left(\frac{t^2}{9\|g(1)\|_{\psi_2}^4}, \frac{t}{3\|g(1)\|_{\psi_2}^2}\right)\widetilde{N}}$$

$$\leq e^{-\frac{1}{18}t^2 N} + e^{-\frac{C_3}{9\|g(1)\|_{\psi_2}^4}t^2\widetilde{N}}$$

$$\leq 2e^{-C_4 t^2 N},$$

where the second line uses (C.2). The fourth line uses the inequalities (A.7) and (A.8) and the fact that $\|G\|_F = \|g\|_2$. The second to last line follows because $\|g(1)\|_{\psi_2} = \sqrt{2/\pi}$ and $t \leq 1$. The above upper bound in combination with (C.1) leads us to the following conclusion:

$$P\left\{\gamma(R) \gtrsim 1 + \sqrt{\frac{J}{N}} + t\right\} \lesssim \widetilde{N} e^{-C_4 t^2 N}.$$

We complete the proof of Lemma 2 by taking $N \geq 2C_4^{-1} t^{-2} \log \widetilde{N}$.

Appendix D. Proof of Lemma 3

Let $T' := F_J \otimes I_P \in \mathbb{C}^{PJ \times PJ}$, where $\otimes$ stands for the Kronecker product. Here, F_J and I_P are the Fourier and canonical orthobases of size, respectively, $J \times J$ and $P \times P$. We consider the natural partitioning

---

[9] This inequality easily follows from the fact that $1 + t \geq (1 + t/3)(1 - t/3)^{-1}$ holds for every $t \leq 1$.

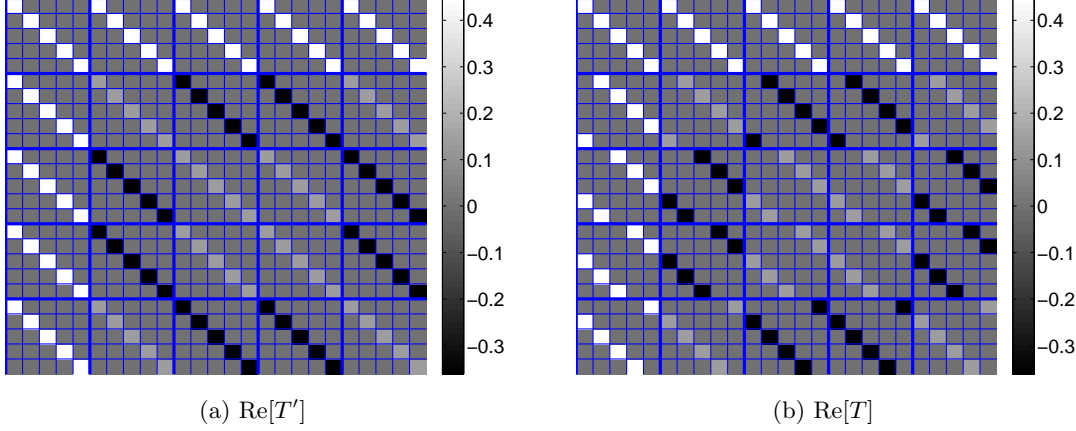(a) Re[T']   (b) Re[T]

Figure D.3: A visual illustration of the transformation from T' to T for $P = J = 5$. See the explanation in Appendix D.

of T' into J row-submatrices of size $P \times PJ$, denoted by T'_j, $j \in [J]$. Now, for every $j \in [J]$, cyclically shift the rows of T'_j by $j - 1$ times upward and create $T_j \in \mathbb{C}^{P \times PJ}$. Then form the matrix $T \in \mathbb{C}^{PJ \times PJ}$ by replacing every T'_j with T_j. The transformation of T' to T is visualized in Figure D.3.

Using the properties of the Kronecker product, it is easily verified that T' is an orthobasis for $\mathbb{C}^{PJ}$. Due to its structure, we also observe that the nonzero entries of the p_1th and p_2th columns of T do not overlap when $(p_1 - p_2) \bmod P \neq 0$, and so these columns are orthogonal. Otherwise, when $(p_1 - p_2) \bmod P = 0$, we need a more subtle argument. Under this condition, the inner product of the p_1th and p_2th columns of T' equals the inner product of the $\lceil p_1/P \rceil$th and $\lceil p_2/P \rceil$th columns of F_J and is indeed zero. The inner product of the p_1th and p_2th columns remains zero under the transformation of T' to T since this transformation only amounts to a permutation in the rows of T'. Therefore, T is an orthobasis for $\mathbb{C}^{PJ}$. As for computing $\gamma(T)$, the structure of T guarantees that, for each j, every column and row of $X(e_j, T)$ has only one nonzero entry with the magnitude of $1/\sqrt{J}$. Therefore, $\gamma(T) = 1$.

Finally, it can be easily verified that $\hat{x}/\sqrt{J} = Tx_e$, where $x_e = [x^T, 0, 0, \cdots, 0]^T \in \mathbb{C}^{PJ}$, and by (17), $\Gamma x = \Xi T x_e$. If x is S-sparse, so is x_e. This completes the proof of Lemma 3.

Appendix E. Proof of Lemma 6

The arguments used in this section are largely adapted from [35]. In what follows, we let $\mathcal{B}_A^{\widetilde{N}}$ denote the unit ball with respect to the norm $\|\cdot\|_A$ in $\mathbb{C}^{\widetilde{N}}$. Also, $\mathcal{B}_1^{\widetilde{N}}$ and $\mathcal{B}_2^{\widetilde{N}}$, respectively, denote the unit ℓ_1-ball and ℓ_2-ball in $\mathbb{C}^{\widetilde{N}}$. For $T \subseteq [\widetilde{N}]$, we let $\mathcal{B}_A^T$ denote the unit $\|\cdot\|_A$ ball in the $\#T$-dimensional subspace of $\mathbb{C}^{\widetilde{N}}$ spanned by $\{e_{\widetilde{n}}\}$, $\widetilde{n} \in T$. We define $\mathcal{B}_1^T$ and $\mathcal{B}_2^T$ similarly.

The first thing to notice is that when $\alpha \in \Omega_S/\sqrt{S}$, then $\|\alpha\|_1 \leq 1$. From the hypothesis of the lemma, we then have that

$$\|\alpha\|_A \leq \frac{\kappa}{\sqrt{\widetilde{M}}}. \tag{E.1}$$

This implies that for every support $T \subset [\widetilde{N}]$ with $\#T = S$, we have

$$\frac{\mathcal{B}_2^T}{\sqrt{S}} \subseteq \frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \mathcal{B}_A^T. \tag{E.2}$$

On the other hand, $\Omega_S/\sqrt{S}$ can be equivalently represented as

$$\frac{\Omega_S}{\sqrt{S}} = \bigcup_{\#T=S} \frac{\mathcal{B}_2^T}{\sqrt{S}}. \tag{E.3}$$

27

Together, (E.2) and (E.3) imply that

$$\frac{\Omega_S}{\sqrt{S}} \subseteq \bigcup_{\#T=S} \frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \mathcal{B}_A^T. \tag{E.4}$$

We also record that

$$\frac{\Omega_S}{\sqrt{S}} \subseteq \frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \mathcal{B}_A^{\widetilde{N}}, \tag{E.5}$$

which dictates that

$$\mathcal{N}(\Omega_S/\sqrt{S}, \|\cdot\|_A, \nu) = 1 \tag{E.6}$$

if $\nu \geq \kappa/\sqrt{\widetilde{M}}$. This proves the second statement in Lemma 6. Otherwise, if $\nu < \kappa/\sqrt{\widetilde{M}}$, we continue with the rest of the argument. In light of (E.4), we have that

$$\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) \leq \mathcal{N}\left(\bigcup_{\#T=S} \frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \mathcal{B}_A^T, \|\cdot\|_A, \nu\right)$$

$$\leq \binom{\widetilde{N}}{S} \cdot \mathcal{N}\left(\frac{\kappa}{\sqrt{\widetilde{M}}} \cdot \mathcal{B}_A^T, \|\cdot\|_A, \nu\right)$$

$$= \binom{\widetilde{N}}{S} \cdot \mathcal{N}\left(\mathcal{B}_A^T, \|\cdot\|_A, \nu\kappa^{-1}\sqrt{\widetilde{M}}\right), \tag{E.7}$$

where the last line uses the second inequality in (A.9). An estimate for the covering number in the last line above is available for $r > 0$, namely[10]

$$\mathcal{N}(\mathcal{B}_A^T, \|\cdot\|_A, r) \leq (1 + 2/r)^{2\#T}. \tag{E.8}$$

Plugging the bound above into (E.7), we arrive at

$$\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) \leq \binom{\widetilde{N}}{S} \cdot \left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right)^{2S} \leq \left(\frac{e\widetilde{N}}{S}\right)^S \cdot \left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right)^{2S}. \tag{E.9}$$

The last inequality holds because $\binom{n}{m} \leq (en/m)^m$ for any pair of integers $m \leq n$. When $\widetilde{N} \gtrsim 1$, (E.9) implies that

$$\log \mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_X, \nu\right) \leq S \log\left(\frac{e\widetilde{N}}{S}\right) + 2S \log\left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right)$$

$$\lesssim S \log \widetilde{N} + S \log\left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right). \tag{E.10}$$

The bound above is less effective for small values of ν. To seek a second bound on the covering number, we begin from the containment

$$\frac{\Omega_S}{\sqrt{S}} \subseteq \mathcal{B}_1^{\widetilde{N}}, \tag{E.11}$$

which immediately dictates that

$$\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) \leq \mathcal{N}\left(\mathcal{B}_1^{\widetilde{N}}, \|\cdot\|_A, \nu\right). \tag{E.12}$$

In order to compute the covering number on the right hand side of (E.12), we use the following result, which is proved in Appendix F.

---

[10] This is proved similar to Lemma 5.2 in [38], after exchanging the Euclidean metric with $\|\cdot\|_A$ and accounting for the complex vector space.

Lemma 10. *Let $B_{1,\widetilde{N}}$ denote the ℓ_1-ball in $\mathbb{R}^{\widetilde{N}}$, and consider the norm $\|\cdot\|_A$ on $\mathbb{C}^{\widetilde{N}}$, which satisfies the hypothesis of Lemma 6. Naturally, $\|\cdot\|_A$ induces a norm on $\mathbb{R}^{\widetilde{N}} \subset \mathbb{C}^{\widetilde{N}}$ (which is represented with the same notation for convenience). For $\nu > 0$ and $\widetilde{M} \gtrsim 1$, it holds that*

$$\log\left(\mathcal{N}\left(B_{1,\widetilde{N}}, \|\cdot\|_A, \nu\right)\right) \lesssim \frac{\kappa^2}{\nu^2 \widetilde{M}} \cdot \log^2 \widetilde{N}.$$

Now consider an arbitrary $\beta \in \mathcal{B}_1^{\widetilde{N}}$, and note that $\mathrm{Re}[\beta], \mathrm{Im}[\beta] \in B_{1,\widetilde{N}}$. Let

$$\mathcal{C}_1 := \mathcal{C}\left(B_{1,\widetilde{N}}, \|\cdot\|_A, \nu/2\right)$$

denote a minimal $(\nu/2)$-cover for $B_{1,\widetilde{N}}$ with respect to the metric $\|\cdot\|_A$. Therefore, there exist $p_1, p_2 \in \mathcal{C}_1$ such that $\|\mathrm{Re}[\beta] - p_1\|_X, \|\mathrm{Im}[\beta] - p_2\|_X \le \nu/2$. It follows by the triangle inequality that

$$\begin{aligned}
\|\beta - (p_1 + \mathrm{i}p_2)\|_A &= \|(\mathrm{Re}[\beta] - p_1) + \mathrm{i}(\mathrm{Im}[\beta] - p_2)\|_A \\
&\le \|\mathrm{Re}[\beta] - p_1\|_A + \|\mathrm{Im}[\beta] - p_2\|_A \\
&\le \nu.
\end{aligned}$$

Therefore, $\{p_1 + \mathrm{i}p_2 : p_1, p_2 \in \mathcal{C}_1\}$ is a cover for $\mathcal{B}_1^{\widetilde{N}}$, and clearly

$$\mathcal{N}(\mathcal{B}_1^{\widetilde{N}}, \|\cdot\|_A, \nu) \le \left(\mathcal{N}(B_{1,\widetilde{N}}, \|\cdot\|_A, \nu/2)\right)^2.$$

It now follows from (E.12), Lemma 10, and the inequality above that

$$\log\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) \lesssim \frac{\kappa^2}{\nu^2 \widetilde{M}} \cdot \log^2 \widetilde{N}. \tag{E.13}$$

Combining (E.10) and (E.13), we finally arrive at

$$\log\mathcal{N}\left(\frac{\Omega_S}{\sqrt{S}}, \|\cdot\|_A, \nu\right) \lesssim \min\left(S\log\widetilde{N} + S\log\left(1 + \frac{2\kappa}{\nu\sqrt{\widetilde{M}}}\right), \frac{\kappa^2}{\nu^2\widetilde{M}} \cdot \log^2 \widetilde{N}\right),$$

which holds when $\widetilde{M} \gtrsim 1$. This completes the proof of Lemma 6.

Appendix F. Proof of Lemma 10

Consider an arbitrary $\beta \in B_{1,\widetilde{N}}$. Also consider a random vector Z that takes a value in $\{0\} \cup \{\mathrm{sgn}\,(\beta(\widetilde{n})) \cdot e_{\widetilde{n}}\}$, $\widetilde{n} \in [\widetilde{N}]$. It takes $\mathrm{sgn}\,(\beta(\widetilde{n})) \cdot e_{\widetilde{n}}$ with probability of $|\beta(\widetilde{n})|$ and zero otherwise. Clearly, $\mathbb{E}Z = \beta$. Now, we wish to approximate β with the average of L independent copies of Z, denoted by $\{Z_l\}$, $l \in [L]$. The expected approximation error, measured in $\|\cdot\|_A$, would be

$$\mathbb{E}\left\|\beta - \frac{1}{L}\sum_{l\in[L]} Z_l\right\|_A.$$

Since the argument of the norm is zero-mean, we can use the symmetrization technique by invoking Lemma 8 from the Toolbox to obtain that

$$\mathbb{E}\left\|\beta - \frac{1}{L}\sum_{l\in[L]} Z_l\right\|_A = \frac{1}{L}\mathbb{E}\left\|\sum_{l\in[L]} Z_l - \mathbb{E}Z_l\right\|_A \le \frac{2}{L}\mathbb{E}\left\|\sum_{l\in[L]} \xi_l Z_l\right\|_A, \tag{F.1}$$

29

where $\{\xi_l\}$, $l \in [L]$, is a Rademacher sequence. According to Lemma 6, $\|\cdot\|_A = \|A(\cdot)\|_2$ and $A(\cdot)$ is a linear map. Therefore

$$\mathbb{E}\left\|\sum_{l\in[L]}\xi_l Z_l\right\|_A = \mathbb{E}\left\|\sum_{l\in[L]}\xi_l A(Z_l)\right\|_2.$$

Also, from the hypothesis of Lemma 6, $\mathrm{Rank}\,(A(Z_l)) \le \widetilde{M}$ for every l. We now invoke a Khintchine-type inequality for the operator norm (stated in Lemma 9 from the Toolbox), which allows us to continue our argument as

$$\mathbb{E}\left\|\sum_{l\in[L]}\xi_l A(Z_l)\right\|_2 = \mathbb{E}^Z\mathbb{E}^\xi\left\|\sum_{l\in[L]}\xi_l A(Z_l)\right\|_2$$

$$\lesssim \sqrt{\log\widetilde{M}}\cdot\mathbb{E}^Z\left(\sum_{l\in[L]}\|A(Z_l)\|_2^2\right)^{1/2}$$

$$\le \sqrt{\log\widetilde{M}}\cdot\left(\sum_{l\in[L]}\mathbb{E}\|A(Z_l)\|_2^2\right)^{1/2}$$

$$= \sqrt{\log\widetilde{M}}\cdot\left(\sum_{l\in[L]}\sum_{\widetilde{n}\in[\widetilde{N}]}|\beta_{\widetilde{n}}|\cdot\|A(e_{\widetilde{n}})\|_2^2\right)^{1/2}$$

$$\le \sqrt{\log\widetilde{M}}\cdot\left(L\cdot\max_{\widetilde{n}\in\widetilde{N}}\|A(e_{\widetilde{n}})\|^2\right)^{1/2}$$

$$= \sqrt{L\log\widetilde{M}}\cdot\max_{\widetilde{n}\in\widetilde{N}}\|e_{\widetilde{n}}\|_A$$

$$\le \kappa\sqrt{\frac{L}{\widetilde{M}}\log\widetilde{M}}, \tag{F.2}$$

where, conditioned on $\{Z_l\}$, $l \in [L]$, the Khintchine inequality was used to produce the second line. The third line follows from the Jensen inequality. The fifth line uses the assumption that $\beta \in B_{1,\widetilde{N}}$. The last line follows from the hypothesis of Lemma 6. Using the above inequality in combination with (F.1) yields

$$\mathbb{E}\left\|\beta - \frac{1}{L}\sum_{l\in[L]}Z_l\right\|_X \lesssim \kappa\sqrt{\frac{\log\widetilde{M}}{L\widetilde{M}}}.$$

To keep the average no larger than ν, it suffices to take

$$L \gtrsim \frac{\kappa^2}{\nu^2\widetilde{M}}\cdot\log\widetilde{M}.$$

With this choice of L, there exists a linear combination of independent copies of Z that falls within a ν distance of β. In other words, we have shown that for an arbitrary $\beta \in B_{1,\widetilde{N}}$, there exists an average of L elements of $\{0\}\cup\{\pm e_{\widetilde{n}}\}$ that is of distance at most ν from β. There are $2\widetilde{N}+1$ elements in the aforementioned set, and so there are $(2\widetilde{N}+1)^L$ possibilities for the average. We therefore conclude that

$$\mathcal{N}\left(B_{1,\widetilde{N}}, \|\cdot\|_A, \nu\right) \le \left(2\widetilde{N}+1\right)^{C_5\log\widetilde{M}\cdot\kappa^2/\nu^2\widetilde{M}},$$

or

$$\log \mathcal{N}\left(B_{1,\widetilde{N}}, \|\cdot\|_A, \nu\right) \lesssim \frac{\kappa^2}{\nu^2 \widetilde{M}} \cdot \log \widetilde{M} \log \left(2\widetilde{N} + 1\right)$$
$$\lesssim \frac{\kappa^2}{\nu^2 \widetilde{M}} \cdot \log \widetilde{M} \log \widetilde{N}$$
$$\leq \frac{\kappa^2}{\nu^2 \widetilde{M}} \cdot \log^2 \widetilde{N}$$

when $\widetilde{N} \gtrsim 1$. This completes the proof of Lemma 10.