Conditioning of Random Block Subdictionaries with Applications to Block-Sparse Recovery and Regression

Waheed U. Bajwa, *Senior Member, IEEE,* Marco F. Duarte, *Senior Member, IEEE,* and Robert Calderbank, *Fellow, IEEE*

***Abstract*—The linear model, in which a set of observations is assumed to be given by a linear combination of columns of a matrix (often termed a dictionary), has long been the mainstay of the statistics and signal processing literature. One particular challenge for inference under linear models is understanding the conditions on the dictionary under which reliable inference is possible. This challenge has attracted renewed attention in recent years since many modern inference problems (e.g., high-dimensional statistics, compressed sensing) deal with the "underdetermined" setting, in which the number of observations is much smaller than the number of columns in the dictionary. This paper makes several contributions for this setting when the set of observations is given by a linear combination of a small number of *groups of columns* of the dictionary, termed the "block-sparse" case. First, it specifies conditions on the dictionary under which most block submatrices of the dictionary (often termed block subdictionaries) are well conditioned. This result is fundamentally different from prior work on block-sparse inference because (i) it provides conditions that can be explicitly computed in polynomial time, (ii) the given conditions translate into near-optimal scaling of the number of columns of the block subdictionaries as a function of the number of observations for a large class of dictionaries, and (iii) it suggests that the spectral norm, rather than the column/block coherences, of the dictionary fundamentally limits the scaling of dimensions of the well-conditioned block subdictionaries. Second, in order to help understand the significance of this result in the context of block-sparse inference, this paper investigates the problems of block-sparse recovery and block-sparse regression in underdetermined settings. In both of these problems, this paper utilizes its result concerning conditioning of block subdictionaries and establishes that near-optimal block-sparse recovery and block-sparse regression is possible for a large class of dictionaries as long as the dictionary satisfies easily computable conditions and the coefficients describing the linear combination of groups of columns can be modeled through a mild statistical prior. Third, the paper reports extensive numerical experiments that highlight the effects of different measures of the dictionary in block-sparse inference problems.**

***Index Terms*—Block-sparse inference, block-sparse recovery, block-sparse regression, compressed sensing, group lasso, group sparsity, high-dimensional statistics, multiple measurement vectors, random block subdictionaries**

I. Introduction

CONSIDER the classical linear forward model $y = \Phi\beta$, which relates a parameter vector $\beta \in \mathbb{R}^p$ to an observation vector $y \in \mathbb{R}^n$ through a linear transformation (henceforth referred to as a *dictionary*) $\Phi \in \mathbb{R}^{n\times p}$. This forward model, despite its apparent simplicity, provides a reasonable mathematical approximation of reality in a surprisingly large number of application areas and scientific disciplines [5–7]. While the operational significance of this linear (forward) model varies from one application to another, the fundamental purpose of it in all applications stays the same: *given knowledge of y and Φ, make an inference about β*. However, before one attempts to solve an inference problem using the linear model, it is important to understand the conditions under which doing so is even feasible. For instance, inferring anything about β will be a moot point if the nullspace of Φ were to contain β. Thus, a large part of the literature on linear models is devoted to characterizing conditions on Φ and β that facilitate reliable inference.

Classical literature on inference using linear models proceeds under the assumption that the number of observations n equals or exceeds the number of parameters p. In this setting, conditions such as Φ being full column rank or $\Phi\Phi^*$ being well conditioned—both of which can be explicitly verified—are common in the inference literature [5, 8, 9]. In contrast, there has recently been a growing interest to study inference under linear models when n is much smaller than p. This setting is the hallmark of high-dimensional statistics [10], arises frequently in many application areas [11], and forms the cornerstone of the philosophy behind compressed sensing [12, 13]. It of course follows from simple linear algebra that inferring about every possible β from $y = \Phi\beta$ is impossible in this setting; instead, the high-dimensional inference literature commonly operates under the assumption that β has only a few nonzero parameters—typically on the order of n—and characterizes corresponding conditions on Φ for reliable inference. Some notable conditions in this regard include the spark [14], the restricted isometry property [15], the irrepresentable condition [16] (and its variant, the incoherence condition [17]), the restricted eigenvalue assumption [18], and the nullspace property [19]. While these and other conditions

The first two authors contributed equally to the paper and are listed in alphabetical order. Portions of this work have previously appeared in technical reports [1, 2], and at the Int. Conf. Sampling Theory and Applications (SAMPTA), 2011 [3] and Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2014 [4]. This work was supported by the NSF under grant CCF-1218942, by the ARO under grant W911NF-14-1-0295, by the ONR under grant N00014-08-1-1110 and by the AFOSR under grants FA9550-09-1-0422, FA9550-09-1-0643, and FA9550-05-0443. MFD was also supported by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch.

W. U. Bajwa is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854. E-mail: waheed.bajwa@rutgers.edu

M. F. Duarte is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003. E-mail: mduarte@ecs.umass.edu

R. Calderbank is with the Departments of Computer Science, Electrical and Computer Engineering, and Mathematics at Duke University, Durham, NC 27708. E-mail: robert.calderbank@duke.edu

Copyright © 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

in the literature differ from each other in one way or the other, they all share one simple fact: *requiring that Φ satisfies one of these conditions implies that one or more column submatrices (subdictionaries) of Φ must be full column rank and/or well conditioned.* Unfortunately, explicitly verifying that Φ satisfies one of these properties is computationally daunting (NP-hard in some cases [20]), while indirect means of verifying these conditions provide rather pessimistic bounds on the dimensions of subdictionaries of Φ that are well conditioned [21].

In a recent series of influential papers, several researchers have managed to circumvent the pessimistic bounds associated with verifiable conditions on Φ for high-dimensional inference by resorting to an *average-case analysis* [22–27]. Representative work by Tropp [24, 25], for instance, shows that *most* subdictionaries of Φ having unit ℓ_2-norm columns are guaranteed to be well conditioned when the number of columns in the subdictionary is proportional to $p/(\|\Phi\|_2^2 \log p)$—provided that correlations between the columns of Φ do not exceed a certain threshold, a condition readily verifiable in polynomial time. In particular, these results imply that if Φ is a unit norm tight frame [21], corresponding to $\|\Phi\|_2^2 = p/n$, then it can be *explicitly verified* that most subdictionaries of Φ of dimension $n \times O(n/\log p)$ are well conditioned.[1] The biggest advantage of such *average-case analysis* results for the conditioning of subdictionaries of Φ lies in their ability to facilitate tighter verifiable conditions for inference under the linear model using an arbitrary (random or deterministic) dictionary Φ. Several works in this regard have been able to leverage the results of [24, 25] to provide tighter verifiable conditions for average-case sparse recovery [28, 29] (i.e., obtaining β from $y = \Phi\beta$ *à la* compressed sensing [12, 13]), average-case model selection [30] (i.e., estimating locations of the nonzero entries of β from $y = \Phi\beta +$ noise), and average-case linear regression [30] (i.e., estimating $\Phi\beta$ from $y = \Phi\beta +$ noise).

A. Our Contributions

Our focus in this paper is on inference under the linear model in the "n smaller than p" setting, in the case when β not only has a few nonzero parameters, but also its nonzero parameters exhibit a certain *block* (or *group*) structure. Specifically, we have $\beta = [\beta_1^* \ \beta_2^* \ \dots \ \beta_r^*]^*$ with $\beta_i \in \mathbb{R}^m$ for $m, r \in \mathbb{Z}_+$, $p = rm$, and only $k \ll r$ of the β_i's are nonzero (sub)vectors. Such setups are often referred to as *block sparse* (or *group sparse*) and arise in various contexts in a number of inference problems [31–35]. The most fundamental challenge for inference in this block-sparse setting then becomes specifying conditions under which one or more *block subdictionaries* of Φ are full column rank and/or well conditioned. A number of researchers have made substantial progress in this regard recently, reporting conditions on Φ in the block setting that mirror many of the ones reported in [14–19] for the classical setup; see, e.g., [31, 32, 36–61]. However, just like in the classical setup, verifying that Φ satisfies one of these properties in the block setting ends up being either computationally intractable or results in rather pessimistic bounds on the dimensions of block subdictionaries of Φ that are well conditioned. In contrast to these works, and in much the same way [22–27] reasoned in the classical case, we are interested in overcoming the pessimistic bounds associated with verifiable conditions on Φ for high-dimensional inference in the block-sparse setting by resorting to an average-case analysis.

Our first main contribution in this regard is a generalization of [24, 25] that establishes that *most block subdictionaries* of Φ having unit ℓ_2-norm columns are guaranteed to be well conditioned with the number of *blocks* in the subdictionary proportional to $r/(\|\Phi\|_2^2 \log p)$ provided that Φ satisfies a polynomial-time verifiable condition that we term the *block incoherence condition*. In particular, these results also imply that if Φ is a unit norm tight frame then it can be explicitly verified that most *block subdictionaries* of Φ of dimension $n \times O(n/\log p)$ are well conditioned.

While our ability to guarantee that most block subdictionaries of a dictionary that satisfies the block incoherence condition are well conditioned makes us optimistic about the use of such dictionaries in inference problems, there remains an analytical gap in going from conditioning of block subdictionaries to performance of inference tasks. Our second main contribution in this regard is applications of the result concerning the conditioning of block subdictionaries to provide tighter verifiable conditions for average-case block-sparse recovery (i.e., obtaining β from $y = \Phi\beta$ with β being block sparse) and average-case block-sparse regression (i.e., estimating $\Phi\beta$ from $y = \Phi\beta +$ noise with β being block sparse).

Last, but not least, we carry out a series of numerical experiments to highlight an aspect of inference under the linear model that is rarely discussed in the related literature: *the spectral norm of the dictionary $\|\Phi\|_2$ influences the inference performance much more than any of its other measures*. Specifically, our results show that performances of block-sparse recovery and regression are inversely proportional to $\|\Phi\|_2^2$ and tend to be independent of correlations between the columns of Φ for the most part—an outcome that also hints at the possible (orderwise) tightness of our results concerning the conditioning of block subdictionaries.

B. Notational Convention and Organization

The following notation will be used throughout the rest of this paper. We use uppercase and lowercase Roman/Greek letters for matrices and vectors/scalars, respectively. Given a vector v, we use $\|v\|_q$ and v^* to denote the usual ℓ_q norm and conjugate transpose of v, respectively. We define the *scalar* sign operator for $x \in \mathbb{R}$ as $\mathrm{sign}(x) := x/|x|$, while we use $\mathrm{sign}(v)$ for a vector v to denote entry-wise sign operation. In addition, we define the *vector* sign operator for a vector v as $\overline{\mathrm{sign}}(v) := v/\|v\|_2$, which returns the unit-norm vector pointing in the direction of v. Given two vectors u and v, we define the inner product between them as $\langle u, v\rangle := \sum_i u_i v_i$. Given a matrix A, we use $\|A\|_2$ and A^* to denote the spectral norm ($\sigma_{\max}(A)$) and the adjoint operator of A, respectively.

[1] Recall Landau's notation: $f(x) = O\big(g(x)\big)$ if there exist some c_0, x_0 such that $f(x) \le c_0 g(x)$ for all $x \ge x_0$.

In addition, assuming A has unit ℓ_2-norm columns and using A_i to denote the i^{th} column of A, the *coherence* of A is defined as $\mu(A) := \max_{i,j:i\neq j} |\langle A_i, A_j\rangle|$. Given a set $\mathcal{S}$, we use $A_{\mathcal{S}}$ (resp. $v_{\mathcal{S}}$) to denote the submatrix (resp. subvector) obtained by retaining the columns of A (resp. entries of v) corresponding to the indices in $\mathcal{S}$. Given a random variable R, we use $\mathbb{E}_q[R]$ to denote $\left(\mathbb{E}[R^q]\right)^{1/q}$. Finally, Id denotes the identity operator and $\otimes$ denotes a Kronecker product.

The rest of this paper is organized as follows. Section II presents the main result of this paper concerning the conditioning of block subdictionaries. Section III leverages the result of Section II and presents an average-case analysis of convex optimization-based block-sparse recovery from noiseless measurements, along with some discussion and numerical experiments. Section IV makes use of the result of Section II to present an average-case analysis of block-sparse regression and the associated numerical experiments. Finally, some concluding remarks are provided in Section V. For the sake of clarity of exposition, we relegate the proofs of most of the lemmas and theorems to several appendices.

II. Conditioning of Random Block Subdictionaries

In this section, we state and discuss the main result of this paper concerning the conditioning of block subdictionaries of the $n\times p$ dictionary Φ. Here, and in the following, it is assumed that Φ has a block structure that comprises $r = p/m$ blocks of dimensions $n \times m$ each; in particular, we can write without loss of generality that $\Phi = [\Phi_1\ \Phi_2\ \ldots\ \Phi_r]$, where each block $\Phi_i = [\phi_{i,1}\ \ldots\ \phi_{i,m}]$ is an $n \times m$ matrix. We also assume throughout this paper that the columns of Φ are normalized: $\|\phi_{i,j}\|_2 = 1$ for all $i = 1,\ldots,r$, $j = 1,\ldots,m$. The problem we are interested in addressing in this section is the following. Let $\mathcal{S} \subset \{1,\ldots,r\}$ with $|\mathcal{S}| = k$ and define an $n \times km$ block subdictionary $X = [\Phi_i : i \in \mathcal{S}]$. Then what are the conditions on Φ that will guarantee that the singular values of X concentrate around unity? Since addressing this question for an *arbitrary* subset $\mathcal{S}$ is known to lead to either nonverifiable conditions or pessimistic bounds on k (cf. Section I-A), our focus here is on a subset $\mathcal{S}$ that is drawn uniformly at random from all $\binom{r}{k}$ possible k-subsets of $\{1,\ldots,r\}$.

A. Main Result

Our main result concerning the conditioning of random block subdictionaries relies on a condition that we term the *block incoherence condition* (BIC).

Definition 1 (Block Incoherence Condition)**.** Define the intra-block coherence of the dictionary Φ as

$$\mu_I := \max_{1\leq i\leq r} \|\Phi_i^*\Phi_i - \mathrm{Id}_m\|_2$$

and the inter-block coherence of the dictionary Φ as[2]

$$\mu_B := \max_{1\leq i\neq j\leq r} \|\Phi_i^*\Phi_j\|_2.$$

[2]See [48] for a related measure of block coherence of a dictionary that is given by μ_B/m.

We say that Φ satisfies the *block incoherence condition* (BIC) with parameters (c_1, c_2) if $\mu_I \leq c_1$ and $\mu_B \leq c_2/\log p$ for some positive numerical constants c_1 and c_2.

Note that μ_I measures the deviation of individual blocks $\{\Phi_i\}$ from being orthonormal and is identically equal to zero for the case of orthonormal blocks. In contrast, μ_B measures the similarity between different blocks and cannot be zero in the n smaller than p setting. Informally, the BIC dictates that individual blocks of Φ do not diverge from being orthonormal in an unbounded fashion and the dissimilarity between different blocks scales as $O(1/\log p)$. The most desirable aspect of the BIC is that it can be verified in polynomial time. We are now ready to state our first result.

Theorem 1. *Suppose that the $n \times p$ dictionary $\Phi = [\Phi_1\ \Phi_2\ \ldots\ \Phi_r]$ satisfies the BIC with parameters (c_1, c_2). Let $\mathcal{S}$ be a k-subset drawn uniformly at random from all $\binom{r}{k}$ possible k-subsets of $\{1,\ldots,r\}$. Then, as long as $k \leq c_0 r/(\|\Phi\|_2^2 \log p)$ for some positive numerical constant c_0 that depends only on (c_1, c_2), the singular values of the block subdictionary $X = [\Phi_i : i \in \mathcal{S}]$ satisfy $\sigma_i(X) \in [\sqrt{1/2}, \sqrt{3/2}]$, $i = 1,\ldots,km$, with probability with respect to the random choice of the subset $\mathcal{S}$ of at least $1 - 2p^{-4\log 2}$.*

Remark 1. The interval $[\sqrt{1/2}, \sqrt{3/2}]$ in Theorem 1 is somewhat arbitrary. In general, it can be replaced with $[\sqrt{1-\epsilon}, \sqrt{1+\epsilon}]$ for any $\epsilon \in (0, 1)$, resulting in the probability of success either increasing ($\epsilon > 1/2$) or decreasing ($\epsilon < 1/2$).

In words, Theorem 1 states that if a dictionary satisfies the BIC then most of its block subdictionaries of dimensions $n \times km$ act as isometries on $\mathbb{R}^{km}$ for $k = O(r/(\|\Phi\|_2^2 \log p))$. In order to better understand the bound $k = O(r/(\|\Phi\|_2^2 \log p))$, notice that $\|\Phi\|_2^2 \geq p/n$ for the case of a normalized dictionary [24], implying $r/(\|\Phi\|_2^2 \log p) = O(n/(m \log p))$. More importantly, the equality $\|\Phi\|_2^2 = p/n$ is achievable by dictionaries with orthogonal rows (also referred to as tight frames [21]), implying Theorem 1 allows optimal scaling of the dimensions of well-conditioned block subdictionaries. Perhaps the most surprising aspect of this theorem, which sets it apart from other works on inference under linear models in block settings [37–39, 45, 48, 58, 60], is the assertion it makes about the effects of different measures of Φ on the conditioning of random block subdictionaries. Roughly, Theorem 1 suggests that as soon as the BIC is satisfied, both μ_I and μ_B stop playing a role in determining the order-wise dimensions of the subdictionaries that are well conditioned; rather, it is the spectral norm of the dictionary $\|\Phi\|_2$ that plays a primary role in this regard. Such an assertion of course needs to be carefully examined, given that Theorem 1 is only concerned with sufficient conditions. Nevertheless, carefully planned numerical experiments carried out in the context of block-sparse recovery (cf. Section III) and block-sparse regression (cf. Section IV) are consistent with this assertion.

B. Proof of Theorem 1

The proof of Theorem 1 leverages the analytical tools employed by Tropp in [25] for conditioning of canonical (i.e.,

non-block) random subdictionaries, coupled with a *Poissonization* argument that is now standard in the literature (see, e.g, [30]). To proceed, we define r independent and identically distributed (i.i.d.) Bernoulli random variables $\zeta_1, \dots, \zeta_r$ with parameter $\delta := k/r$ (i.e., $\mathbb{P}(\zeta_i = 1) = \delta$) and a random set $\mathcal{S}' := \{i : \zeta_i = 1\}$. Next, we define a random block subdictionary $X' := [\Phi_i : i \in \mathcal{S}']$ and use $G := \Phi^*\Phi - \mathrm{Id}$ and $F := X'^*X' - \mathrm{Id}$ to denote the hollow Gram matrix of Φ and the hollow Gram matrix of X', respectively. Finally, define $\Sigma := \mathrm{diag}(\zeta_1, \dots, \zeta_r)$ to be a random diagonal matrix, $R := \Sigma \otimes \mathrm{Id}_m$ to be a block masking matrix, and notice from definition of the spectral norm that $\|F\|_2 = \|RGR\|_2$. Using this notation, we can show that the L_q norm of the random variable $\|RGR\|_2$ for $q = 4\log p$ is controlled by μ_I, μ_B, and $\|\Phi\|_2$.

Lemma 1. *For $\delta = k/r$ and $q = 4\log p$, the L_q norm of the random variable $\|RGR\|_2 = \|F\|_2$ can be bounded as*

$$\mathbb{E}_q\|RGR\|_2 \le 48\mu_B \log p + 17\sqrt{\delta \log p(1+\mu_I)}\|\Phi\|_2 + 2\delta\|\Phi\|_2^2 + 3\mu_I.$$

The proof of Lemma 1, which is fundamental to the proof of Theorem 1 and comprises novel generalizations of some of the results in [25, 62–64] to the block setting of this paper, is provided in Appendix A. We are now ready to provide a proof of Theorem 1.

Proof of Theorem 1: Define $Z := \|X^*X - \mathrm{Id}\|_2$ and notice that $\sigma_i(X) \in [\sqrt{1/2}, \sqrt{3/2}]$, $i = 1, \dots, km$, if and only if $Z \le 1/2$. Instead of studying Z directly, however, we first study the related random variable $Z' := \|X'^*X' - \mathrm{Id}\|_2 = \|F\|_2$, where X' is the random subdictionary defined in relation to Lemma 1. It then follows from the Markov inequality and Lemma 1 that

$$\begin{aligned}\mathbb{P}(Z' > 1/2) &\le (1/2)^{-q}(\mathbb{E}_q[Z'])^q \\ &\le 2^q\Big(48\mu_B \log p + 17\sqrt{\delta \log p(1+\mu_I)}\|\Phi\|_2 + 2\delta\|\Phi\|_2^2 + 3\mu_I\Big)^q, \qquad (1)\end{aligned}$$

where $q := 4\log p$. Next, our goal is to show that for all $t > 0$,

$$\mathbb{P}(Z > t) \le 2\mathbb{P}(Z' > t), \qquad (2)$$

using the Poissonization argument from [30]. Toward this end, we explicitly write $X' = \Phi_{\mathcal{S}'}$ and note that

$$\begin{aligned}&\mathbb{P}(\|\Phi_{\mathcal{S}'}^*\Phi_{\mathcal{S}'} - \mathrm{Id}\|_2 > t) \\ &\quad= \sum_{\ell=0}^{r} \mathbb{P}\left(\|\Phi_{\mathcal{S}'}^*\Phi_{\mathcal{S}'} - \mathrm{Id}\|_2 > t \middle| |\mathcal{S}'| = \ell\right)\mathbb{P}(|\mathcal{S}'| = \ell) \\ &\quad\ge \sum_{\ell=k}^{r} \mathbb{P}\left(\|\Phi_{\mathcal{S}'}^*\Phi_{\mathcal{S}'} - \mathrm{Id}\|_2 > t \middle| |\mathcal{S}'| = \ell\right)\mathbb{P}(|\mathcal{S}'| = \ell) \\ &\quad= \sum_{\ell=k}^{r} \mathbb{P}\left(\|\Phi_{\mathcal{S}_\ell}^*\Phi_{\mathcal{S}_\ell} - \mathrm{Id}\|_2 > t\right)\mathbb{P}(|\mathcal{S}'| = \ell), \qquad (3)\end{aligned}$$

where $\mathcal{S}_\ell$ is a subset drawn uniformly at random from all $\binom{r}{\ell}$ possible ℓ-subsets of $\{1, \dots, r\}$. We now make two observations. First, $|\mathcal{S}'|$ is a binomial random variable with parameters $(r, k/r)$ and therefore $\mathbb{P}(|\mathcal{S}'| \ge k) \ge 1/2$ due to k being the median of $|\mathcal{S}'|$. Second, since each (random) submatrix $\Phi_{\mathcal{S}_{\ell'}}^*\Phi_{\mathcal{S}_{\ell'}} - \mathrm{Id}$ for a given value of $\ell' \le \ell$ is a submatrix of some (random) $\Phi_{\mathcal{S}_\ell}^*\Phi_{\mathcal{S}_\ell} - \mathrm{Id}$ and the spectral norm of a matrix is lower bounded by that of its submatrices, we have that $\mathbb{P}(\|\Phi_{\mathcal{S}_\ell}^*\Phi_{\mathcal{S}_\ell} - \mathrm{Id}\|_2 > t)$ is a nondecreasing function of ℓ. Therefore we can write

$$\begin{aligned}&\mathbb{P}(\|\Phi_{\mathcal{S}'}^*\Phi_{\mathcal{S}'} - \mathrm{Id}\|_2 > t) \\ &\qquad\ge \mathbb{P}\left(\|\Phi_{\mathcal{S}_k}^*\Phi_{\mathcal{S}_k} - \mathrm{Id}\|_2 > t\right)\sum_{\ell=k}^{r}\mathbb{P}(|\mathcal{S}'| = \ell) \\ &\qquad\ge \mathbb{P}\left(\|\Phi_{\mathcal{S}_k}^*\Phi_{\mathcal{S}_k} - \mathrm{Id}\|_2 > t\right)\mathbb{P}(|\mathcal{S}'| \ge k) \\ &\qquad\ge \frac{1}{2}\mathbb{P}\left(\|\Phi_{\mathcal{S}_k}^*\Phi_{\mathcal{S}_k} - \mathrm{Id}\|_2 > t\right) \\ &\qquad= \frac{1}{2}\mathbb{P}\left(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}} - \mathrm{Id}\|_2 > t\right), \qquad (4)\end{aligned}$$

where the last equality follows since $\mathcal{S}_k$ and $\mathcal{S}$ have the same probability distribution. By combining (1) and (4), we therefore obtain

$$\begin{aligned}\mathbb{P}(Z > 1/2) \le 2^{q+1}\Big(&48\mu_B \log p + 17\sqrt{\delta \log p(1+\mu_I)}\|\Phi\|_2 \\ &+ 2\delta\|\Phi\|_2^2 + 3\mu_I\Big)^q. \qquad (5)\end{aligned}$$

Finally, the expression inside parentheses in the above equation can be bounded by $1/4$ for small-enough constants c_0, c_1, and c_2, resulting in $\mathbb{P}(Z > 1/2) \le 2(1/2)^{4\log p} = 2p^{-4\log 2}$. ∎

C. Discussion

Among existing works in the literature focusing on the conditioning of random (non-block) subdictionaries [22–27], [25] and [26] are the ones with the most general and strongest results. Specifically, [22, 23] deal with the case of the dictionary Φ being a concatenation of two orthonormal bases, while [27] studies the case of Φ being a disjoint union of orthonormal bases. The results in [25] and [26] are related to each other in the sense that [26] extends [25] to the case when the subdictionaries of Φ are not necessarily selected uniformly at random. The proof technique employed in this paper for conditioning of random block subdictionaries is inspired by [25] and is rather tight in the sense that in the case of $m = 1$, $\mu_I = 0$, and for a unit-norm dictionary Φ, Lemma 1 reduces to [25, Corollary 5.2]. While we believe our result can be extended to the case when the random block subdictionaries of Φ are selected with a more "structured randomness" by leveraging the insights offered by [26], we leave this for future work.

It is also instructive to note that while Theorem 1 is the most general incarnation of results concerning conditioning of random block subdictionaries, it is rather straightforward to specialize this result for conditioning of random block subdictionaries of *structured* dictionaries. Next, we specialize Theorem 1 to one such structure that corresponds to Φ being a Kronecker product of an arbitrary unit-norm dictionary and a dictionary with orthonormal columns. Such *Kronecker-structured* dictionaries arise in many contexts [35, 65] and have a special connection to the literature on multiple measurement vectors (MMV) [36–39, 44, 45, 53, 54, 59, 61, 66, 67] and multivariate linear regression [55, 68–70] problems. The following

section therefore will also help understand our work in the context of these two research areas.

1) Random block subdictionaries of Kronecker-structured dictionaries with application to multiple measurement vectors problem: Consider an arbitrary, unit-norm $n_1 \times r$ dictionary P and an $n_2 \times m$ dictionary Q with orthonormal columns (i.e., $Q^*Q = \mathrm{Id}_m$), where $n_1 > r$, $n_2 \leq m$, and $n_1 n_2 = n$. Then a corollary of Theorem 1 is that conditioning of random block subdictionaries of the Kronecker-structured dictionary $\Phi = P \otimes Q$ is simply a function of the coherence, $\mu(P) = \max_{i,j:i\neq j} |\langle P_i, P_j\rangle|$, and spectral norm, $\|P\|_2$, of P, where P_i denotes the i^{th} column of P. Formally, this corollary has the following statement.

Corollary 1. *Suppose that the $n\times p$ dictionary $\Phi = P\otimes Q$ with $Q^*Q = \mathrm{Id}_m$ and $\mu(P) \leq c_2/\log p$ for a positive numerical constant c_2. Let $\mathcal{S}$ be a k-subset drawn uniformly at random from all $\binom{r}{k}$ possible k-subsets of $\{1,\ldots,r\}$. Then, as long as $k \leq c_0 r/(\|P\|_2^2 \log p)$ for some positive numerical constant $c_0 := c_0(c_2)$, the singular values of the block subdictionary $X = [\Phi_i = P_i \otimes Q : i \in \mathcal{S}]$ satisfy $\sigma_i(X) \in [\sqrt{1/2}, \sqrt{3/2}]$, $i = 1,\ldots,km$, with probability at least $1 - 2p^{-4\log 2}$. Here, the probability is with respect to the random choice of the subset $\mathcal{S}$.*

Corollary 1 is a simple consequence of properties of Kronecker product. In terms of the spectral norm of Φ, we have $\|P \otimes Q\|_2 = \|P\|_2\|Q\|_2 = \|P\|_2$. In terms of the intra- and inter-block coherences, we note that

$$\begin{aligned}\Phi_i^*\Phi_j &= (P_i \otimes Q)^*(P_j \otimes Q) = (P_i^* \otimes Q^*)(P_j \otimes Q)\\ &= (P_i^*P_j) \otimes (Q^*Q) = \langle P_i, P_j\rangle \mathrm{Id}_m, \end{aligned} \tag{6}$$

which trivially leads to

$$\begin{aligned}\mu_I &= \max_{1\leq i\leq r} \|\Phi_i^*\Phi_i - \mathrm{Id}_m\|_2\\ &= \max_{1\leq i\leq r} \|(\langle P_i, P_i\rangle - 1)\mathrm{Id}_m\|_2 = 0, \quad \text{and} \end{aligned} \tag{7}$$

$$\mu_B = \max_{1\leq i\neq j\leq r} \|\Phi_i^*\Phi_j\|_2 = \max_{1\leq i\neq j\leq r} |\langle P_i, P_j\rangle| = \mu(P). \tag{8}$$

Note that Corollary 1 is not the tightest possible result for Kronecker-structured dictionaries since Theorem 1 does not exploit any dictionary structure. In particular, one can obtain a variant of Corollary 1 in which the $\log p$ terms are replaced with the $\log r$ terms by explicitly accounting for the Kronecker structure in the proof of Lemma 1.

We conclude this section by connecting Corollary 1 to the MMV/multivariate linear regression problem, which will help clarify the similarities and differences between our work and MMV-related works.[3] The inference problems studied under the MMV setting are essentially special cases of inference in the block-sparse setting studied here. In the MMV setting, it is assumed there are m parameter vectors, $b_1,\ldots,b_m$, collected as m columns of an $r \times m$ matrix B. In addition, each b_i is observed using the same $n \times r$ dictionary A, $y_i := Ab_i$, and the observation vectors $y_1,\ldots,y_m$ are collected as m columns of an $n \times m$ matrix $Y := AB$. A typical assumption in this MMV setting states that the m different parameter vectors share locations of their $k \ll r$ nonzero entries, resulting in B having no more than k nonzero *rows*. It is however easy to see that if we define $y := \mathrm{vec}(Y^T)$ and $b := \mathrm{vec}(B^T)$ then $y = (A \otimes \mathrm{Id}_m)b$, where the vector b exhibits block sparsity. In other words, inference in the MMV setting requires understanding the conditioning of block subdictionaries of $A \otimes \mathrm{Id}_m$. In this case, we already know from Corollary 1 that the conditioning of random block subdictionaries of $A \otimes \mathrm{Id}_m$ is simply a function of the coherence and spectral norm of A. Interestingly, while there exists a significant body of literature in the MMV setting [36–39, 44, 45, 53–55, 59, 61, 66–70], most of these works do not provide near-optimal, verifiable conditions for guaranteeing success of MMV-based inference problems. The most notable exception to this is the recent work [45], which studies the problem of noiseless recovery in the MMV setting. Nonetheless, our block-sparsity results (including the forthcoming noiseless recovery results) are much more general than the ones in [45] because of the MMV setting being just a special case of Corollary 1 in the block-sparse setting.

[3] In the operational sense, MMV and multivariate linear regression are two distinct inference problems. For ease of exposition, however, we use the term MMV in here to refer to both problems.

III. Application: Recovery of Block-Sparse Signals from Noiseless Measurements

We now shift our focus to the applicability of Theorem 1 in the context of inference problems. We first begin with the problem of recovery of β from $y = \Phi\beta$ when the signal β is block sparse. Block sparsity is one of the most popular structures used in sparse signal recovery problems. It is also intrinsically linked with the multiple measurement vectors (MMV) problem described in Section II-C, as there is an equivalent block-sparse formulation for each MMV problem. Block sparsity arises in many applications, including union-of-subspaces models [44, 71], multiband communications [33, 72], array processing [73, 74], and multi-view medical imaging [74–76].

Because of the relevance of block sparsity in these and other applications, significant efforts have been made toward development of block-sparse signal recovery methods/algorithms and matching guarantees on the number of measurements required for successful recovery [36–39, 43–48, 50–54, 59–61, 67]. However, the results reported in some of these works are only applicable in the case of randomized dictionary constructions [43, 44, 46, 47, 61], while those reported in other works rely on dictionary conditions that either cannot be explicitly verified in polynomial time [36, 37, 39, 44, 46, 50, 53, 54, 59, 60, 67] or result in a suboptimal scaling of the number of measurements due to their focus on the worst-case performance [37–39, 48, 50–52, 60].

To the best of our knowledge, the only work that does not have the aforementioned limitations is [45]. Nonetheless, the focus in [45] is only on the restrictive MMV problem, rather than the general block-sparse signal recovery problem. In addition, the analytical guarantees provided in [45] rely on the nonzero entries of β following either Gaussian or spherical distributions. In contrast, we make use of the main result

of Section II in the following to state a result for average-case recovery of block-sparse signals that suffers from none of these and earlier limitations. Our result depends primarily on the spectral norm of Φ, while it has a mild dependence on the intra- and inter-block coherence through the BIC; all three of these quantities can be explicitly computed in polynomial time. It further requires only weak assumptions on the distribution of the nonzero entries of β. Equally important, the forthcoming result does not suffer from the so-called "square-root bottleneck" [26]; specifically, it allows near-optimal scaling of the sparsity level km as a function of the number of measurement n for dictionaries Φ with small spectral norms (e.g., tight frames).

A. Recovery of Block-Sparse Signals: Problem Formulation

Our exposition throughout the rest of this section will be based upon the following formulation. We are interested in recovering a block-sparse signal $\beta \in \mathbb{R}^p$ from noiseless measurements $y = \Phi\beta$, where the dictionary Φ denotes an $n \times p$ observation matrix with $n \ll p$ and $y \in \mathbb{R}^n$ denotes the observation vector. We assume β comprises a total of r blocks, each of size m (yielding $p = rm$), and represent it without loss of generality as $\beta = [\beta_1^* \ \beta_2^* \ \dots \ \beta_r^*]^*$ with each block $\beta_i \in \mathbb{R}^m$. In order to make this problem well posed, we require that β is k-block sparse with $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll r$. Finally, we impose a mild statistical prior on β, as described below.

- M1) The *block support* of β, $\mathcal{S} = \{i : \beta_i \neq \mathbf{0}\}$, has a uniform distribution over all k-subsets of $\{1, \dots, r\}$,
- M2) Entries in β have zero median (i.e., the nonzero entries are equally likely to be positive and negative): $\mathbb{E}(\mathrm{sign}(\beta)) = \mathbf{0}$, and
- M3) Nonzero blocks of the block-sparse signal β have statistically independent "directions." Specifically, we require $\mathbb{P}\left(\bigcap_{i\in\mathcal{S}} \left(\overline{\mathrm{sign}}(\beta_i) \in \mathcal{A}_i\right)\right) = \prod_{i\in\mathcal{S}} \mathbb{P}\left(\overline{\mathrm{sign}}(\beta_i) \in \mathcal{A}_i\right)$, where $\mathcal{A}_i \subset \mathbb{S}^{m-1}$ with $\mathbb{S}^{m-1}$ denoting the unit sphere in $\mathbb{R}^m$.

Note that M2 and M3 are trivially satisfied in the case of the nonzero blocks of β drawn independently from either Gaussian or spherical distributions. However, it is easy to convince oneself that many other distributions—including those that are not absolutely continuous—will satisfy these two conditions. Conditions M1–M3 provide a probabilistic characterization of block-sparse β that is inspired by Tropp [24] and Candès and Plan [30] in which a related characterization of non block-sparse β helped them overcome some analytical hurdles in relation to performance specifications of sparse recovery and regression problems, respectively.

B. Main Result and Discussion

In this section, we are interested in understanding the average-case performance of the following mixed-norm convex optimization program for recovery of block-sparse signals satisfying M1–M3:

$$\widehat{\beta} = \arg\min_{\bar{\beta}\in\mathbb{R}^p} \|\bar{\beta}\|_{2,1} \text{ such that } y = \Phi\bar{\beta}, \tag{9}$$

where the $\ell_{2,1}$ norm of a vector $\beta \in \mathbb{R}^p$ containing r blocks of m entries each is defined as $\|\beta\|_{2,1} := \sum_{i=1}^r \|\beta_i\|_2$. While (9) has been utilized in the past for recovery of block sparse signals (see, e.g., [43–45]), an average-case analysis result along the following lines is novel. The following theorem is proven in Appendix B.

Theorem 2. *Suppose that $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that β is observed according to the linear model $y = \Phi\beta$, where the $n \times p$ matrix Φ satisfies the BIC with some parameters (c_1, c_2). Then, as long as $k \leq c_0 r/\|\Phi\|_2^2 \log p$ for some positive numerical constant $c_0 := c_0(c_1, c_2)$, the minimization (9) results in $\widehat{\beta} = \beta$ with probability at least $1 - 4p^{-4\log 2}$.*

Interestingly, Theorem 2 specialized to the case of non-block sparse signals (by setting $m = 1$ and $r = p$) gives us an average-case analysis result for recovery of sparse signals that has never been explicitly stated in prior works. The optimization program (9) in this case reduces to the standard *basis pursuit* program [77]:

$$\widehat{\beta} = \arg\min_{\bar{\beta}\in\mathbb{R}^p} \|\bar{\beta}\|_1 \text{ such that } y = \Phi\bar{\beta}, \tag{10}$$

the BIC reduces to a bound on the coherence of Φ, and Theorem 2 reduces to the following corollary.[4]

Corollary 2. *Suppose $\beta \in \mathbb{R}^p$ is k-sparse, its support (i.e., locations of its nonzero entries) is a k-subset drawn uniformly at random from all $\binom{p}{k}$ possible k-subsets of $\{1, \dots, p\}$, its nonzero entries are drawn from a multivariate distribution with zero median (i.e., the nonzero entries are equally likely to be positive and negative), and the signs of the nonzero entries are independent. Then, as long as $k \leq c_0' p/\|\Phi\|_2^2 \log p$ and $\mu(\Phi) \leq c_1'/\log p$ for some positive numerical constants $c_0' := c_0'(c_1')$ and c_1', the minimization (10) successfully recovers β from $y = \Phi\beta$ with probability at least $1 - 4p^{-4\log 2}$.*

We now elaborate on the similarities and differences between our (average-case) guarantees for recovery of block-sparse (Theorem 2) and non-block sparse (Corollary 2) signals. In terms of similarities, both results allow for the same scaling of the *total number of nonzero entries* in β: $km = O(p/\|\Phi\|_2^2 \log p)$ in the case of block-sparse signals and $k = O(p/\|\Phi\|_2^2 \log p)$ in the case of sparse signals. However, while Corollary 2 requires that the inner product of any two columns in Φ be $O(1/\log p)$, Theorem 2 allows for less restrictive inner products of columns *within* blocks as long as $\mu_I = O(1)$. Stated differently, explicit exploitation of the block structure in (9) enables us to shift the $O(1/\log p)$ scaling requirement from $\mu(\Phi)$ in Corollary 2 to μ_B in Theorem 2. Similarly, while Corollary 2 requires that the signs of the nonzero entries in β be independent, Theorem 2 allows for correlations among the signs of entries *within* nonzero blocks. With the caveat that Theorem 2 and Corollary 2 only specify

[4]We refer the reader to the forthcoming discussion for the difference between the average-case analysis result in [24] and Corollary 2. While it is possible to leverage the results in [25] for obtaining Corollary 2, rather than obtaining Corollary 2 from Theorem 2 in this paper, such a result does not explicitly exist in prior literature to the best of our knowledge.

sufficient conditions, these two results seem to suggest that explicitly accounting for block structures in sparse signals allows one to expand the classes of sparse signals β *and* dictionaries Φ under which successful (average-case) recovery can be guaranteed.

Next, we comment on the tightness of the scaling on the number of nonzero entries in Theorem 2 and Corollary 2. Assuming appropriate conditions on statistical properties of β and (intra-/inter-block) coherence of Φ are satisfied, both results allow for the number of nonzero entries to scale like $O(n/\log p)$ for dictionaries Φ that are "approximately" tight frames [21]: $\|\Phi\|_2^2 \approx p/n$. This suggests a near-optimal nature of both results (modulo perhaps $\log$ factors) as one cannot expect better than linear scaling of the number of nonzero entries as a function of the number of observations. In particular, existing literature on frame theory [78] can be leveraged to specialize these results for oft-used classes of random dictionaries (e.g., Gaussian, random partial Fourier) and to establish that in such cases the scaling of our guarantee matches that obtained using nonverifiable conditions such as the restricted isometry property [15, 79].

Additionally, we note that when Corollary 2 is specialized to the case of (approximately) tight frames we obtain average-case guarantees somewhat similar to the ones reported in [24, Theorem 14]. The main difference between the two results is the role that the coherence $\mu(\Phi)$ plays in the guarantees. In [24, Theorem 14], the maximum allowable sparsity k is required to be inversely proportional to $\mu^2(\Phi)$. In contrast, we assert that the maximum allowable sparsity scaling is not fundamentally determined by the coherence. Numerical experiments reported in the following section suggest that this is indeed the case.

Remark 2. Despite the order-wise tightness of Theorem 2 for tight frames, it is important to note that the bound on k in that theorem is only a sufficient condition. In particular, we do not have a converse that shows the impossibility of recovering β from y when this bound is violated. To the best of our knowledge, however, no such converses exist even in the non block-sparse setting for arbitrary dictionaries and/or non-asymptotic analysis (cf. [80, 81]).

C. Numerical Experiments

One of the fundamental takeaways of this section is that the spectral norm of the dictionary, rather than the (intra-/inter-block) coherence of the dictionary, determines the maximum allowable sparsity in (block)-sparse signal recovery problems. In order to experimentally verify this insight, we performed a set of block-sparse signal recovery experiments with carefully designed dictionaries having varying spectral norms and coherence values. Throughout our experiments, we set the signal length to $p = 5000$, the block size and the number of blocks to $m = 10$ and $r = 500$, respectively, and the number of observations to $n = 858$ (computed from the bound in [82] for $k = 20$ nonzero blocks). In order to design our dictionaries, we first used Matlab's random number generator to obtain 2000 matrices with unit-norm columns. Next, we manipulated the singular values of each of these matrices to increase their spectral norms by a set of integer multipliers $\mathcal{T}$. Finally, for each of the $2000 \cdot |\mathcal{T}|$ resulting matrices, we normalized their columns to obtain our dictionaries and recorded their spectral norms $\|\Phi\|_2$, coherences $\mu(\Phi)$, inter-block coherences $\mu_B(\Phi)$, and intra-block coherences $\mu_I(\Phi)$.

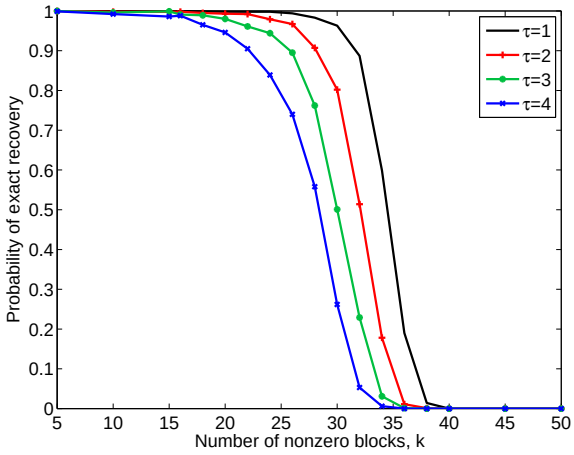

Fig. 1. Performances of dictionaries Φ with varying spectral norms and roughly equal coherences in block-sparse signal recovery as a function of the number of nonzero blocks k; $\tau \in \mathcal{T}$ denotes the value of the spectral norm multiplier used to generate the dictionary.

τ	1	2	3	4
$\|\Phi_\tau\|_2$	3.3963	6.7503	10.0547	13.2034
$\mu(\Phi_\tau)$	0.1992	0.2026	0.2000	0.2207
$\mu_B(\Phi_\tau)$	0.2973	0.3431	0.5573	0.8490
$\mu_I(\Phi_\tau)$	0.1992	0.2026	0.2177	0.3787

TABLE I
SPECTRAL NORMS AND COHERENCES FOR THE DICTIONARIES USED IN THE EXPERIMENTS OF FIGURE 1.

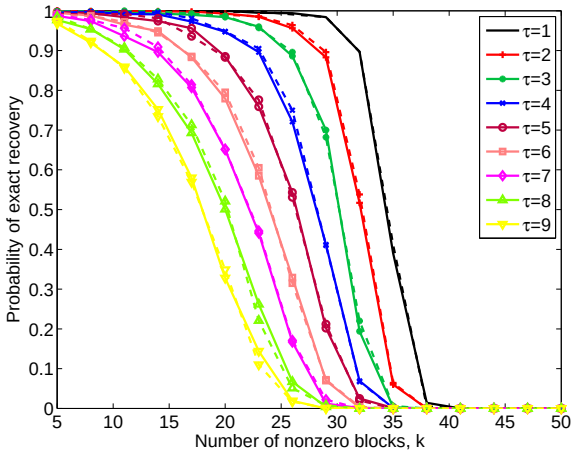

Fig. 2. Performances of dictionaries Φ with varying spectral norms and extremal coherence values (cf. Table II) in block-sparse signal recovery as a function of the number of nonzero blocks k; $\tau \in \mathcal{T}$ denotes the value of the spectral norm multiplier used. Solid lines correspond to dictionaries with minimum coherence, while dashed lines correspond to dictionaries with maximum coherence.

We evaluate the block-sparse signal recovery performance of each resulting dictionary Φ using Monte Carlo trials, corresponding to the generation of 1000 block-sparse signals with k nonzero blocks. Each signal has block support selected uniformly at random according to M1 and nonzero entries drawn independently from the standard Gaussian distribution $\mathcal{N}(0, \mathrm{Id})$. We then obtain the observations $y = \Phi\beta$ using the dictionary Φ under study for each one of these signals and perform recovery using the minimization (9).[5] We define successful recovery to be the case when the block support of $\widehat{\beta}$ matches the block support of β and the submatrix of Φ with columns corresponding to the block support of β has full rank.

Figure 1 shows the performances of dictionaries Φ of increasing spectral norms ($\mathcal{T} = \{1, 2, 3, 4\}$), where we choose the dictionary (among the 2000 available options) whose coherence value is closest to 0.2. The spectral norms, coherences, inter-block coherences, and intra-block coherences for these chosen dictionaries are collected in Table I. The performance is shown as a function of the number of nonzero blocks k in the signal. The figure shows a consistent improvement in the values of k for which successful recovery is achieved as the spectral norm of the dictionary decays, even though $\mu(\Phi)$ does not significantly change among the dictionaries.

To further emphasize strong dependence of sparse-signal recovery on spectral norm and weak dependence on (intra-/inter-block) coherences, Figure 2 shows the performance of dictionaries Φ with increasing spectral norms ($\mathcal{T} = \{1, \ldots, 9\}$), where we choose dictionaries with the largest and smallest coherence values for each $\tau \in \mathcal{T}$ (among the 2000 available options). The spectral norms, coherences, inter-block coherences, and intra-block coherences for these chosen dictionaries are collected in Table II. The figure shows not only the same consistent improvement as the spectral norm of the dictionary decays, but also that significant changes in the values of the (intra-/inter-block) coherences do not significantly affect the recovery performance. This behavior agrees with our expectation from Theorem 2 that the role of the intra-/inter-block coherences in performance guarantees is limited to the BIC and is decoupled from the scaling of the number of nonzero blocks k (equivalently, number of nonzero entries km) in the signal.

IV. Application: Linear Regression of Block-Sparse Vectors

In this section, we leverage Theorem 1 to obtain average-case results for linear regression of block-sparse vectors, defined as estimating $\Phi\beta$ from $y = \Phi\beta + \text{noise}$ when β has a block-sparse structure. In particular, we focus on two popular convex optimization-based methods, the lasso [84] and the group lasso [31], for characterizing results for linear regression of block-sparse vectors. Our focus on these two methods is due to their widespread adoption by the signal processing and statistics communities. In the signal processing literature, these methods are typically used for efficient sparse approximations of arbitrary signals in overcomplete dictionaries. In the statistics literature, they are mostly used for efficient variable selection and reliable regression under the linear model assumption. Nonetheless, ample empirical evidence in both fields suggests that an appropriately regularized group lasso can outperform the lasso whenever there is a natural grouping of the dictionary atoms/regression variables in terms of their contributions to the observations [31, 32]. In this section, we analytically characterize the linear regression performances of both the lasso and the group lasso for block-sparse vectors, which helps us highlight one of the ways in which the group lasso might outperform the lasso for regression problems.

Note that analytical characterization of the group lasso using ℓ_1/ℓ_2 regularization for the "underdetermined" setting, in which one can have far more regression variables than observations ($n \ll p$), has received attention recently in the statistics literature [32, 40–42, 49, 55–58]. However, prior analytical work on the performance of the group lasso either studies an asymptotic regime [32, 40–42], focuses on random design matrices (i.e., dictionaries) [32, 41, 55, 56], and/or relies on conditions that are either computationally prohibitive to verify [40, 42, 49, 57] or that do not allow for near-optimal scaling of the number of observations with the number of active blocks of regression variables k [58]. In contrast, our analysis for the regression performance of the group lasso using ℓ_1/ℓ_2 regularization in the underdetermined case for block-sparse vectors circumvents these shortcomings of existing works by adopting a probabilistic model, described by the conditions M1–M3 in Section III-A, for the blocks of regression coefficients in β. To the best of our knowledge, the result stated in the sequel concerning the linear regression performances of the group lasso[6] for block-sparse vectors is the first one for block linear regression that is non-asymptotic in nature and applicable to arbitrary design matrices through verifiable conditions, while still allowing for near-optimal scaling of the number of observations with the number of blocks of nonzero regression coefficients. Our proof techniques are natural extensions of the ones used in [30] for the non-block setting and rely on Theorem 1 for many of the key steps.

A. Regression of Block-Sparse Vectors: Problem Formulation

This section concerns regression in the "underdetermined" setting for the case when the observations $y \in \mathbb{R}^n$ can be approximately explained by a linear combination of a small number of blocks ($k < n \ll p$) of regression variables (predictors). Mathematically, we have that $y = \Phi\beta + z$, where Φ denotes the design matrix (dictionary) containing one regression variable per column, $\beta \in \mathbb{R}^p = [\beta_1^* \ \beta_2^* \ \ldots \ \beta_r^*]^*$ denotes the k-block sparse vector of regression coefficients corresponding to these variables (i.e., $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll r$), and $z \in \mathbb{R}^n$ denotes the modeling error. Here, we assume without loss of generality that Φ has unit-norm columns, while we assume the modeling error z to be an i.i.d. Gaussian vector with variance σ^2. Finally, in keeping with the earlier discussion, we impose a mild statistical prior on the vector of regression coefficients β that is given by the conditions M1,

[5] We used the `SPGL1` Matlab package [83] in all simulations in this section.

[6] We refer to the group lasso using ℓ_1/ℓ_2 regularization as "group lasso" throughout the rest of this paper for brevity.

τ	1	2	3	4	5	6	7	8	9
$\|\Phi_{\tau,\min}\|_2$	3.4064	6.7726	10.0536	13.2034	16.3421	19.2980	22.1413	24.6710	27.2951
$\|\Phi_{\tau,\max}\|_2$	3.3963	6.7503	10.0543	13.2250	16.2747	19.1506	21.9975	24.7026	27.3199
$\mu(\Phi_{\tau,\min})$	0.1230	0.1198	0.1500	0.2207	0.2964	0.3760	0.4583	0.5337	0.6000
$\mu(\Phi_{\tau,\max})$	0.1992	0.2026	0.2698	0.3816	0.4863	0.5778	0.6566	0.7225	0.7758
$\mu_B(\Phi_{\tau,\min})$	0.2887	0.3177	0.5357	0.8490	1.2917	1.7372	2.2263	2.5989	3.1204
$\mu_B(\Phi_{\tau,\max})$	0.2973	0.3431	0.6516	1.0287	1.4419	1.6616	2.0230	2.4479	2.8737
$\mu_I(\Phi_{\tau,\min})$	0.1487	0.2002	0.3368	0.3787	0.3472	0.4385	0.5462	1.0551	1.3095
$\mu_I(\Phi_{\tau,\max})$	0.1992	0.2026	0.2698	0.3816	0.4863	0.5778	0.8273	1.0415	1.2723

TABLE II
SPECTRAL NORMS AND BLOCK COHERENCES FOR THE DICTIONARIES USED IN THE EXPERIMENTS OF FIGURE 2 AND FIGURE 3.

M2, and M3 in Section III-A. The fundamental goal in here then is to obtain an estimate $\widehat{\beta}$ from the observations y such that $\Phi\widehat{\beta}$ is as close to $\Phi\beta$ as possible, where the closeness is measured in terms of the ℓ_2 regression error, $\|\Phi\beta - \Phi\widehat{\beta}\|_2$.

B. Main Results and Discussion

In this section, we are interested in understanding the average-case regression performance of two methods in the block-sparse setting. The first one of these methods is the lasso [84], which ignores any grouping of the regression variables and estimates the vector of regression coefficients as

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - \Phi\beta\|_2^2 + 2\lambda\sigma\|\beta\|_1, \tag{11}$$

where $\lambda > 0$ is a tuning parameter. In terms of a baseline result for the lasso, we can extend the probabilitic model of Candès and Plan [30] for non-block linear regression to the block setting and state the following theorem that follows from Theorem 1 in this paper and the proof of [30, Theorem 1.2].

Theorem 3 ([30, Theorem 1.2] and Theorem 1)**.** *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and that the observation vector can be modeled as $y = \Phi\beta + z$ with the modeling error z being i.i.d. Gaussian with variance σ^2. Further, assume that β is drawn according to the statistical model M1 and M2 with the signs of its nonzero entries being i.i.d., and the $n \times p$ matrix Φ satisfies (i) $\mu(\Phi) = O(1/\log p)$ and (ii) the BIC with some parameters (c_1', c_2'). Then, as long as $k \leq c_0' r/\|\Phi\|_2^2 \log p$ for some positive numerical constant $c_0' := c_0'(c_1', c_2')$, the lasso estimate $\widehat{\beta}$ in* (11) *computed with $\lambda = \sqrt{2\log p}$ obeys*

$$\|\Phi\beta - \Phi\widehat{\beta}\|_2^2 \leq C' mk\sigma^2 \log p$$

with probability at least $1 - O(p^{-1})$, where $C' > 0$ is a constant independent of the problem parameters.

The proof of this theorem is omitted here because it is a straightforward extension. While this theorem suggests that the lasso solution in the block setting enjoys many of the optimality properties of the lasso solution in the non-block setting (see, e.g., the discussion in [30]), it fails to extend to the case when the independence assumption on the signs of the nonzero regression coefficients is replaced by the less restrictive condition M3. In particular, one expects that allowing for arbitrary correlations within the blocks of regression coefficients will limit the usefulness of the lasso for linear regression in the presence of large blocks. While such an insight can be difficult to confirm in the case of arbitrary design matrices and average-case analysis, we provide an extension of Theorem 3 in the following that highlights the challenges for the lasso in the case of regression of block-sparse vectors with arbitrarily correlated blocks.

Theorem 4. *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that the observation vector can be modeled as $y = \Phi\beta + z$, where the $n \times p$ matrix Φ satisfies $\mu_I(\Phi) \leq c_1''$ and $\mu_B(\Phi) \leq c_2''/(\sqrt{m}\log p)$ for some positive numerical constants c_1'', c_2'', and the modeling error z is i.i.d. Gaussian with variance σ^2. Then, as long as $k \leq c_0'' r/\|\Phi\|_2^2 m \log p$ for some positive numerical constant $c_0'' := (c_1'', c_2'')$, the lasso estimate $\widehat{\beta}$ in* (11) *computed with $\lambda = \sqrt{2\log p}$ obeys*

$$\|\Phi\beta - \Phi\widehat{\beta}\|_2^2 \leq C'' mk\sigma^2 \log p$$

with probability at least $1 - p^{-1}(2\pi\log p)^{-1/2} - 8p^{-4\log 2}$, where $C'' > 0$ is a constant independent of the problem parameters.

The proof of this theorem is omitted here since it shares many similarities with the proof of the main result of this section for the group lasso. Nonetheless, interested readers can read it in [85, Appendix C]. It can be seen from Theorems 3 and 4 that while both the theorems guarantee same scaling of the regression error, the scalings of the maximum number of allowable nonzero blocks and the block coherence in Theorem 4 match the ones in Theorem 3 only for the case of $m = O(1)$; otherwise, Theorem 4 with correlated blocks results in less-desirable scalings of k and $\mu_B(\Phi)$. It can be seen from the proof of Theorem 4 in [85, Appendix C] that this dependence upon m—the size of the blocks—is a direct consequence of allowing for arbitrary correlations within blocks. A natural question to ask then is whether it is possible to return to the scalings of Theorem 3 *without* sacrificing intra-block correlations. The answer to this question is in the affirmative as long as one explicitly accounts for the block structure of the regression problem.

Specifically, the group lasso explicitly accounts for the grouping of the regression variables in its formulation and estimates the vector of regression coefficients as

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - \Phi\beta\|_2^2 + 2\lambda\sigma\sqrt{m}\|\beta\|_{2,1}, \tag{12}$$

where $\lambda > 0$ is once again a tuning parameter. The following theorem shows that the group lasso can achieve the

same scaling results as the lasso for block-sparse vectors (cf. Theorem 3), while allowing for arbitrary correlations among the regression coefficients within blocks. The proof of this theorem is given in Appendix C.

Theorem 5. *Suppose that the vector of regression coefficients $\beta \in \mathbb{R}^p$ is k-block sparse and it is drawn according to the statistical model M1, M2, and M3. Further, assume that the observation vector can be modeled as $y = \Phi\beta + z$, where the $n \times p$ matrix Φ satisfies the BIC with some parameters (c_1, c_2), and the modeling error z is i.i.d. Gaussian with variance σ^2. Then, as long as $k \leq c_0 r / \|\Phi\|_2^2 \log p$ for some positive numerical constant $c_0 := (c_1, c_2)$, the group lasso estimate $\widehat{\beta}$ in* (12) *computed with $\lambda = \sqrt{2 \log p}$ obeys*

$$\|\Phi\beta - \Phi\widehat{\beta}\|_2^2 \leq Cmk\sigma^2 \log p$$

with probability at least $1 - p^{-1}(2\pi \log p)^{-1/2} - 8p^{-4 \log 2}$, where $C > 0$ is a constant independent of the problem parameters.

Remark 3. Note that if one has $m = 1$ then block sparsity reduces to the canonical sparsity, the group lasso (12) reduces to the lasso (11), the block coherence $\mu_B(\Phi)$ reduces to the coherence $\mu(\Phi)$, and Theorem 5 essentially reduces to [30, Theorem 1.2].

With the caveat that both Theorems 4 and 5 are concerned with sufficient conditions for average-case regression performance, we now comment on the strengths and weaknesses of these two results. Assuming appropriate conditions are satisfied for the two theorems, we have that both the lasso and the group lasso result in the same scaling of the regression error, $\|\Phi\beta - \Phi\widehat{\beta}\|_2^2 = O(mk\sigma^2 \log p)$, in the presence of intra-block correlations. This scaling of the regression error is indeed the best that any method can achieve, modulo the logarithmic factor, since we are assuming that the observations are described by a total of mk regression variables. Unlike the lasso, however, the group lasso also allows for a more favorable scaling of the maximum number of regression variables contributing to the observations, $km = O(p/\|\Phi\|_2^2 \log p)$, even when arbitrary intra-block correlations are permitted. In fact, similar to the discussion in Section III, it is easy to conclude that this scaling of the number of nonzero regression coefficients is near-optimal since it leads to a linear relationship (modulo logarithmic factors) between the number of observations n and the number of active regression variables km for the case of design matrices that are approximately tight frames: $\|\Phi\|_2^2 \approx p/n$. The other main difference between Theorems 4 and 5 is the role that the inter-block coherence $\mu_B(\Phi)$ plays in guarantees for the lasso and the group lasso. Specifically, Theorem 4 requires the inter-block coherence to be smaller, $\mu_B(\Phi) = O(1/\sqrt{m} \log p)$, than Theorem 5 for the lasso to yield near-optimal regression error in the case of intra-block correlations. This discussion suggests that reliable linear regression of block-sparse vectors can be carried out using the group lasso for a larger class of regression vectors and design matrices than the lasso. We plan to provide a more rigorous mathematical understanding of these and other subtle but important differences between the lasso and the group lasso in future works.

Remark 4 (Beyond canonical block sparsity)**.** While the block-sparse structures of Theorems 1–5 can be found in many applications, there exist other applications in which canonical block sparsity does not adequately capture the sparsity structure of β. Consider, for instance, wavelet expansions of piecewise smooth signals. Nonzero wavelet coefficients in these cases appear for chains of wavelets at multiple scales and overlapping offsets, as captured through parent–children relationships in wavelet trees [46]. More general structured-sparsity models (also known as model-based sparsity and union-of-subspaces models) are often used in the literature to express these kinds of sparsity structures by allowing some supports and disallowing others [46]. The support of structured-sparse signals in many instances can be expressed in terms of unions of groups of indices $\{\Omega_1, \ldots, \Omega_M\}$, where each group $\Omega_m \subset \{1, \ldots, p\}$ contains indices of k coefficients that become active simultaneously in structured-sparse signals, and the groups Ω_m's may or may not be disjoint. Therefore, despite the non-block sparse nature of β in this setting, one can reorganize the columns of the dictionary Φ and the entries of β as $\Phi' = [\Phi_{\Omega_1}\ \Phi_{\Omega_2}\ \ldots\ \Phi_{\Omega_M}]$ and $\beta' = [\beta_{\Omega_1}\ \beta_{\Omega_2}\ \ldots\ \beta_{\Omega_M}]$, respectively, so that $y = \Phi\beta = \Phi'\beta'$. (In the case of overlapping Ω_m's, this does require minor corrections to β' to ensure each nonzero entry in β appears only once in β'; see, e.g., [82, 86].) Doing so converts the structured-sparse problem involving β into a block-sparse problem involving β' and results of Theorems 1–5 can still be utilized in this case as long as the number of sets M is not prohibitively large.

C. Numerical Experiments

One of the most important implications of this section is that, similar to the case of recovery of block-sparse signals, the number of maximum allowable active regression variables in regression of block-sparse vectors is fundamentally a function of the spectral norm of the design matrix, provided the inter- and intra-block coherence of the design matrix are not too large. However, such a claim needs to be carefully investigated since our results are only concerned with sufficient conditions on design matrices. To this end, we construct numerical experiments that help us evaluate the regression performance of the group lasso for a range of design matrices with varying spectral norms, coherences, inter-block coherences, and intra-block coherences. In order to generate these design matrices, we reuse the experimental setup described in Section III-C (corresponding to $n = 858$, $m = 10$, and $r = 500$).

For the sake of brevity, we focus only on the performance of the group lasso (12) for regression of block-sparse vectors.[7] This performance is evaluated for different design matrices using Monte Carlo trials, corresponding to generation of 1000 block-sparse β with k nonzero blocks. Each vector of regression coefficients has block support selected uniformly at random according to M1 in Section III-A and nonzero entries drawn independently from the Gaussian distribution. We then

[7]We used the `SpaRSA` Matlab package [87] with `debias` option turned on in all simulations in this section.

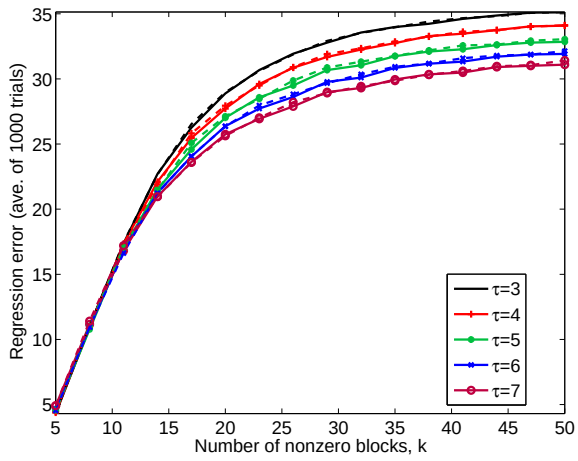

Fig. 3. Performances of the group lasso for design matrices Φ with varying spectral norms and extremal coherence values (cf. Table II) in regression of block-sparse vectors as a function of the number of nonzero regression blocks k; τ denotes the value of the spectral norm multiplier used. Solid lines correspond to matrices with minimum coherence, while dashed lines correspond to matrices with maximum coherence.

obtain the observations $y = \Phi\beta + z$ using the design matrix (dictionary) Φ under study for each one of the block-sparse β, where the variance σ^2 of the modeling error z is selected such that $\|\beta\|_2^2/n\sigma^2 \approx 0.84$. Finally, we carry out linear regression using the group lasso by setting $\lambda \approx 1.4592$ and we then record the regression error $\|\Phi\beta - \Phi\widehat{\beta}\|_2^2$.

Figure 3 shows the regression performance of the group lasso for design matrices Φ with increasing spectral norms ($\tau = \{3, \ldots, 7\}$), where we once again choose matrices with the largest and smallest coherence values for each τ (among the 2000 available options). The spectral norms, coherences, inter-block coherences, and intra-block coherences for these chosen design matrices are still given by Table II in Section III-C. Similar to the case of block-sparse recovery, we observe that significant changes in the values of the (intra-/inter-block) coherences do not significantly affect the regression performance. This behavior is clearly in agreement with our expectation from Theorem 5 that the role of (intra-/inter-block) coherences in regression is limited to the BIC and is decoupled from the scaling of the number of nonzero blocks k (equivalently, number of nonzero regression coefficients km). In addition, we observe that as the spectral norm of the matrix decreases, the range of values of k for which the regression error exhibits a linear trend also increases. This is again consistent with the statement of Theorem 5.

V. Conclusion

In this paper, we have provided conditions under which most block subdictionaries of a dictionary are well conditioned. In contrast to prior works, these conditions are explicitly computable in polynomial time, they lead to near-optimal scaling of the dimensions of the well-conditioned subdictionaries for dictionaries that are approximately tight frames, and they suggest that the spectral norm plays a far important role than the (inter-/intra-block) coherences of the dictionary in determining the order-wise dimensions of the well-conditioned subdictionaries. In addition, we have utilized this result to investigate the average-case performances of convex optimization-based methods for block-sparse recovery and regression of block-sparse vectors. Our average-case performance results significantly improve upon the existing results for both block-sparse recovery and regression of block-sparse vectors. Finally, numerical experiments conducted in the context of block-sparse recovery and regression problems validate the insight offered by our results in relation to the effects of spectral norm and (inter-/intra-block) coherences of the dictionary on inference problems in block-sparse settings.

VI. Acknowledgements

We gratefully acknowledge many helpful comments provided by Dr. Lorne Applebaum in relation to a preliminary draft of this paper. We are also thankful to the anonymous reviewers for their many valuable remarks that helped improve the quality of this paper.

Appendix A
Proof of Lemma 1

The proof of this lemma relies on many lemmas and tools, some of which are generalizations of the corresponding results in [25, 62–64] to the block setting of this paper. To begin, we denote the matrix G in block-partitioned fashion:

$$G = [G_1\ G_2\ \ldots\ G_r] = \begin{bmatrix} G_{1,1} & G_{1,2} & \ldots & G_{1,r} \\ G_{2,1} & G_{2,2} & \ldots & G_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ G_{r,1} & G_{r,2} & \ldots & G_{r,r} \end{bmatrix},$$

where $G_{i,j} = \Phi_i^*\Phi_j - 1_{\{i=j\}}\mathrm{Id}$ for $1 \le i, j \le r$. We then split $G = H + D$, where D contains the diagonal blocks $G_{i,i}$, and H contains only the non-diagonal blocks. We next define the following "norms" for block matrices:

- When we block only columns of a matrix M, we define $\|M\|_{B,1} := \max_{1\le i\le r} \|M_i\|_2$, and
- When we block both columns and rows of a matrix M, we define $\|M\|_{B,2} := \max_{1\le i,j\le r} \|M_{i,j}\|_2$.

Finally, we make use of some standard inequalities in the following, including:

- Cauchy-Schwarz Inequality: $|\mathbb{E}(XY)|^2 \le \mathbb{E}(X^2)\mathbb{E}(Y^2)$.
- Hölder's Inequality: $\|fg\|_1 \le \|f\|_s\|g\|_q$, $1 \le s, q \le \infty$ and $1/s + 1/q = 1$.
- Jensen's Inequality for a convex function f: $f(\mathbb{E}X) \le \mathbb{E}f(X)$.
- Scalar Khintchine Inequality: Let $\{a_i\}$ be a finite sequence of complex numbers and $\{\epsilon_i\}$ be a Rademacher (uniformly random ± 1 binary, i.i.d.) sequence. For each $q \ge 0$, we have

$$\mathbb{E}_q \left| \sum_i \epsilon_i a_i \right| \le C_q \left(\sum_i |a_i|^2 \right)^{1/2},$$

where $C_q \le 2^{1/4}\sqrt{q/e}$.

- Noncommutative Khintchine Inequality [62]: Let $\{M_i\}$ be a finite sequence of matrices of the same dimensions and $\{\epsilon_i\}$ be a Rademacher sequence. For each $q \geq 2$,

$$\mathbb{E}_q \left\| \sum_i \epsilon_i M_i \right\|_{S_q} \leq W_q \max \left\{ \left\| \left(\sum_i M_i M_i^* \right)^{1/2} \right\|_{S_q}, \left\| \left(\sum_i M_i^* M_i \right)^{1/2} \right\|_{S_q} \right\},$$

where $\|M\|_{S_q} := \|\sigma(M)\|_q$ denotes the Schatten q-norm for a matrix M (equal to the ℓ_q-norm of the vector $\sigma(M)$, which contains singular values of the matrix M) and $W_q \leq 2^{-1/4}\sqrt{\pi q/e}$.

We need the following five lemmas in our proof of Lemma 1. The first two lemmas here are used to prove the later ones.

Lemma A.1. *Let $X = [X_1 \ X_2 \ \dots \ X_r]$ be a block matrix and D_X be its block diagonalization, i.e., a block-diagonal matrix $D_X = \mathrm{diag}(X_1, X_2, \dots, X_r)$ containing the matrices $\{X_i\}$ in its diagonal, with all other elements being equal to zero. Then, we have*

$$\|D_X\|_2 \leq \|X\|_{B,1}.$$

Proof: For a vector a of appropriate length, we evaluate the ratio $\frac{\|D_X a\|_2^2}{\|a\|_2^2}$. We partition $a = [a_1^* \ a_2^* \ \dots \ a_r^*]^*$ into its pieces a_i matching the number of columns of the blocks X_i, $1 \leq i \leq r$. Then, we have

$$\begin{aligned}\frac{\|D_X a\|_2^2}{\|a\|_2^2} &= \frac{\sum_{i=1}^r \|X_i a_i\|_2^2}{\sum_{i=1}^r \|a_i\|_2^2} \leq \frac{\sum_{i=1}^r \|X_i\|_2^2 \|a_i\|_2^2}{\sum_{i=1}^r \|a_i\|_2^2} \\ &\leq \frac{\max_{1\leq i\leq r} \|X_i\|_2^2 \sum_{i=1}^r \|a_i\|_2^2}{\sum_{i=1}^r \|a_i\|_2^2} = \max_{1\leq i\leq r} \|X_i\|_2^2.\end{aligned}$$

Thus, the spectral norm obeys

$$\|D_X\|_2 = \max_a \frac{\|D_X a\|_2}{\|a\|_2} \leq \max_{1\leq i\leq r} \|X_i\|_2 = \|X\|_{B,1}.$$

■

The next lemma is a generalization of the lemma in [63, Sec. 2] to our block setting.

Lemma A.2. *Let $X = [X_1 \ X_2 \ \dots \ X_r]$ be a block matrix where each block X_i has m columns with $mr = p$ and let $\{\epsilon_i\}$ be a Rademacher sequence. For any $q \geq 2\log p$, we have*

$$\mathbb{E}_q \left\| \sum_{i=1}^r \epsilon_i X_i X_i^* \right\|_2 \leq 1.5\sqrt{q} \|X\|_{B,1} \|X\|_2.$$

Proof: We start by bounding the spectral norm by the Schatten q-norm:

$$E := \mathbb{E}_q \left\| \sum_{i=1}^r \epsilon_i X_i X_i^* \right\|_2 \leq \mathbb{E}_q \left\| \sum_{i=1}^r \epsilon_i X_i X_i^* \right\|_{S_q}.$$

Now we use the noncommutative Khintchine inequality (noting that the two terms in the inequality's $\max$ are equal in this case) to get

$$E \leq W_q \left\| \left(\sum_{i=1}^r X_i X_i^* X_i X_i^* \right)^{1/2} \right\|_{S_q}.$$

We can bound the Schatten q-norm by the spectral norm by paying a multiplicative penalty of $p^{1/q}$, where p is the maximum rank of the matrix sum. By the hypothesis $q \geq 2\log p$, this penalty does not exceed $\sqrt{e}$, resulting in

$$\begin{aligned}E &\leq W_q \sqrt{e} \left\| \left(\sum_{i=1}^r X_i X_i^* X_i X_i^* \right)^{1/2} \right\|_2 \\ &\leq W_q \sqrt{e} \left\| \sum_{i=1}^r X_i X_i^* X_i X_i^* \right\|_2^{1/2}.\end{aligned}$$

Finally, we note that the sum term is a quadratic form that can be expressed in terms of X and its block diagonalization, as follows:

$$\begin{aligned}E &\leq W_q \sqrt{e} \|X D_X^* D_X X^*\|_2^{1/2} \leq W_q \sqrt{e} \|D_X X^*\|_2 \\ &\leq W_q \sqrt{e} \|D_X\|_2 \|X\|_2 \\ &\leq W_q \sqrt{e} \|X\|_{B,1} \|X\|_2,\end{aligned}$$

where the last step used Lemma A.1. Now replace $W_q \leq 2^{-1/4}\sqrt{\pi q/e}$ to complete the proof. ■

The next lemma is a generalization of [25, Proposition 2.1] to our block setting.

Lemma A.3. *Let H be a Hermitian matrix with zero blocks on the diagonal. Then $\mathbb{E}_q \|RHR\|_2 \leq 2\mathbb{E}_q \|RHR'\|_2$, where $R' := \Sigma' \otimes \mathrm{Id}_m$ with Σ' denoting an independent realization of the random matrix Σ.*

Proof: We establish the result for $q = 1$ for simplicity and without loss of generality. Denote by $\widetilde{H}_{i,j}$ the masking of the matrix H that preserves only the subblock $H_{i,j}$ and makes other entries of H zero. Then, we have

$$\mathbb{E}\|RHR\|_2 = \mathbb{E} \left\| \sum_{1\leq i<j\leq r} \zeta_i \zeta_j (\widetilde{H}_{i,j} + \widetilde{H}_{j,i}) \right\|_2.$$

Let η_i be i.i.d. Bernoulli random variables with parameter $1/2$. We then use Jensen's inequality on this new (multivariate) random variable $\eta = \{\eta_i\}_{1\leq i\leq r}$. Specifically, we define $M_{i,j}(\eta) = \eta_i(1-\eta_j) + \eta_j(1-\eta_i)$, and note that $\mathbb{E}_\eta M_{i,j}(\eta) = 1/2$ for all i, j, where $\mathbb{E}_\eta$ denotes expectation over the random variable η (in contrast to the notation $\mathbb{E}_q$, where q is a constant). We also define the function

$$f(M_{i,j}(\eta)) = \mathbb{E}_\zeta \left\| \sum_{1\leq i<j\leq r} 2\zeta_i \zeta_j M_{i,j}(\eta)(\widetilde{H}_{i,j} + \widetilde{H}_{j,i}) \right\|_2.$$

Then, by applying Jensen's inequality to $f(\mathbb{E}_\eta M_{i,j}(\eta)) = \mathbb{E}\|RHR\|_2$, we obtain

$$\mathbb{E}\|RHR\|_2 \leq 2\mathbb{E}_\eta\mathbb{E}_\zeta \Bigg\| \sum_{1\leq i<j\leq r} [\eta_i(1-\eta_j) + \eta_j(1-\eta_i)]\zeta_i\zeta_j(\widetilde{H}_{i,j}+\widetilde{H}_{j,i}) \Bigg\|_2.$$

There is a 0-1 vector η^* for which the expression exceeds its expectation over η. Letting $T = \{i : \eta_i^* = 1\}$, we get

$$\begin{aligned}
\mathbb{E}\|RHR\|_2 &\leq 2\mathbb{E}_\zeta \left\| \sum_{i\in T, j\in T^C} \zeta_i\zeta_j(\widetilde{H}_{i,j}+\widetilde{H}_{j,i}) \right\|_2 \\
&= 2\mathbb{E}_\zeta \left\| \sum_{i\in T, j\in T^C} \zeta_i\zeta_j\widetilde{H}_{i,j} + \sum_{i\in T, j\in T^C} \zeta_i\zeta_j\widetilde{H}_{j,i} \right\|_2 \\
&= 2\mathbb{E}_\zeta \left\| \sum_{i\in T, j\in T^C} \zeta_i\zeta_j\widetilde{H}_{i,j} \right\|_2 \\
&= 2\mathbb{E}_\zeta \left\| \sum_{i\in T, j\in T^C} \zeta_i\zeta_j'\widetilde{H}_{i,j} \right\|_2, \qquad (13)
\end{aligned}$$

where $\{\zeta_j'\}$ is an independent realization of the sequence $\{\zeta_i\}$. The equality in (13) is a combination of the following facts: (i) $\widetilde{H}_{i,j} = \widetilde{H}_{j,i}^*$ and, therefore, defining $A := \sum_{i\in T, j\in T^C} \zeta_i\zeta_j\widetilde{H}_{i,j}$ and $B := \sum_{i\in T, j\in T^C} \zeta_i\zeta_j\widetilde{H}_{j,i}$ we have $B = A^*$; (ii) we can reorder the columns and rows of A and B to have those corresponding to the set T first, followed by those corresponding to the set T^C later, giving us matrices of the form

$$\check{A} = \begin{bmatrix} 0 & \tilde{A} \\ 0 & 0 \end{bmatrix} \text{ and } \check{B} = \begin{bmatrix} 0 & 0 \\ \tilde{B} & 0 \end{bmatrix},$$

respectively, where $\tilde{A} = \tilde{B}^*$; (iii) permuting the columns and rows of a matrix does not affect its spectral norm; (iv) the Hermitian dilation map of a matrix M,

$$\mathcal{D}(M) : M \mapsto \begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix},$$

preserves the spectral norm of M: $\|M\|_2 = \|\mathcal{D}(M)\|_2$; and (v) removal of all-zero rows and columns from a matrix preserves its spectral norm. Combining these five facts together, we have

$$\begin{aligned}
\|A+B\|_2 = \|\check{A}+\check{B}\|_2 &= \left\| \begin{bmatrix} 0 & \tilde{A} \\ \tilde{B} & 0 \end{bmatrix} \right\|_2 = \|\mathcal{D}(\tilde{A})\|_2 \\
&= \|\tilde{A}\|_2 = \|\check{A}\|_2 = \|A\|_2.
\end{aligned}$$

Finally, since the norm of a submatrix does not exceed the norm of the matrix, we re-introduce the missing blocks to complete the argument:

$$\mathbb{E}\|RHR\|_2 \leq 2\mathbb{E}_\zeta \left\| \sum_{1\leq i,j\leq r, i\neq j} \zeta_i\zeta_j'\widetilde{H}_{i,j} \right\|_2 = 2\mathbb{E}_\zeta\|RHR'\|_2.$$

■

The next lemma is adapted to our problem setup of block matrices from [64, Theorem 3.1], [62, Proposition 12].

Lemma A.4. *Let $M = [M_1\ M_2\ \dots\ M_r]$ be a matrix with r column blocks, and suppose $q \geq 4\log p \geq 2$. Then*

$$\mathbb{E}_q\|MR\|_2 \leq 3\sqrt{\frac{q}{2}}\mathbb{E}_q\|MR\|_{B,1} + \sqrt{\delta}\|M\|_2.$$

Proof: We denote $E := \mathbb{E}_q\|MR\|_2$ and note that

$$\begin{aligned}
E^2 = \mathbb{E}_{q/2}\|MRM^*\|_2 &= \mathbb{E}_{q/2}\left\| \sum_{1\leq i\leq r} \zeta_i M_i M_i^* \right\|_2 \\
&\leq \mathbb{E}_{q/2}\left\| \sum_{1\leq i\leq r} (\zeta_i-\delta) M_i M_i^* \right\|_2 + \delta\left\| \sum_{1\leq i\leq p} M_i M_i^* \right\|_2.
\end{aligned}$$

Next, we replace δ by $\delta = \mathbb{E}\zeta_i'$, with $\{\zeta_i'\}$ denoting an independent copy of the sequence $\{\zeta_i\}$. We then take the expectation out of the norm by applying Jensen's inequality to get

$$E^2 \leq \mathbb{E}_{q/2}\left\| \sum_{1\leq i\leq r} (\zeta_i-\zeta_i') M_i M_i^* \right\|_2 + \delta\|MM^*\|_2.$$

We now symmetrize the distribution by introducing a Rademacher sequence $\{\epsilon_i\}$, noticing that the expectation does not change due to the symmetry of the random variables $\zeta_i - \zeta_i'$:

$$E^2 \leq \mathbb{E}_{q/2}\left\| \sum_{1\leq i\leq r} \epsilon_i(\zeta_i-\zeta_i') M_i M_i^* \right\|_2 + \delta\|M\|_2^2.$$

We apply the triangle inequality to separate ζ_i and ζ_i', and by noticing that they have the same distribution, we obtain

$$E^2 \leq 2\mathbb{E}_{q/2}\left\| \sum_{1\leq i\leq r} \epsilon_i\zeta_i M_i M_i^* \right\|_2 + \delta\|M\|_2^2.$$

Writing $\Omega = \{i : \zeta_i = 1\}$, we see that

$$E^2 \leq 2\mathbb{E}_{q/2,\zeta}\left(\mathbb{E}_{q/2,\epsilon}\left\| \sum_{i\in\Omega} \epsilon_i M_i M_i^* \right\|_2 \right) + \delta\|M\|_2^2,$$

where we have split the expectation on the random variables $\{\zeta_i\}$ and $\{\epsilon_i\}$. Now we use Lemma A.2 on the term in parentheses to get

$$E^2 \leq 3\sqrt{\frac{q}{2}}\mathbb{E}_{q/2}(\|MR\|_{B,1}\|MR\|_2) + \delta\|M\|_2^2.$$

Using the Cauchy-Schwarz inequality, we get

$$E^2 \leq 3\sqrt{\frac{q}{2}}\mathbb{E}_q\|MR\|_{B,1}\mathbb{E}_q\|MR\|_2 + \delta\|M\|_2^2.$$

This inequality takes the form $E^2 \leq bE + c$. We bound E by the largest solution of this quadratic form:

$$E \leq \frac{b+\sqrt{b^2+4c}}{2} \leq b + \sqrt{c},$$

thereby proving the lemma. ■

The last lemma that we need for our proof is a generalization of [62, Proposition 13] to our block setting.

Lemma A.5. *Let*

$$M = [M_1\ M_2\ \dots\ M_r] = \begin{bmatrix} M_{1,1} & M_{1,2} & \dots & M_{1.r} \\ M_{2,1} & M_{2,2} & \dots & M_{2.r} \\ \vdots & \vdots & \ddots & \vdots \\ M_{r,1} & M_{r,2} & \dots & M_{r.r} \end{bmatrix}$$

be a block matrix, where each block $M_{i,j}$ has size $m \times m$. Assume $q \geq 2\log r$. Then, we have

$$\mathbb{E}_q \|RM\|_{B,1} \leq 2^{1.5}\sqrt{q}\|M\|_{B,2} + \sqrt{\delta}\|M\|_{B,1}.$$

Proof: We begin by seeing that

$$\begin{aligned} E^2 &:= (\mathbb{E}_q\|RM\|_{B,1})^2 = \left[\mathbb{E}_q\left(\max_{1\leq j\leq r}\|RM_j\|_2\right)\right]^2 \\ &= \mathbb{E}_{q/2}\left(\max_{1\leq j\leq r}\sum_{i=1}^{r}\zeta_i\|M_{i,j}\|_2^2\right) \end{aligned}$$

In the sequel, we abbreviate $t = q/2$ and $y_{i,j} = \|M_{i,j}\|_2^2$. We continue by using the same technique as in the proof of Lemma A.4: we split a term for the mean value of the sequence $\{\zeta_i\}$, then replace the term by $\mathbb{E}\zeta_i'$ — an independent copy of the sequence, then exploit symmetrization by introducing a Rademacher sequence $\{\epsilon_i\}$, and then finish by merging the two terms due to their identical distributions:

$$\begin{aligned} E^2 &\leq \mathbb{E}_t\left(\max_{1\leq j\leq r}\sum_{i=1}^{r}(\zeta_i-\delta)y_{i,j}\right) + \delta\max_{1\leq j\leq r}\sum_{i=1}^{r}\|M_{i,j}\|_2^2 \\ &\leq \mathbb{E}_t\left(\max_{1\leq j\leq r}\sum_{i=1}^{r}(\zeta_i-\zeta_i')y_{i,j}\right) + \delta\|M\|_{B,1}^2 \\ &= \mathbb{E}_t\left(\max_{1\leq j\leq r}\sum_{i=1}^{r}\epsilon_i(\zeta_i-\zeta_i')y_{i,j}\right) + \delta\|M\|_{B,1}^2 \\ &\leq 2\mathbb{E}_t\left(\max_{1\leq j\leq r}\sum_{i=1}^{r}\epsilon_i\zeta_i y_{i,j}\right) + \delta\|M\|_{B,1}^2. \end{aligned}$$

Now we bound the maximum by the sum and separate the expectations on the two sequences:

$$E^2 \leq 2\left(\mathbb{E}_\zeta\sum_{j=1}^{r}\left(\mathbb{E}_{t,\epsilon}\sum_{i=1}^{r}\epsilon_i\zeta_i y_{i,j}\right)^t\right)^{1/t} + \delta\|M\|_{B,1}^2.$$

For the inner term, we can use the scalar Khintchine inequality to obtain

$$E^2 \leq 2C_t\left(\mathbb{E}_\zeta\sum_{j=1}^{r}\left(\sum_{i=1}^{r}\zeta_i y_{i,j}^2\right)^{t/2}\right)^{1/t} + \delta\|M\|_{B,1}^2.$$

We continue by bounding the outer sum by the maximum term times the number of terms:

$$E^2 \leq 2C_t r^{1/t}\left(\mathbb{E}_\zeta\max_{1\leq j\leq r}\left(\sum_{i=1}^{r}\zeta_i y_{i,j}^2\right)^{t/2}\right)^{1/t} + \delta\|M\|_{B,1}^2.$$

Since $t \geq \log r$, it holds that $r^{1/t} \leq e$, which implies that $2C_t r^{1/t} \leq 4\sqrt{t}$. We now use Hölder's inequality inside the sum term $\zeta_i y_{i,j}^2 = y_{i,j}\cdot\zeta_i y_{i,j}$ with $s=\infty$, $q=1$:

$$\begin{aligned} E^2 &\leq 4\sqrt{t}\left(\max_{1\leq i,j\leq r} y_{i,j}\right)^{1/2}\left(\mathbb{E}_\zeta\max_{1\leq j\leq r}\left(\sum_{i=1}^{r}\zeta_i y_{i,j}\right)^{t/2}\right)^{1/t} \\ &\qquad + \delta\|M\|_{B,1}^2 \\ &\leq 4\sqrt{t}\left(\max_{1\leq i,j\leq r} y_{i,j}\right)^{1/2}\left(\mathbb{E}_\zeta\max_{1\leq j\leq r}\left(\sum_{i=1}^{r}\zeta_i y_{i,j}\right)^{t}\right)^{1/2t} \\ &\qquad + \delta\|M\|_{B,1}^2. \end{aligned}$$

Now we recall that $t = q/2$ and $y_{i,j} = \|M_{i,j}\|_2^2$, to get

$$\begin{aligned} E^2 &\leq 2^{1.5}\sqrt{q}\max_{1\leq i,j\leq r}\|M_{i,j}\|_2 \times \\ &\qquad \left(\mathbb{E}_\zeta\max_{1\leq j\leq r}\left(\sum_{i=1}^{r}\zeta_i\|M_{i,j}\|_2^2\right)^{q/2}\right)^{1/q} + \delta\|M\|_{B,1}^2 \\ &\leq 2^{1.5}\sqrt{q}\max_{1\leq i,j\leq r}\|M_{i,j}\|_2\left(\mathbb{E}_{q/2}\max_{1\leq j\leq r}\sum_{i=1}^{r}\zeta_i\|M_{i,j}\|_2^2\right)^{1/2} \\ &\qquad + \delta\|M\|_{B,1}^2 \\ &\leq 2^{1.5}\sqrt{q}\|M\|_{B,2}\mathbb{E}_q\max_{1\leq j\leq r}\|RM_j\|_2 + \delta\|M\|_{B,1}^2 \\ &\leq 2^{1.5}\sqrt{q}\|M\|_{B,2}\mathbb{E}_q\|RM\|_{B,1} + \delta\|M\|_{B,1}^2 \end{aligned}$$

and notice that E has appeared on the right hand side. By following the same argument that ends the proof of Lemma A.4, we complete the proof. ∎

We now have all the required results to prove Lemma 1. Split G into its diagonal blocks D (containing $\Phi_i^*\Phi_i - \mathrm{Id}$, $1 \leq i \leq r$) and off-diagonal blocks H (containing $\Phi_i^*\Phi_j$, $1 \leq i \neq j \leq r$) and apply Lemma A.3:

$$\mathbb{E}_q\|RGR\|_2 \leq 2\mathbb{E}_q\|RHR'\|_2 + \mathbb{E}_q\|RDR\|_2.$$

To estimate the first term, we apply Lemma A.4 twice; once for R, and once for R':

$$\begin{aligned} &\mathbb{E}_q\|RHR'\|_2 \\ &\quad\leq 3\sqrt{\frac{q}{2}}\mathbb{E}_q\|RHR'\|_{B,1} + \sqrt{\delta}\mathbb{E}_q\|HR'\|_2 \\ &\quad\leq 3\sqrt{\frac{q}{2}}\mathbb{E}_q\|RHR'\|_{B,1} + 3\sqrt{\frac{\delta q}{2}}\mathbb{E}_q\|HR'\|_{B,1} + \delta\|H\|_2. \end{aligned}$$

By applying Lemma A.5 on the first term, we obtain

$$\begin{aligned} \mathbb{E}_q\|RHR'\|_2 &\leq 3\sqrt{\frac{q}{2}}\Big[2^{1.5}\sqrt{q}\mathbb{E}_q\|HR'\|_{B,2} \\ &\qquad + \sqrt{\delta}\mathbb{E}_q\|HR'\|_{B,1}\Big] + 3\sqrt{\frac{\delta q}{2}}\mathbb{E}_q\|HR'\|_{B,1} + \delta\|H\|_2. \end{aligned}$$

Since R and R' have the same distribution, we can collect terms to get

$$\begin{aligned} \mathbb{E}_q\|RGR\|_2 &\leq 12q\mathbb{E}_q\|HR\|_{B,2} + 12\sqrt{\frac{\delta q}{2}}\mathbb{E}_q\|HR\|_{B,1} \\ &\qquad + 2\delta\|H\|_2 + \mathbb{E}_q\|RDR\|_2. \end{aligned}$$

Next, in order to bound $\|HR\|_{B,1}$, we use the notation $\Phi_{\{i\}^C} = [\Phi_1 \ \dots \ \Phi_{i-1} \ \Phi_{i+1} \ \dots \ \Phi_r]$; we then have

$$\begin{aligned}\|HR\|_{B,1} \leq \|H\|_{B,1} &= \max_{1\leq i\leq r} \|\Phi_i^*\Phi_{\{i\}^C}\|_2 \leq \max_{1\leq i\leq r} \|\Phi_i^*\Phi\|_2 \\ &\leq \max_{1\leq i\leq r} \|\Phi_i\|_2\|\Phi\|_2 = \sqrt{1+\mu_I}\|\Phi\|_2.\end{aligned}$$

Now we use the facts $\|HR\|_{B,2} \leq \mu_B$, $\|H\|_2 \leq \|G\|_2 + \|D\|_2 = \|\Phi\|_2^2 + \|D\|_2$ and, using Lemma A.1,

$$\mathbb{E}_q\|RDR\|_2 \leq \|D\|_2 = \max_{1\leq i\leq r} \|\Phi_i^*\Phi_i - \mathrm{Id}\|_2 \leq \mu_I$$

to complete the proof of the lemma:

$$\begin{aligned}\mathbb{E}_q\|RGR\|_2 &\leq 12q\mu_B + 12\sqrt{\frac{\delta q(1+\mu_I)}{2}}\|\Phi\|_2 \\ &\quad + 2\delta(\|\Phi\|_2^2 + \mu_I) + \mu_I \\ &\leq 48\mu_B \log p + 17\sqrt{\delta \log p(1+\mu_I)}\|\Phi\|_2 \\ &\quad + 2\delta\|\Phi\|_2^2 + 3\mu_I.\end{aligned}$$

Appendix B
Proof of Theorem 2

In this appendix, we will prove that the minimization (9) successfully recovers a k-block sparse β from $y = \Phi\beta$ with high probability. Mathematically, this is equivalent to showing that $\|\beta\|_{2,1} < \|\beta'\|_{2,1}$ for all $\beta' \neq \beta$ such that $y = \Phi\beta'$. In the following, we will argue that this is true as long as there exists a vector $h \in \mathbb{R}^{km}$ such that (i) $\Phi_{\mathcal{S}}^* h = \overline{\mathrm{sign}}(\beta_{\mathcal{S}})$, where $\mathcal{S}$ denotes the block support of β, $\Phi_{\mathcal{S}}^*$ denotes the adjoint of $\Phi_{\mathcal{S}}$, and $\overline{\mathrm{sign}}(\beta_{\mathcal{S}})$ denotes the block-wise extension of $\overline{\mathrm{sign}}(\cdot)$ to the blocks in $\mathcal{S}$, and (ii) $\|\Phi_j^* h\|_2 < 1$ for all $j \notin \mathcal{S}$. Note that these two conditions on the vector h imply that (iii) $\|\Phi_j^* h\|_2 \leq 1$ for all $1 \leq j \leq r$.

To prove the sufficiency of conditions (i) and (ii) above, we follow the same ideas as in [45, 88, 89]. To begin, we need the following lemma; its proof is a simple exercise using Hölder's Inequality.

Lemma B.1. *Consider two block vectors a and b such that the blocks of a have nonidentical ℓ_2-norms and the blocks of b are nonzero. Then $\langle a, b\rangle < \|a\|_{2,\infty}\|b\|_{2,1}$, where $\|a\|_{2,\infty} := \max_{j=1,\dots,r} \|a_j\|_2$.*

Proof: We can write

$$\begin{aligned}\langle a, b\rangle &= \sum_{j=1}^{r}\sum_{l=1}^{m} a_{jl}b_{jl} = \sum_{j=1}^{r}\langle a_j, b_j\rangle \\ &\leq \sum_{j=1}^{r} \|a_j\|_2\|b_j\|_2 = \langle \overline{a}, \overline{b}\rangle, \qquad (14)\end{aligned}$$

where the vectors $\overline{a} = [\|a_1\|_2 \ \|a_2\|_2 \ \dots \ \|a_r\|_2]$ and $\overline{b} = [\|b_1\|_2 \ \|b_2\|_2 \ \dots \ \|b_r\|_2]$ are defined as ones that contain the ℓ_2-norms of the blocks of a and b, respectively. Using the conditions of [89, Lemma 6], we have

$$\langle a, b\rangle \leq \langle \overline{a}, \overline{b}\rangle < \|\overline{a}\|_\infty\|\overline{b}\|_1 = \|a\|_{2,\infty}\|b\|_{2,1},$$

thereby proving the lemma. ∎

Remark 5. Note that if we remove the requirements on a and b, it can be shown that $\langle a, b\rangle \leq \|a\|_{2,\infty}\|b\|_{2,1}$; that is, the conditions on a and b remove the possibility of equality.

In addition to this lemma, we will also need to use the vector Bernstein inequality from [90, 91].

Lemma B.2. *Let $\{v_k\} \in \mathbb{R}^m$ be a finite sequence of independent random vectors. Suppose that $\mathbb{E}(v_k) = 0$ and $\|v_k\|_2 \leq B$ almost surely, and put $\sigma^2 \geq \sum_k \mathbb{E}\|v_k\|_2^2$. Then for all $0 \leq t \leq \sigma^2/B$,*

$$\mathbb{P}\left(\left\|\sum_k v_k\right\|_2 \geq t\right) \leq e^{-\frac{t^2}{8\sigma^2}+\frac{1}{4}}.$$

We are ready now to formally prove Theorem 2. We begin by writing

$$\begin{aligned}\|\beta\|_{2,1} = \|\beta_{\mathcal{S}}\|_{2,1} &= \sum_{j\in\mathcal{S}} \|\beta_j\|_2 \\ &= \sum_{j\in\mathcal{S}} \frac{\beta_j^*\beta_j}{\|\beta_j\|_2} = \sum_{j\in\mathcal{S}} \left\langle \frac{\beta_j}{\|\beta_j\|_2}, \beta_j \right\rangle \\ &= \sum_{j\in\mathcal{S}} \left\langle \overline{\mathrm{sign}}(\beta_j), \beta_j \right\rangle = \left\langle \overline{\mathrm{sign}}(\beta_{\mathcal{S}}), \beta_{\mathcal{S}} \right\rangle,\end{aligned}$$

Next, we assume that conditions (i) and (ii) (which together imply condition (iii)) are true in our case and consider any $\beta' \neq \beta$ such that $y = \Phi\beta'$. Then since $\overline{\mathrm{sign}}(\beta_{\mathcal{S}}) = \Phi_{\mathcal{S}}^* h$, we have

$$\begin{aligned}\|\beta\|_{2,1} &= \langle \Phi_{\mathcal{S}}^* h, \beta_{\mathcal{S}}\rangle = \langle h, \Phi_{\mathcal{S}}\beta_{\mathcal{S}}\rangle \\ &= \langle h, y\rangle = \langle h, \Phi_{\mathcal{S}'}\beta'_{\mathcal{S}'}\rangle = \langle \Phi_{\mathcal{S}'}^* h, \beta'_{\mathcal{S}'}\rangle,\end{aligned}$$

where $\mathcal{S}'$ denotes the support of a different solution β' as described earlier. We now consider two cases. If not all norms $\|\Phi_j^* h\|_2$ are identical over $j \in \mathcal{S}'$, then we apply Lemma B.1 to obtain

$$\|\beta\|_{2,1} < \|\Phi_{\mathcal{S}'}^* h\|_{2,\infty}\|\beta'_{\mathcal{S}'}\|_{2,1} \leq \|\beta'_{\mathcal{S}'}\|_{2,1} = \|\beta'\|_{2,1},$$

where the last inequality is due to condition (iii). If all the norms $\|\Phi_j^* h\|_2$ are identical over $j \in \mathcal{S}'$, note that since $\beta \neq \beta'$ and since Theorem 1 guarantees that $\Phi_{\mathcal{S}}$ has linearly independent columns with high probability (noting that we will come back to Theorem 1 later), then there must exist a block index $j_0 \in \mathcal{S}'$ such that $j_0 \notin \mathcal{S}$. From condition (ii), we know that for such a j_0 we have $\|\Phi_{j_0}^* h\|_2 < 1$, meaning that $\|\Phi_j^* h\|_2 < 1$ for all $j \in \mathcal{S}'$. We then leverage (14) to obtain

$$\begin{aligned}\|\beta\|_{2,1} &\leq \sum_{j\in\mathcal{S}'} \|\Phi_j^* h\|_2\|\beta'_j\|_2 = \|\Phi_{j_0}^* h\|_2 \sum_{j\in\mathcal{S}'} \|\beta'_j\|_2 \\ &< \|\beta'_{\mathcal{S}'}\|_{2,1} = \|\beta'\|_{2,1}.\end{aligned}$$

In order to complete the proof of the theorem, the only thing that remains to be shown now is that conditions (i) and (ii) hold in our case.

To simplify conditions (i) and (ii), we can define the vector $h = (\Phi_{\mathcal{S}}^\dagger)^*\overline{\mathrm{sign}}(\beta_{\mathcal{S}})$, where $(\cdot)^\dagger$ denotes the Moore–Penrose inverse of a matrix. Note that such an h trivially satisfies condition (i). Condition (ii) then reduces to

$$\|\Phi_{\mathcal{S}^C}^*(\Phi_{\mathcal{S}}^\dagger)^*\overline{\mathrm{sign}}(\beta_{\mathcal{S}})\|_{2,\infty} < 1. \qquad (15)$$

It remains to prove (15). Denote the vector

$$Z_{0,i} = \Phi_i^* \Phi_{\mathcal{S}} (\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \overline{\text{sign}}(\beta_{\mathcal{S}})$$

for each $i \in \mathcal{S}^C$. Further denote

$$Z_0 = \max_{i \notin \mathcal{S}} \|Z_{0,i}\|_2 = \|\Phi_{\mathcal{S}^C}^* \Phi_{\mathcal{S}} (\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \overline{\text{sign}}(\beta_{\mathcal{S}})\|_{2,\infty}.$$

We then simply need to show that with large probability $Z_0 < 1$. Define the matrix $W_i = (\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \Phi_{\mathcal{S}}^* \Phi_i$ for $i \notin \mathcal{S}$. Further, denote by W_i^j the submatrix of W_i containing the block of rows that corresponds to $j \in \mathcal{S}$. We can then write $Z_{0,i} = \sum_{j \in \mathcal{S}} W_i^{j*} \overline{\text{sign}}(\beta_j)$, where W_i^{j*} is the adjoint of W_i^j. The sum terms have norms bounded by

$$\left\| W_i^{j*} \overline{\text{sign}}(\beta_j) \right\|_2 \leq \left\| W_i^j \right\|_2 \left\| \overline{\text{sign}}(\beta_j) \right\|_2 = \left\| W_i^j \right\|_2.$$

We now use Lemma B.2 by setting $B = \max_{j \in \mathcal{S}} \|W_i^j\|_2$ and $\sigma^2 = \sum_{j \in \mathcal{S}} \|W_i^j\|_2^2 = \|W_i\|_2^2$ to obtain

$$\mathbb{P}(|Z_{0,i}| \geq t) \leq 2e^{-t^2/8 \max_{i \notin \mathcal{S}} \|W_i\|_2^2}$$

for $0 \leq t \leq 1$ (as $\sigma^2 > B$). A union bound then gives us $\mathbb{P}(Z_0 \geq t) \leq 2pe^{-t^2/8\kappa^2}$, where $\kappa > \max_{i \notin \mathcal{S}} \|W_i\|_2$. We can also see from the statement of Theorem 1 concerning the conditioning of random block subdictionaries that

$$\begin{aligned} \max_{i \notin \mathcal{S}} \|W_i\|_2 &= \max_{i \notin \mathcal{S}} \|(\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \Phi_{\mathcal{S}}^* \Phi_i\|_2 \\ &\leq 2 \max_{i \notin \mathcal{S}} \|\Phi_{\mathcal{S}}^* \Phi_i\|_2 = 2\|\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}^C}\|_{B,1} \end{aligned}$$

with probability at least $1 - 2p^{-4 \log 2}$. Thus, conditioning on the probability events

$$\gamma \geq \|\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}^C}\|_{B,1} \tag{16}$$

and $\|(\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1}\|_2 \leq 2$ (which also accounts for the use of Theorem 1 earlier in the proof), and replacing $t = 1$, the probability that $Z_0 \geq 1$ is at most $2pe^{-1/32\gamma^2}$. By seeing that $\|\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}^C}\|_{B,1} \leq \mu_B$, the BIC allows us to set $\gamma = c_2/\sqrt{\log(p)}$ so that this probability of failure is upper bounded by $2p^{-4 \log 2}$ for sufficiently small values of c_2. The proof of the theorem now follows by taking a final union bound over the events $\|(\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1}\|_2 > 2$ and $Z_0 \geq 1$.

Appendix C
Proof of Theorem 5

To begin, we need the following lemma concerning the behavior of the group lasso.

Lemma C.1. *The group lasso estimate $\widehat{\beta}$ satisfies the inequality $\|\Phi^*(y - \Phi\widehat{\beta})\|_{2,\infty} \leq 2\lambda\sigma\sqrt{m}$.*

Proof: Since $\widehat{\beta}$ minimizes the objective function over β, then 0 must be a subgradient of the objective function at $\widehat{\beta}$. The subgradients of the group lasso objective function are of the form [31]

$$\Phi_i^*(\Phi\beta - y) + 2\lambda\sigma\sqrt{m}\epsilon_i = 0, \ i = 1, \ldots, r,$$

where $\epsilon_i \in \mathbb{R}^m$ is given by $\epsilon_i = \overline{\text{sign}}(\beta_i)$ if $\beta_i \neq 0$ and $\|\epsilon_i\|_2 \leq 1$ otherwise. Hence, since 0 is a subgradient at $\widehat{\beta}$, there exists $\epsilon = [\epsilon_1^* \ \ldots \ \epsilon_r^*]^*$ such that

$$\Phi^*(\Phi\widehat{\beta} - y) = -2\lambda\sigma\sqrt{m}\epsilon.$$

The conclusion follows from the fact that $\|\epsilon\|_{2,\infty} \leq 1$. ∎

Similar to the proof in [30] for linear regression in the non-block setting, the proof of Theorem 5 will rely on three conditions involving the design matrix Φ, the vector of regression coefficients β, and the modeling error z. We once again use $\mathcal{S}$ to denote the block support of the k-block sparse β and use $\Phi_{\mathcal{S}}$ to denote the matrix containing columns of the blocks indexed by $\mathcal{S}$, i.e., an $n \times km$ submatrix of Φ. Additionally, note that $\Phi_{\mathcal{S}}^*$ denotes the adjoint of $\Phi_{\mathcal{S}}$ rather than a column submatrix of Φ^*. Finally, we assume in this appendix that $\sigma = 1$ without loss of generality. First, we pose the following three conditions, which we will show are sufficient for the theorem statement to hold:

- *Invertibility condition:* The submatrix $\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}}$ is invertible and obeys $\|(\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1}\|_2 \leq 2$.
- *Orthogonality condition:* The vector z satisfies the following inequality: $\|\Phi^* z\|_{2,\infty} \leq \sqrt{2m} \cdot \lambda$.
- *Complementary size condition:* The following inequality holds:

$$\begin{aligned} &\|\Phi_{\mathcal{S}^C}^* \Phi_{\mathcal{S}} (\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \Phi_{\mathcal{S}}^* z\|_{2,\infty} \\ &\qquad + 2\lambda\sqrt{m}\|\Phi_{\mathcal{S}^C}^* \Phi_{\mathcal{S}} (\Phi_{\mathcal{S}}^* \Phi_{\mathcal{S}})^{-1} \overline{\text{sign}}(\beta_{\mathcal{S}})\|_{2,\infty} \\ &\leq (2 - \sqrt{2})\lambda\sqrt{m}. \end{aligned}$$

We assume that these three conditions hold. Since $\widehat{\beta}$ minimizes the group lasso objective function, we must have

$$\frac{1}{2}\|y - \Phi\widehat{\beta}\|_2^2 + 2\lambda\sqrt{m}\|\widehat{\beta}\|_{2,1} \leq \frac{1}{2}\|y - \Phi\beta\|_2^2 + 2\lambda\sqrt{m}\|\beta\|_{2,1}.$$

Define $h := \widehat{\beta} - \beta$, and note that

$$\begin{aligned} \|y - \Phi\widehat{\beta}\|_2^2 &= \|(y - \Phi\beta) - \Phi h\|_2^2 \\ &= \|\Phi h\|_2^2 + \|y - \Phi\beta\|_2^2 - 2\langle \Phi h, y - \Phi\beta \rangle. \end{aligned}$$

Plugging this identity with $z = y - \Phi\beta$ into the above inequality and rearranging the terms gives

$$\frac{1}{2}\|\Phi h\|_2^2 \leq \langle \Phi h, z \rangle + 2\lambda\sqrt{m}(\|\beta\|_{2,1} - \|\widehat{\beta}\|_{2,1}).$$

Next, break up $\widehat{\beta}$ into $\widehat{\beta}_{\mathcal{S}}$ and $\widehat{\beta}_{\mathcal{S}^C} = h_{\mathcal{S}^C}$ and rewrite the above equation as

$$\begin{aligned} \frac{1}{2}\|\Phi h\|_2^2 &\leq \langle h, \Phi^* z \rangle \\ &\quad + 2\lambda\sqrt{m}(\|\beta_{\mathcal{S}}\|_{2,1} - \|\widehat{\beta}_{\mathcal{S}}\|_{2,1} - \|h_{\mathcal{S}^C}\|_{2,1}). \end{aligned} \tag{17}$$

For each $i \in \mathcal{S}$, we have

$$\begin{aligned} \|\widehat{\beta}_i\|_2 = \|\beta_i + h_i\|_2 &\geq \left| \left\langle \beta_i + h_i, \frac{\beta_i}{\|\beta_i\|_2} \right\rangle \right| \\ &\geq \left\langle \beta_i + h_i, \frac{\beta_i}{\|\beta_i\|_2} \right\rangle = \|\beta_i\|_2 + \langle h_i, \overline{\text{sign}}(\beta_i) \rangle, \end{aligned}$$

where the first inequality is due to the projection of $\beta_i + h_i$ on $\text{span}\{\beta_i\}$ having magnitude at most $\|\beta_i + h_i\|_2$, and the second inequality is due to $|x| \geq x$ for all x. Thus, we can write $\|\widehat{\beta}_{\mathcal{S}}\|_{2,1} \geq \|\beta_{\mathcal{S}}\|_{2,1} + \langle h_{\mathcal{S}}, \overline{\text{sign}}(\beta_{\mathcal{S}}) \rangle$. Merging this inequality with (17) gives us

$$\begin{aligned} \frac{1}{2}\|\Phi h\|_2^2 &\leq \langle h, \Phi^* z \rangle + 2\lambda\sqrt{m}(-\langle h_{\mathcal{S}}, \overline{\text{sign}}(\beta_{\mathcal{S}}) \rangle - \|h_{\mathcal{S}^C}\|_{2,1}) \\ &= \langle h_{\mathcal{S}}, \Phi_{\mathcal{S}}^* z \rangle + \langle h_{\mathcal{S}^C}, \Phi_{\mathcal{S}^C}^* z \rangle \\ &\quad - 2\lambda\sqrt{m}(\langle h_{\mathcal{S}}, \overline{\text{sign}}(\beta_{\mathcal{S}}) \rangle + \|h_{\mathcal{S}^C}\|_{2,1}). \end{aligned} \tag{18}$$

The orthogonality condition and Lemma B.1 also imply

$$\langle h_{\mathcal{S}^C}, \Phi_{\mathcal{S}^C}^* z\rangle \leq \|h_{\mathcal{S}^C}\|_{2,1}\|\Phi_{\mathcal{S}^C}^* z\|_{2,\infty} \leq \sqrt{2m}\cdot\lambda\|h_{\mathcal{S}^C}\|_{2,1}.$$

Merging this result with (18) results in

$$\begin{aligned}\frac{1}{2}\|\Phi h\|_2^2 &\leq \langle h_{\mathcal{S}}, v\rangle - (2-\sqrt{2})\lambda\sqrt{m}\|h_{\mathcal{S}^C}\|_{2,1},\\ &\leq |\langle h_{\mathcal{S}}, v\rangle| - (2-\sqrt{2})\lambda\sqrt{m}\|h_{\mathcal{S}^C}\|_{2,1}, \qquad (19)\end{aligned}$$

where $v = \Phi_{\mathcal{S}}^* z - 2\lambda\sqrt{m}\cdot\overline{\mathrm{sign}}(\beta_{\mathcal{S}})$. We aim to bound each of the terms on the right hand side independently. For the first term, we have

$$\begin{aligned}|\langle h_{\mathcal{S}}, v\rangle| &= |\langle (\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}} h_{\mathcal{S}}, v\rangle|\\ &= |\langle \Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}} h_{\mathcal{S}}, (\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\rangle|\\ &\leq |\langle \Phi_{\mathcal{S}}^*\Phi h, (\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\rangle|\\ &\qquad + |\langle \Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C} h_{\mathcal{S}^C}, (\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\rangle|.\end{aligned}$$

Denote the two terms on the right hand side as A_1 and A_2, respectively. For A_1 we use Lemma B.1 to obtain

$$A_1 \leq \|(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\|_{2,1}\|\Phi_{\mathcal{S}}^*\Phi h\|_{2,\infty}.$$

Now we bound these two terms. For the first term, we get

$$\begin{aligned}\|(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\|_{2,1} &\leq \sqrt{k}\|(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\|_2\\ &\leq \sqrt{k}\|(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\|_2\|v\|_2 \leq 2k\|v\|_{2,\infty}\end{aligned}$$

due to the invertibility condition. Using the orthogonality condition, we also get

$$\begin{aligned}\|v\|_{2,\infty} &= \|\Phi_{\mathcal{S}}^* z - 2\lambda\sqrt{m}\cdot\overline{\mathrm{sign}}(\beta_{\mathcal{S}})\|_{2,\infty}\\ &\leq \|\Phi_{\mathcal{S}}^* z\|_{2,\infty} + 2\lambda\sqrt{m} \leq (2+\sqrt{2})\lambda\sqrt{m}.\end{aligned}$$

For the second term $\|\Phi_{\mathcal{S}}^*\Phi h\|_{2,\infty}$, we use Lemma C.1 and the orthogonality condition to get

$$\begin{aligned}\|\Phi_{\mathcal{S}}^*\Phi h\|_{2,\infty} &\leq \|\Phi_{\mathcal{S}}^*(\Phi\beta - y)\|_{2,\infty} + \|\Phi_{\mathcal{S}}^*(y - \Phi\widehat{\beta})\|_{2,\infty}\\ &= \|\Phi_{\mathcal{S}}^* z\|_{2,\infty} + \|\Phi_{\mathcal{S}}^*(y - \Phi\widehat{\beta})\|_{2,\infty}\\ &\leq (2+\sqrt{2})\lambda\sqrt{m}.\end{aligned}$$

Combining, we finally get $A_1 \leq 2(2+\sqrt{2})^2\lambda^2 mk$. For A_2, we have from Lemma B.1 that

$$\begin{aligned}A_2 &\leq \|h_{\mathcal{S}^C}\|_{2,1}\|\Phi_{\mathcal{S}^C}^*\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}v\|_{2,\infty}\\ &\leq (2-\sqrt{2})\lambda\sqrt{m}\|h_{\mathcal{S}^C}\|_{2,1},\end{aligned}$$

because of the complementary size condition. Using now these bounds on A_1, A_2, we have

$$|\langle h_{\mathcal{S}}, v\rangle| \leq 2(2+\sqrt{2})^2\lambda^2 mk + (2-\sqrt{2})\lambda\sqrt{m}\|h_{\mathcal{S}^C}\|_{2,1}.$$

Plugging this into (19) gives

$$\frac{1}{2}\|\Phi(\beta - \widehat{\beta})\|_2^2 \leq 2(2+\sqrt{2})^2\lambda^2 mk,$$

which suffices to prove the theorem, modulo the three conditions.

To finish the proof of the theorem, we now must evaluate the probability of each condition failing to hold under the assumed statistical model. The invertibility condition in this regard simply follows from Theorem 1, which means that it fails to hold with probability at most $2p^{-4\log 2}$. Next, note that $\|\Phi^* z\|_{2,\infty} \leq \sqrt{2}\cdot\lambda\sqrt{m}$ is implied by $\|\Phi^* z\|_\infty \leq \sqrt{2}\cdot\lambda$. Further, it is shown in [30] that $\|\Phi^* z\|_\infty > \sqrt{2}\cdot\lambda$ with probability at most $p^{-1}(2\pi\log p)^{-1/2}$. This implies that the orthogonality condition fails to holds with probability at most $p^{-1}(2\pi\log p)^{-1/2}$ in the case of the group lasso. Therefore, we only need to evaluate the complementary size condition.

In order to study the complementary size condition, we partition it into two statements:

$$Z_0 := \|\Phi_{\mathcal{S}^C}^*\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\overline{\mathrm{sign}}(\beta_{\mathcal{S}})\|_{2,\infty} \leq \frac{1}{4}, \qquad (20)$$

$$\begin{aligned}Z_1 &:= \|\Phi_{\mathcal{S}^C}^*\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\Phi_{\mathcal{S}}^* z\|_{2,\infty}\\ &\leq \left(\frac{3}{2} - \sqrt{2}\right)\lambda\sqrt{m}. \qquad (21)\end{aligned}$$

In order to evaluate (20), we compare it to (15) and note that the only difference between the two expressions is a change from $1/4$ to 1. Given that both Theorem 2 and this theorem operate under the same statistical model, it is therefore straightforward to argue from the analysis of (15) that (20) holds except with probability at most $2pe^{-1/512\gamma^2}$, where γ is defined as any positive scalar that satisfies $\gamma \geq \|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1}$. The second condition (21) is implied by the inequality

$$Z_2 := \|\Phi_{\mathcal{S}^C}^*\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\Phi_{\mathcal{S}}^* z\|_\infty \leq \left(\frac{3}{2} - \sqrt{2}\right)\lambda. \qquad (22)$$

In order to establish (22), we use the following result that is a simple consequence of the Chernoff bound on the tail probability of the Gaussian distribution [6] and the union bound (see, e.g., [30, Lemma 3.3]).

Lemma C.2. *Let $(W_j')_{j\in J}$ be a fixed collection of vectors in $\mathbb{R}^n$ and set $Z_2 = \max_{j\in J}|\langle W_j', z\rangle|$. We then have $\mathbb{P}(Z_2 > t) \leq 2|J|e^{-t^2/2(\kappa')^2}$ for any $\kappa' \geq \max_{j\in J}\|W_j'\|_2$.*

We now denote $W_{ij}' = \Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\Phi_{\mathcal{S}}^*\Phi_{ij}$ for $i\notin\mathcal{S}$, $1\leq j\leq m$, where Φ_{ij} represents the j^{th} column of the i^{th} block Φ_i. Then we can write

$$Z_2 = \|\Phi_{\mathcal{S}^C}^*\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}z\|_\infty = \max_{i\notin\mathcal{S}, 1\leq j\leq m}|\langle W_{ij}', z\rangle|.$$

To use Lemma C.2 in this case, we assume that the invertibility condition holds, $\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} \leq \gamma$, and search for a bound on κ':

$$\begin{aligned}\kappa' &= \max_{i\notin\mathcal{S}, 1\leq j\leq m}\|\Phi_{\mathcal{S}}(\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}})^{-1}\Phi_{\mathcal{S}}^*\Phi_{ij}\|_2\\ &\leq \sqrt{6}\max_{i\notin\mathcal{S}, 1\leq j\leq m}\|\Phi_{\mathcal{S}}^*\Phi_{ij}\|_2\\ &\leq \sqrt{6}\max_{i\notin\mathcal{S}}\|\Phi_{\mathcal{S}}^*\Phi_i\|_2 = \sqrt{6}\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} \leq \sqrt{6}\gamma.\end{aligned}$$

Thus, we have that conditioned on the bound $\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} \leq \gamma$ and the invertibility condition, and replacing $t = (3/2 - \sqrt{2})\lambda$, the inequality (22) holds except with probability at most $2pe^{-(3/2-\sqrt{2})^2\lambda^2/12\gamma^2}$.

To conclude, we define $Z = \|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}} - \mathrm{Id}\|_2$ and define the event

$$E = \{Z \leq 1/2\}\cup\{\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} \leq \gamma\}.$$

Then we have that the probability P of the complementary size condition not being met is upper bounded by

$$\begin{aligned} P &\leq \mathbb{P}(\{Z_0 > 1/4\} \cup \{Z_1 \geq (3/2-\sqrt{2})\lambda\}|E) + \mathbb{P}(E^C) \\ &\leq 2pe^{-1/512\gamma^2} + 2pe^{-(3/2-\sqrt{2})^2\lambda^2/12\gamma^2} + \mathbb{P}(Z > 1/2) \\ &\qquad + \mathbb{P}(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} > \gamma) \\ &\leq 2pe^{-1/512\gamma^2} + 2pe^{-(3/2-\sqrt{2})^2\lambda^2/12\gamma^2} + 2p^{-4\log 2} \\ &\qquad + \mathbb{P}(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} > \gamma). \end{aligned}$$

We set $\gamma = c_3/\sqrt{\log p}$ for small enough c_3 so that each of the first two terms of the right hand side is upper bounded by $2p^{-4\log 2}$. In order to bound the last term, we appeal to Lemma A.5 together with the Markov inequality and a Poissonization argument (see (2) and (4) for an example) to obtain

$$\begin{aligned} &\mathbb{P}(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} > \gamma) \\ &\qquad \leq 2\gamma^{-q}\mathbb{E}(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1}^q) \\ &\qquad \leq 2\gamma^{-q}(2^{2.5}\sqrt{q}\mu_B + 2\sqrt{\delta(1+\mu_I)}\|\Phi\|_2)^q. \quad (23) \end{aligned}$$

Then replacing $\gamma = c_3/\sqrt{\log p}$ and $q = 4\log p$ as well as the bounds on k, μ_I and μ_B from the theorem and the BIC in (23), we obtain

$$\begin{aligned} &\mathbb{P}\left(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} > \frac{c_3}{\sqrt{\log p}}\right) \\ &\qquad \leq 2\left(\frac{8\sqrt{2}c_2}{c_3} + \frac{2\sqrt{c_0(1+c_1)}}{c_3}\right)^{4\log p}. \end{aligned}$$

By picking the constants c_0, c_1, c_2 small enough so that the base of the exponential term on the right hand side is less than $1/2$, we get $\mathbb{P}(\|\Phi_{\mathcal{S}}^*\Phi_{\mathcal{S}^C}\|_{B,1} > c_3/\sqrt{\log p}) < 2p^{-4\log 2}$. Thus, the complementary size condition fails to hold with probability at most $8p^{-4\log 2}$.

By combining the failures of the three conditions (and noting that the third condition already accounts for the first one), we have that Theorem 5 holds with probability at least $1 - 8p^{-4\log 2} - p^{-1}(2\pi\log p)^{-1/2}$.

REFERENCES

[1] W. U. Bajwa, R. Calderbank, and M. F. Duarte, "On the conditioning of random block subdictionaries," Duke University, Department of Computer Science, Durham, NC, Tech. Rep. TR-2010-06, Sept. 2010.

[2] M. F. Duarte, W. U. Bajwa, and R. Calderbank, "The performance of group lasso for linear regression of grouped variables," Duke University, Dept. Computer Science, Durham, NC, Technical Report TR-2010-10, Feb. 2011.

[3] ——, "Regression performance of group lasso for arbitrary design matrices," in *Proc. Int. Conf. Sampling Theory and its Applications (SampTA)*, Singapore, June 2011.

[4] W. U. Bajwa, M. F. Duarte, and R. Calderbank, "Average case analysis of high-dimensional block-sparse recovery and regression for arbitrary designs," in *Proc. 17th Intl. Conf. Artificial Intelligence and Statistics (AISTATS'14)*, Reykjavik, Iceland, Apr. 2014, pp. 57–67.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall, 1993.

[6] ——, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ: Prentice Hall, 1998.

[7] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2008.

[8] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *Ann. Statist.*, vol. 28, pp. 1356–1378, 2000.

[9] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.

[10] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Proc. AMS Conf. Math Challenges of the 21st Century*, Los Angeles, CA, Aug. 2000. [Online]. Available: http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint." *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[12] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, Sept. 2006.

[13] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematicians*, vol. 3, Madrid, Spain, 2006, pp. 1433–1452.

[14] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[15] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," in *Compte Rendus de l'Academie des Sciences, Paris, Series I*, vol. 346, 2008, pp. 589–592.

[16] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2563, Nov. 2006.

[17] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[18] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.

[19] A. Cohen, W. Dahmen, and R. A. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 211–231, Jan. 2009.

[20] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3448–3450, June 2013.

[21] W. U. Bajwa and A. Pezeshki, "Finite frames for sparse signal processing," in *Finite Frames*, P. Casazza and G. Kutyniok, Eds. Cambridge, MA: Birkhuser Boston, 2012, ch. 10, pp. 303–335.

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

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

[24] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 1–24, 2008.

[25] ——, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, no. 23–24, pp. 1271–1274, 2008.

[26] P. Kuppinger, G. Durisi, and H. Bölcskei, "Uncertainty relations and sparse signal recovery for pairs of general signal sets," *IEEE Trans. Inform. Theory*, vol. 58, no. 1, pp. 263–277, Jan. 2012.

[27] S. Gurevich and R. Hadani, "The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries," Mar. 2009, unpublished manuscript. [Online]. Available: http://arxiv.org/abs/0903.3627

[28] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2275–2284, Jun. 2009.

[29] G. E. Pfander and H. Rauhut, "Sparsity in time-frequency representations," *J. Fourier Anal. Appl.*, vol. 16, pp. 233–260, 2010.

[30] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.

[31] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statist. Soc. B*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[32] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Machine Learning Research*, vol. 9, no. 6, pp. 1179–1225, June 2008.

[33] M. Mishali and Y. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Proc.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

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

[35] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[36] S. Cotter, B. Rao, E. Kjersti, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005.

[37] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Apr. 2006.

[38] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, Apr. 2006.

[39] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 655–687, 2008.

[40] Y. Nardi and A. Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Statistics*, vol. 2, pp. 605–633, 2008.

[41] L. Meier, S. van de Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Royal Statist. Soc. B*, vol. 70, no. 1, pp. 53–71, Jan. 2008.

[42] H. Liu and J. Zhang, "Estimation consistency of the group lasso and its applications," in *Proc. Int. Conf. Artificiall Intelligence and Statistics (AISTATS)*, Clearwater Beach, FL, Apr. 2009, pp. 376–383.

[43] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[44] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[45] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Info. Theory*, vol. 6, no. 1, pp. 505–519, Jan. 2010.

[46] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[47] M. Stojnic, "ℓ_2/ℓ_1-optimization in block-sparse compressed sensing and its strong thresholds," *IEEE J. Select. Top. Signal Processing*, vol. 4, no. 2, pp. 350–357, Apr. 2010.

[48] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[49] J. Huang and T. Zhang, "The benefit of group sparsity," *Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, Aug. 2010.

[50] P. T. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," *IEEE Trans. Info. Theory*, vol. 57, no. 6, pp. 3864–3876, June 2011.

[51] J. Fang and H. Li, "Recovery of block-sparse representations from noisy observations via orthogonal matching pursuit," Sep. 2011, unpublished manuscript. [Online]. Available: http://arxiv.org/abs/1109.5430

[52] Z. Ben-Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE J. Select. Top. Signal Processing*, vol. 5, no. 5, pp. 1032–1047, Sep. 2011.

[53] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inform. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.

[54] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[55] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Annals of Statistics*, vol. 39, no. 1, pp. 1–47, Jan. 2011.

[56] M. Kolar, J. Lafferty, and L. Wasserman, "Union support recovery in multi-task learning," *J. Machine Learning Res.*, vol. 12, no. 7, pp. 2415–2435, July 2011.

[57] Z. Fang, "Sparse group selection through co-adaptive penalties," Nov. 2011, unpublished manuscript. [Online]. Available: http://arxiv.org/abs/1111.4416

[58] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *Ann. Statist.*, vol. 39, no. 4, pp. 2164–2204, 2011.

[59] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inform. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[60] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[61] M. F. Duarte, M. B. Wakin, D. Baron, S. Sarvotham, and R. G. Baraniuk, "Measurement bounds for sparse signal ensembles via graphical models," *IEEE Trans. Info. Theory*, vol. 59, no. 7, pp. 4280–4289, July 2013.

[62] J. A. Tropp, "The random paving property for uniformly bounded matrices," *Studia Math.*, vol. 185, no. 1, pp. 67–82, 2008.

[63] M. Rudelson, "Random vectors in the isotropic position," *J. Functional Anal.*, vol. 164, no. 1, pp. 60–72, May 1999.

[64] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *J. Assoc. Comput. Mach.*, vol. 54, no. 4, pp. 1–19, 2007.

[65] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Proc.*, vol. 21, no. 2, pp. 494–504, Feb. 2012.

[66] M. Fornasier and H. Rauhut, "Recovery algorithms for vector valued data with joint sparsity constraints," *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[67] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Processing*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.

[68] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[69] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., 2007, pp. 41–48.

[70] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statistics and Computing*, vol. 20, no. 2, pp. 231–252, 2010.

[71] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 4, pp. 1872–1882, Apr. 2009.

[72] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices and Systems*, vol. 5, no. 1, pp. 8–20, Jan. 2011.

[73] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Proc.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[74] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Proc.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[75] J. W. Phillips, R. M. Leahy, and J. C. Mosher, "MEG-based imaging of focal neuronal current sources," *IEEE Trans. Medical Imaging*, vol. 16, no. 3, pp. 338–348, June 1997.

[76] W. Ou, M. S. Hämäläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver," *Neuroimage*, vol. 44, no. 3, pp. 932–946, Feb. 2009.

[77] S. Chen, D. L. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, 1998.

[78] W. U. Bajwa, R. Calderbank, and D. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Appl. Comput. Harmon. Anal.*, vol. 33, no. 1, pp. 58–78, July 2012.

[79] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, Aug. 2008.

[80] D. L. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *Proc. Natl. Acad. Sci.*, vol. 102, no. 27, pp. 9452–9457, July 2005.

[81] ——, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, Jan. 2009.

[82] N. Rao, B. Recht, and R. Nowak, "Universal measurement bounds for structured sparse signal recovery," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, La Palma, Spain, Apr. 2012, pp. 942–950.

[83] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

[84] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc B*, vol. 58, no. 1, pp. 267–288, 1996.

[85] W. U. Bajwa, M. F. Duarte, and R. Calderbank, "Conditioning of random block subdictionaries with applications to block-sparse recovery and regression," Oct. 2014, extended online version. [Online]. Available: http://arxiv.org/abs/1309.5310v3

[86] F. R. Bach, "Structured sparsity-inducing norms through submodular functions," in *Neural Information Processing Systems (NIPS)*, Vancouver, BC, Dec. 2010, pp. 118–126.

[87] S. Wright, R. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2479–2493, July 2009.

[88] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, June 2004.

[89] J. A. Tropp, "Recovery of short complex linear combinations via ℓ_1 minimization," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1568–1570, Apr. 2005.
[90] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.
[91] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Info. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

Waheed U. Bajwa (S'98–M'09–SM'13) received BE (with Honors) degree in electrical engineering from the National University of Sciences and Technology, Pakistan in 2001 and MS and PhD degrees in electrical engineering from the University of Wisconsin-Madison in 2005 and 2009, respectively. He was a Postdoctoral Research Associate in the Program in Applied and Computational Mathematics at Princeton University from 2009 to 2010, and a Research Scientist in the Department of Electrical and Computer Engineering at Duke University from 2010 to 2011. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Rutgers University. His research interests include high-dimensional inference and inverse problems, sampling theory, statistical signal processing, computational harmonic analysis, machine learning, wireless communications, and applications in biological sciences, complex networked systems, and radar & image processing.

Dr. Bajwa has more than three years of industry experience, including a summer position at GE Global Research, Niskayuna, NY. He received the Best in Academics Gold Medal and President's Gold Medal in Electrical Engineering from the National University of Sciences and Technology (NUST) in 2001, the Morgridge Distinguished Graduate Fellowship from the University of Wisconsin-Madison in 2003, the Army Research Office Young Investigator Award in 2014, and the National Science Foundation CAREER Award in 2015. He co-guest edited a special issue of Elsevier Physical Communication Journal on "Compressive Sensing in Communications" (2012), co-organized 1st CPS Week Workshop on Signal Processing Advances in Sensor Networks (2013), and co-chaired IEEE GlobalSIP Symposium on New Sensing and Statistical Inference Methods (2013). He is currently an Associate Editor of the IEEE Signal Processing Letters, Publicity & Publications Chair of IEEE CAMSAP 2015, and a Senior Member of the IEEE.

Marco F. Duarte (S'99–M'09–SM'14) received the B.Sc. degree in computer engineering (with distinction) and the M.Sc. degree in electrical engineering from the University of Wisconsin-Madison in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from Rice University, Houston, TX, in 2009. He was an NSF/IPAM Mathematical Sciences Postdoctoral Research Fellow in the Program of Applied and Computational Mathematics at Princeton University, Princeton, NJ, from 2009 to 2010, and in the Department of Computer Science at Duke University, Durham, NC, from 2010 to 2011. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Massachusetts, Amherst, MA. His research interests include machine learning, compressed sensing, sensor networks, and computational imaging.

Dr. Duarte received the Presidential Fellowship and the Texas Instruments Distinguished Fellowship in 2004 and the Hershel M. Rich Invention Award in 2007, all from Rice University. He is also a member of Tau Beta Pi.

Robert Calderbank (M'89–SM'97–F'98) received the BSc degree in 1975 from Warwick University, England, the MSc degree in 1976 from Oxford University, England, and the PhD degree in 1980 from the California Institute of Technology, all in mathematics.

Dr. Calderbank is Professor of Electrical Engineering at Duke University where he now directs the Information Initiative at Duke (iiD) after serving as Dean of Natural Sciences (2010-2013). Dr. Calderbank was previously Professor of Electrical Engineering and Mathematics at Princeton University where he directed the Program in Applied and Computational Mathematics. Prior to joining Princeton in 2004, he was Vice President for Research at AT&T, responsible for directing the first industrial research lab in the world where the primary focus is data at scale. At the start of his career at Bell Labs, innovations by Dr. Calderbank were incorporated in a progression of voiceband modem standards that moved communications practice close to the Shannon limit. Together with Peter Shor and colleagues at AT&T Labs he showed that good quantum error correcting codes exist and developed the group theoretic framework for quantum error correction. He is a co-inventor of space-time codes for wireless communication, where correlation of signals across different transmit antennas is the key to reliable transmission.

Dr. Calderbank served as Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1995 to 1998, and as Associate Editor for Coding Techniques from 1986 to 1989. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2006 to 2008. Dr. Calderbank was honored by the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z_4 linearity of Kerdock and Preparata Codes (joint with A.R. Hammons Jr., P.V. Kumar, N.J.A. Sloane, and P. Sole), and again in 1999 for the invention of space-time codes (joint with V.Tarokh and N. Seshadri). He has received the 2006 IEEE Donald G. Fink Prize Paper Award, the IEEE Millennium Medal, the 2013 IEEE Richard W. Hamming Medal, the 2015 Shannon Award, and he was elected to the US National Academy of Engineering in 2005.