Reconstruction from Anisotropic Random Measurements

Mark Rudelson, and Shuheng Zhou, *Member, IEEE*

***Abstract*—Random matrices are widely used in sparse recovery problems, and the relevant properties of matrices with i.i.d. entries are well understood. The current paper discusses the recently introduced Restricted Eigenvalue (RE) condition, which is among the most general assumptions on the matrix, guaranteeing recovery. We prove a reduction principle showing that the RE condition can be guaranteed by checking the restricted isometry on a certain family of low-dimensional subspaces. This principle allows us to establish the RE condition for several broad classes of random matrices with dependent entries, including random matrices with subgaussian rows and non-trivial covariance structure, as well as matrices with independent rows, and uniformly bounded entries.**

***Index Terms*—ℓ_1 minimization, Sparsity, Restricted Eigenvalue conditions, Subgaussian random matrices, Design matrices with uniformly bounded entries.**

I. Introduction

In a typical high dimensional setting, the number of variables p is much larger than the number of observations n. This challenging setting appears in statistics and signal processing, for example, in regression, covariance selection on Gaussian graphical models, signal reconstruction, and sparse approximation. Consider a simple setting, where we try to recover a vector $\beta \in \mathbb{R}^p$ in the following linear model:

$$Y = X\beta + \epsilon. \tag{1}$$

Here X is an $n \times p$ design matrix, Y is a vector of noisy observations, and ϵ is the noise term. Even in the noiseless case, recovering β (or its support) from (X, Y) seems impossible when $n \ll p$, given that we have more variables than observations.

A line of recent research shows that when β is sparse, that is, when it has a relatively small number of nonzero coefficients, it is possible to recover β from an underdetermined system of equations. In order to ensure reconstruction, the design matrix X needs to behave sufficiently nicely in a sense that it satisfies certain incoherence conditions. One notion of the incoherence which has been formulated in the sparse reconstruction literature [1]–[3] bears the name of Restricted Isometry Property (RIP). It states that for all s-sparse sets T, the matrix X restricted to the columns from T acts as an almost isometry. Let X_T, where $T \subset \{1, \ldots, p\}$ be the $n \times |T|$ submatrix obtained by extracting columns of X indexed by T. For each integer $s = 1, 2, \ldots$ such that $s < p$, the s-restricted isometry constant θ_s of X is the smallest quantity such that

$$(1 - \theta_s) \|c\|_2^2 \leq \|X_T c\|_2^2 / n \leq (1 + \theta_s) \|c\|_2^2, \tag{2}$$

for all $T \subset \{1, \ldots, p\}$ with $|T| \leq s$ and coefficients sequences $(c_j)_{j \in T}$. Throughout this paper, we refer to a vector $\beta \in \mathbb{R}^p$ with at most s non-zero entries, where $s \leq p$, as a **s-sparse** vector.

To understand the formulation of the RIP, consider the simplest noiseless case as mentioned earlier, where we assume $\epsilon = 0$ in (1). Given a set of values $(\langle X^i, \beta \rangle)_{i=1}^n$, where $X^1, X^2, \ldots, X^n$ are independent random vectors in $\mathbb{R}^p$, the basis pursuit program [4] finds $\widehat{\beta}$ which minimizes the ℓ_1-norm of β' among all β' satisfying $X\beta' = X\beta$, where X is a $n \times p$ matrix with rows $X^1, X^2, \ldots, X^n$. This can be cast as a linear program and thus is computationally efficient. Under variants of such conditions, the exact recovery or approximate reconstruction of a sparse β using the basis pursuit program has been shown in a series of powerful results [1]–[3], [5]–[9]. We refer to these papers for further references on earlier results for sparse recovery.

In other words, under the RIP, the design matrix X is taken as a $n \times p$ measurement ensemble through which one aims to recover both the unknown non-zero positions and the magnitude of a s-sparse signal β in $\mathbb{R}^p$ efficiently (thus the name for compressed sensing). Naturally, we wish n to be as small as possible for given values of p and s. It is well known that for random matrices, RIP holds for $s = O(n/\log(p/n))$ with i.i.d. Gaussian random entries, Bernoulli, and in general subgaussian entries [1], [2], [7], [10]–[12]. Recently, it has been shown [13] that RIP holds for $s = O(n/\log^2(p/n))$ when X is a random matrix composed of columns that are independent isotropic vectors with log-concave densities. For a random Fourier ensemble, or randomly sampled rows of orthonormal matrices, it is shown that [8], [9] the RIP holds for $s = O(n/\log^c p)$ for $c = 4$, which improves upon the earlier result of [2] where $c = 6$. To be able to prove RIP for random measurements or design matrix, the isotropicity condition (cf. Definition 5) has been assumed in all literature cited above. This assumption is not always reasonable in statistics and machine learning, where we often come across high dimensional data with correlated entries.

The work of [14] formulated the restricted eigenvalue (RE) condition and showed that it is among the weakest and hence

M. Rudelson is with the Department of Mathematics, University of Michigan, Ann Arbor, MI, 48109 USA email: rudelson@umich.edu. Research was supported in part by NSF grants DMS-0907023 and DMS-1161372 .

S. Zhou is with the Department of Statistics, University of Michigan, Ann Arbor, MI, 48109 USA email: shuhengz@umich.edu

The paper was presented in part at COLT 2012.

Manuscript received January 9, 2012; accepted June 20, 2012. Communicated by A. Tulino, Associate Editor.

the most general conditions in literature imposed on the Gram matrix in order to guarantee nice statistical properties for the Lasso estimator [15] as well as the Dantzig selector [3]. In particular, it is shown to be a relaxation of the RIP under suitable choices of parameters involved in each condition; see [14]. We now state one version of the *Restricted Eigenvalue condition* as formulated in [14]. For some integer $0 < s_0 < p$ and a positive number k_0, $\mathsf{RE}(s_0, k_0, X)$ for matrix X requires that the following holds $\forall \upsilon \neq 0$,

$$\min_{\substack{J \subset \{1,\ldots,p\},\ \|\upsilon_{J^c}\|_1 \leq k_0 \|\upsilon_J\|_1 \\ |J| \leq s_0}} \frac{\|X\upsilon\|_2}{\|\upsilon_J\|_2} > 0, \tag{3}$$

where υ_J represents the subvector of $\upsilon \in \mathbb{R}^p$ confined to a subset J of $\{1,\ldots,p\}$, and the strict inequality signifies that the entity on the left hand side is bounded away from 0. In the context of compressed sensing, RE condition can also be taken as a way to guarantee recovery for anisotropic measurements. We refer to [16] for other conditions which are closely related to the RE condition.

Consider now the linear regression model in (1). For a chosen penalization parameter $\lambda_n \geq 0$, regularized estimation with the ℓ_1-norm penalty, also known as the Lasso [15] refers to the following convex optimization problem

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \tag{4}$$

where the scaling factor $1/(2n)$ is chosen for convenience. Under i.i.d Gaussian noise and the RE condition, bounds on ℓ_2 prediction loss and on ℓ_q, $1 \leq q \leq 2$, loss for estimating the parameter β in (1) for both the Lasso and the Dantzig selector have all been derived in [14]. For a given $\lambda_n \geq 0$, the Dantzig selector is defined as:

$$\text{(DS)} \quad \arg\min_{\widehat{\beta} \in \mathbb{R}^p} \left\|\widehat{\beta}\right\|_1 \quad \text{subject to} \quad \left\|\frac{1}{n}X^T(Y - X\widehat{\beta})\right\|_\infty \leq \lambda_n$$

In particular, ℓ_2 loss of $\Theta(\lambda\sigma\sqrt{s})$ were obtained for the Lasso under $\mathsf{RE}(s, 3, X)$ and the Dantzig selector under $\mathsf{RE}(s, 1, X)$ respectively in [14], where it is shown that $\mathsf{RE}(s, 1, X)$ condition is weaker than the RIP used in [3].

RE condition with parameters s_0 and k_0 was shown to hold for random Gaussian measurements / design matrix which consists of $n = O(s_0 \log p)$ independent copies of a p-dimensional Gaussian random vector Y with covariance matrix Σ in [17], assuming that condition (3) holds for the square root of Σ. The matrix Σ is called the population covariance matrix in this context. As we show below, the bound $n = O(s_0 \log p)$ can be improved to the optimal one $n = O(s_0 \log(p/s_0))$ when $\mathsf{RE}(s_0, k_0, \Sigma^{1/2})$ is replaced with $\mathsf{RE}(s_0, (1+\varepsilon)k_0, \Sigma^{1/2})$ for any $\varepsilon > 0$. The work by [17] has motivated the investigation for a non-iid subgaussian random design by [18], as well as the present work. The proof of [17] relies on a deep result from the theory of Gaussian random processes – Gordon's Minimax Lemma [19]. However, this result relies on the properties of the normal random variables, and is not available beyond the Gaussian setting. To establish the RE condition for more general classes of random matrices we had to introduce a new approach based on geometric functional analysis. We defer the comparison of the present paper with [18] to Section I-B. Both [20] and [16] obtained weaker (but earlier) results which are based on bounding the maximum entry-wise difference between sample and the population covariance matrices. We refer to [17] for a more elaborate comparison.

A. Notation and definitions

Let $e_1, \ldots, e_p$ be the canonical basis of $\mathbb{R}^p$. For a set $J \subset \{1,\ldots,p\}$, denote $E_J = \text{span}\{e_j : j \in J\}$. We denote by $[1,p]$ the set $\{1,\ldots,p\}$. For a matrix A, we use $\|A\|_2$ to denote its operator norm. For a set $V \subset \mathbb{R}^p$, we let conv V denote the convex hull of the set V and absconv V denote the absolutely convex hull of the set V. For a finite set Y, the cardinality is denoted by $|Y|$. Let B_2^p and S^{p-1} be the unit Euclidean ball and the unit sphere respectively. For a vector $u \in \mathbb{R}^p$, let T_0 denote the locations of the s_0 largest coefficients of u in absolute values, and u_{T_0} be the subvector of u confined to the locations of its s_0 largest coefficients in absolute values. In this paper, C, c, etc, denote various absolute constants which may change line by line. Occasionally, we use $u_T \in \mathbb{R}^{|T|}$, where $T \subseteq \{1,\ldots,p\}$, to also represent its 0-extended version $u' \in \mathbb{R}^p$ such that $u'_{T^c} = 0$ and $u'_T = u_T$.

We define $\mathcal{C}(s_0, k_0)$, where $0 < s_0 < p$ and k_0 is a positive number, as the set of vectors in $\mathbb{R}^p$ which satisfy the following cone constraint:

$$\mathcal{C}(s_0, k_0) = \left\{x \in \mathbb{R}^p \mid \exists I \in [1,p], |I| = s_0 \text{ s.t. } \|x_{I^c}\|_1 \leq k_0 \|x_I\|_1\right\}. \tag{5}$$

Let β be a s-sparse vector and $\widehat{\beta}$ be the solution from either the Lasso or the Dantzig selector. One of the common properties of the Lasso and the Dantzig selector is: for an appropriately chosen λ_n and under i.i.d. Gaussian noise, the condition

$$\upsilon := \widehat{\beta} - \beta \in \mathcal{C}(s, k_0) \tag{6}$$

holds with high probability. Here $k_0 = 1$ for the Dantzig selector, and $k_0 = 3$ for the Lasso; see [14] and [3] for example. The combination of the cone property (6) and the RE condition leads to various nice convergence results as stated earlier.

We now define some parameters related to the RE and sparse eigenvalue conditions.

Definition 1: Let $1 \leq s_0 \leq p$, and let k_0 be a positive number. We say that a $q \times p$ matrix A satisfies $\mathsf{RE}(s_0, k_0, A)$ condition with parameter $K(s_0, k_0, A)$ if for any $\upsilon \neq 0$,

$$\frac{1}{K(s_0, k_0, A)} := \min_{\substack{J \subseteq \{1,\ldots,p\},\ \|\upsilon_{J^c}\|_1 \leq k_0 \|\upsilon_J\|_1 \\ |J| \leq s_0}} \frac{\|A\upsilon\|_2}{\|\upsilon_J\|_2} > 0. \tag{7}$$

It is clear that when s_0 and k_0 become smaller, this condition is easier to satisfy.

Definition 2: For $m \leq p$, we define the largest and smallest m-sparse eigenvalue of a $q \times p$ matrix A to be

$$\rho_{\max}(m, A) := \max_{t \in \mathbb{R}^p, t \neq 0; m-\text{sparse}} \|At\|_2^2 / \|t\|_2^2, \tag{8}$$

$$\rho_{\min}(m, A) := \min_{t \in \mathbb{R}^p, t \neq 0; m-\text{sparse}} \|At\|_2^2 / \|t\|_2^2. \tag{9}$$

B. Main results

The main purpose of this paper is to show that the RE condition holds with high probability for systems of random measurements/random design matrices of a general nature. To establish such result with high probability, one has to assume that it holds in average. So, our problem boils down to showing that, under some assumptions on random variables, the RE condition on the covariance matrix implies a similar condition on a random design matrix with high probability when n is sufficiently large (cf. Theorems 6 and Theorem 8). This generalizes the results on RIP mentioned above, where the covariance matrix is assumed to be identity. Denote by A a fixed $q \times p$ matrix. We consider the design matrix

$$X = \Psi A, \tag{10}$$

where the rows of the matrix Ψ are isotropic random vectors. An example of such a random matrix X consists of independent rows, each being a random vector in $\mathbb{R}^p$ that follows a multivariate normal distribution $N(0, \Sigma)$, when we take $A = \Sigma^{1/2}$ in (10). Our first main result is related to this setup. We consider a matrix represented as $\tilde{X} = \tilde{\Psi} A$, where the matrix A satisfies the RE condition. The result is purely geometric, so we consider a *deterministic* matrix $\tilde{\Psi}$.

We prove a general reduction principle showing that if the matrix $\tilde{\Psi}$ acts as almost isometry on the images of the sparse vectors under A, then the product $\tilde{\Psi} A$ satisfies the RE condition with a smaller parameter k_0. More precisely, we prove Theorem 3.

Theorem 3: Let $1/5 > \delta > 0$. Let $0 < s_0 < p$ and $k_0 > 0$. Let A be a $q \times p$ matrix such that $\mathsf{RE}(s_0, 3k_0, A)$ holds for $0 < K(s_0, 3k_0, A) < \infty$. Set

$$d = s_0 + s_0 \max_j \|Ae_j\|_2^2 \times \frac{16K^2(s_0, 3k_0, A)(3k_0)^2(3k_0+1)}{\delta^2}. \tag{11}$$

Let $E = \cup_{|J|=d} E_J$ for $d < p$ and E denotes $\mathbb{R}^p$ otherwise. Let $\tilde{\Psi}$ be a matrix such that

$$\forall x \in AE \quad (1-\delta)\|x\|_2 \le \left\|\tilde{\Psi} x\right\|_2 \le (1+\delta)\|x\|_2 . \tag{12}$$

Then $\mathsf{RE}(s_0, k_0, \tilde{\Psi} A)$ condition holds with

$$0 < K(s_0, k_0, \tilde{\Psi} A) \le K(s_0, k_0, A)/(1-5\delta).$$

Remark 4: We note that this result does not involve $\rho_{\max}(s_0, A)$, nor the global parameters of the matrices A and $\tilde{\Psi}$, such as the norm or the smallest singular value. We refer to [17] for an example of matrix A satisfying the RE condition, such that $\rho_{\max}(s_0, A)$ grows linearly with s_0 while the maximum of $\|Ae_j\|_2$ is bounded above.

The assumption $\mathsf{RE}(s_0, 3k_0, A)$ can be replaced by $\mathsf{RE}(s_0, (1+\varepsilon)k_0, A)$ for any $\varepsilon > 0$ by appropriately increasing d. See Remark 15 for details.

We apply the reduction principle to analyze different classes of random design matrices. This analysis is reduced to checking that the almost isometry property holds for all vectors from some low-dimensional subspaces, which is easier than checking the RE property directly. The first example is the matrix Ψ whose rows are independent isotropic vectors with *subgaussian* marginals as in Definition 5. This result extends a theorem of [17] to a non-Gaussian setting, in which the entries of the design matrix may even not have a density.

Definition 5: Let Y be a random vector in $\mathbb{R}^p$

1) Y is called isotropic if for every $y \in \mathbb{R}^p$,
$$\mathbb{E}\,|\langle Y, y\rangle|^2 = \|y\|_2^2 .$$
2) Y is ψ_2 with a constant α if for every $y \in \mathbb{R}^p$,
$$\|\langle Y, y\rangle\|_{\psi_2} := \inf\{t : \mathbb{E}\exp(\langle Y, y\rangle^2/t^2) \le 2\} \le \alpha \|y\|_2 .$$

The ψ_2 condition on a scalar random variable V is equivalent to the subgaussian tail decay of V, which means for some constant c,

$$\mathbb{P}(|V| > t) \le 2\exp(-t^2/c^2), \text{ for all } t > 0.$$

We use ψ_2, vector with subgaussian marginals and subgaussian vector interchangeably. Examples of isotropic random vectors with subgaussian marginals are:

- The random vector Y with i.i.d $N(0,1)$ random coordinates.
- Discrete Gaussian vector, which is a random vector taking values on the integer lattice $\mathbb{Z}^p$ with distribution $\mathbb{P}(X = m) = C\exp(-\|m\|_2^2/2)$ for $m \in \mathbb{Z}^p$.
- A vector with independent centered bounded random coordinates. The subgaussian property here follows from the Hoeffding inequality for sums of independent random variables. This example includes, in particular, vectors with random symmetric Bernoulli coordinates, in other words, random vertices of the discrete cube.

It is hard to argue that such multivariate Gaussian or Bernoulli random designs are not relevant for statistical applications.

Theorem 6: Set $0 < \delta < 1$, $k_0 > 0$, and $0 < s_0 < p$. Let A be a $q \times p$ matrix satisfying $\mathsf{RE}(s_0, 3k_0, A)$ condition as in Definition 1. Let d be as defined in (11), and let $m = \min(d, p)$. Let Ψ be an $n \times q$ matrix whose rows are independent isotropic ψ_2 random vectors in $\mathbb{R}^q$ with constant α. Suppose the sample size satisfies

$$n \ge \frac{2000 m \alpha^4}{\delta^2} \log\left(\frac{60ep}{m\delta}\right). \tag{13}$$

Then with probability at least $1 - 2\exp(-\delta^2 n/2000\alpha^4)$, $\mathsf{RE}\left(s_0, k_0, \frac{1}{\sqrt{n}}\Psi A\right)$ condition holds for matrix $\frac{1}{\sqrt{n}}\Psi A$ with

$$0 < K\left(s_0, k_0, \frac{1}{\sqrt{n}}\Psi A\right) \le \frac{K(s_0, k_0, A)}{1-\delta}. \tag{14}$$

Remark 7: We note that all constants in Theorem 6 are explicit, although they are not optimized.

The reconstruction of sparse signals by subgaussian design matrices was analyzed in [12] and [11]. Note however that both papers used the RIP assumptions and estimate the deviation of the restricted operator from identity. These methods are not applicable in our contexts since the matrix A may be far from identity.

Theorem 6 is applicable in various contexts. We describe two examples. The first example concerns cases which have

been considered in [17], [18]. They show that the RE condition on the covariance matrix Σ implies a similar condition on a random design matrix $X = \Psi\Sigma^{1/2}$ with high probability when n is sufficiently large. In particular, in [18], the author considered subgaussian random matrices of the form $X = \Psi\Sigma^{1/2}$ where Σ is a $p \times p$ positive semidefinite matrix satisfying $\mathsf{RE}(s_0, k_0, \Sigma^{1/2})$ condition, and Ψ is as in Theorem 6. Unlike the current paper, the author allowed $\rho_{\max}(s_0, \Sigma^{1/2})$ as well as $K^2(s_0, k_0, \Sigma^{1/2})$ to appear in the lower bound on n, and showed that $X/\sqrt{n}$ satisfies the RE condition as in (14) with overwhelming probability whenever

$$n > \frac{9c'\alpha^4}{\delta^2}(2+k_0)^2K^2(s_0,k_0,\Sigma^{1/2})\times \\ \min(4\rho_{\max}(s_0,\Sigma^{1/2})s_0\log(5ep/s_0), s_0\log p) \quad (15)$$

where the first term was given in [18, Theorem 1.6] explicitly, and the second term is an easy consequence by combining arguments in [18] and [17]. Analysis there used Corollary 2.7 in [21] crucially. In the present work, we get rid of the dependency of the sample size on $\rho_{\max}(s_0, \Sigma^{1/2})$, although under a slightly stronger $\mathsf{RE}(s_0, 3k_0, \Sigma^{1/2})$ (See Remarks 4 and 15). More precisely, let Σ be a $p \times p$ covariance matrix satisfying $\mathsf{RE}(s_0, 3k_0, \Sigma^{1/2})$ condition. Then, (14) implies that with probability at least $1 - 2\exp(-\delta^2 n/2000\alpha^4)$,

$$0 < K\left(s_0, k_0, \frac{1}{\sqrt{n}}\Psi\Sigma^{1/2}\right) \le \frac{K(s_0,k_0,\Sigma^{1/2})}{1-\delta} \quad (16)$$

where n satisfies (13) for d defined in (11), with A replaced by $\Sigma^{1/2}$. In particular, bounds developed in the present paper can be applied to obtain tight convergence results for covariance estimation for a multivariate Gaussian model [22].

Another application of Theorem 6 is given in [23]. The $q \times p$ matrix A can be taken as a data matrix with p attributes (e.g., weight, height, age, etc), and q individual records. The data are compressed by a random linear transformation $X = \Psi A$. Such transformations have have been called "matrix masking" in the privacy literature [24]. We think of X as "public," while Ψ, which is a $n \times q$ random matrix, is private and only needed at the time of compression. However, even with Ψ known, recovering A from Ψ requires solving a highly under-determined linear system and comes with information theoretic privacy guarantees when $n \ll q$, as demonstrated in [23]. On the other hand, sparse recovery using X is highly feasible given that the RE conditions are guaranteed to hold by Theorem 6 with a small n. We refer to [23] for a detailed setup on regression using compressed data as in (10).

The second application of the reduction principle is to the design matrices with uniformly bounded entries. As we mentioned above, if the entries of such matrix are independent, then its rows are subgaussian. However, the independence of entries is not assumed, so the decay of the marginals can be arbitrary slow. Indeed, if all coordinates of the vector equal to the same symmetric Bernoulli random variable, then the maximal ψ_2-norm of the marginals is of the order $\sqrt{p}$.

A natural example for compressed sensing would be measurements of random Fourier coefficients, when some of the coefficients cannot be measured.

Theorem 8: Let $0 < \delta < 1$ and $0 < s_0 < p$. Let $Y \in \mathbb{R}^p$ be a random vector such that $\|Y\|_\infty \le M$ a.s and denote $\Sigma = \mathbb{E}YY^T$. Let X be an $n \times p$ matrix, whose rows $X_1, \dots, X_n$ are independent copies of Y. Let Σ satisfy the $\mathsf{RE}(s_0, 3k_0, \Sigma^{1/2})$ condition as in Definition 1. Let d be as defined in (11), where we replace A with $\Sigma^{1/2}$. Assume that $d \le p$ and $\rho = \rho_{\min}(d, \Sigma^{1/2}) > 0$. Suppose the sample size satisfies for some absolute constant C

$$n \ge \frac{CM^2d\cdot\log p}{\rho\delta^2}\cdot\log^3\left(\frac{CM^2d\cdot\log p}{\rho\delta^2}\right).$$

Then with probability at least $1 - \exp\left(-\delta\rho n/(6M^2d)\right)$, $\mathsf{RE}(s_0, k_0, X)$ condition holds for matrix $\frac{1}{\sqrt{n}}X$ with

$$0 < K\left(s_0, k_0, \frac{1}{\sqrt{n}}X\right) \le \frac{K(s_0,k_0,\Sigma^{1/2})}{1-\delta}.$$

Remark 9: Note that unlike the case of a random matrix with subgaussian marginals, the estimate of Theorem 8 contains the minimal sparse singular value ρ. We will provide an example illustrating that this is necessary in Remark 25.

We will prove Theorems 3, 6, and 8 in Sections II, III, and IV respectively.

We note that the reduction principle can be applied to other types of random variables. One can consider the case of heavy-tailed marginals. In this case the estimate for the images of sparse vectors can be proved using the technique developed by [25], [26]. One can also consider random vectors with log-concave densities, and obtain similar estimates following the methods of [13], [27].

To make our exposition complete, we will show some immediate consequences in terms of statistical inference on high dimensional data that satisfy such RE and sparse eigenvalue conditions. As mentioned, the restricted eigenvalue (RE) condition as formulated by [14] are among the weakest and hence the most general conditions in literature imposed on the Gram matrix in order to guarantee nice statistical properties for the Lasso and the Dantzig selector. For random design as considered in the present paper, one can show that various oracle inequalities in terms of ℓ_2 convergence hold for the Lasso and the Dantzig selector as long as n satisfies the lower bounds above. Let $s = |\text{supp}\ \beta|$ for β in (1). Under $\mathsf{RE}(s, 9, \Sigma^{1/2})$, a sample size of $n = O(s\log(p/s))$ is sufficient for us to derive bounds corresponding to those in [14, Theorem 7.2]; see also [18, Theorem 3.1, 3.2]. As a consequence, we see that this setup requires only $\Theta(\log(p/s))$ observations per nonzero value in β where Θ hides a constant depending on $K^2(s, 9, \Sigma^{1/2})$ for the family of random matrices with subgaussian marginals which satisfies $\mathsf{RE}(s, 9, \Sigma^{1/2})$ condition. Similarly, we note that for random matrix X with a.s. bounded entries of size M, $n = O(sM^2\log p\log^3(s\log p))$ samples are sufficient in order to achieve accurate statistical estimation. We say this is a *linear or sublinear sparsity*. For $p \gg n$, this is a desirable property as it implies that accurate statistical estimation is feasible given a very limited amount of data.

II. Reduction principle

We first reformulate the reduction principle in the form of restrictive isometry: we show that if the matrix $\tilde{\Psi}$ acts as

almost isometry on the images of the sparse vectors under A, then it acts the same way on the images of a set of vectors which satisfy the cone constraint (5). We then prove Theorem 3 as a corollary of Theorem 10 The proof of Theorem 10 itself uses several auxiliary results, which will be established in the next two subsections.

Theorem 10: Let $1/5 > \delta > 0$. Let $0 < s_0 < p$ and $k_0 > 0$. Let A be a $q \times p$ matrix such that $\mathsf{RE}(s_0, 3k_0, A)$ condition holds for $0 < K(s_0, 3k_0, A) < \infty$. Set

$$d = s_0 + s_0 \max_j \|Ae_j\|_2^2 \left(\frac{16K^2(s_0, 3k_0, A)(3k_0)^2(3k_0+1)}{\delta^2} \right),$$

and let $E = \cup_{|J|=d} E_J$ for $d < p$ and $E = \mathbb{R}^p$ otherwise. Let $\tilde{\Psi}$ be a matrix such that

$$\forall x \in AE \quad (1-\delta)\|x\|_2 \le \left\|\tilde{\Psi}x\right\|_2 \le (1+\delta)\|x\|_2 . \tag{17}$$

Then for any $x \in A\Big(\mathcal{C}(s_0, k_0)\Big) \cap S^{q-1}$,

$$(1-5\delta) \le \left\|\tilde{\Psi}x\right\|_2 \le (1+3\delta) \tag{18}$$

Proof of Theorem 3. By the $\mathsf{RE}(s_0, 3k_0, A)$ condition, $\mathsf{RE}(s_0, k_0, A)$ condition holds as well. Hence for $u \in \mathcal{C}(s_0, k_0)$ such that $u \neq 0$,

$$\|Au\|_2 \ge \frac{\|u_{T_0}\|_2}{K(s_0, k_0, A)} > 0,$$

and by (18)

$$\left\|\tilde{\Psi}Au\right\|_2 \ge (1-5\delta)\|Au\|_2 \ge (1-5\delta)\frac{\|u_{T_0}\|_2}{K(s_0, k_0, A)} > 0.$$

□

A. Preliminary results

Our first lemma is based on Maurey's empirical approximation argument [28]. We show that any vector belonging to the convex hull of many vectors can be approximated by a convex combination of a few of them.

Lemma 11: Let $u_1, \ldots, u_M \in \mathbb{R}^q$. Let $y \in \operatorname{conv}(u_1, \ldots, u_M)$. Then, there exists a set $L \subset \{1, 2, \ldots, M\}$ such that

$$|L| \le m = \frac{4\max_{j\in\{1,\ldots,M\}} \|u_j\|_2^2}{\varepsilon^2}$$

and a vector $y' \in \operatorname{conv}(u_j, j \in L)$ such that

$$\|y' - y\|_2 \le \varepsilon.$$

Proof: Assume that

$$y = \sum_{j\in\{1,\ldots,M\}} \alpha_j u_j \quad \text{where} \quad \alpha_j \ge 0, \text{ and } \sum_j \alpha_j = 1.$$

Let Y be a random vector in $\mathbb{R}^q$ such that

$$\mathbb{P}(Y = u_\ell) = \alpha_\ell, \ \ell \in \{1, \ldots, M\}$$

Then

$$\mathbb{E}Y = \sum_{\ell\in\{1,\ldots,M\}} \alpha_\ell u_\ell = y.$$

Let $Y_1, \ldots, Y_m$ be independent copies of Y and let $\varepsilon_1, \ldots, \varepsilon_m$ be ± 1 i.i.d. mean zero Bernoulli random variables, chosen independently of $Y_1, \ldots, Y_m$. By the standard symmetrization argument [30, Section 6.1], we have

$$\begin{aligned}
\mathbb{E}\left\|y - \frac{1}{m}\sum_{j=1}^m Y_j\right\|_2^2 &\le 4\mathbb{E}\left\|\frac{1}{m}\sum_{j=1}^m \varepsilon_j Y_j\right\|_2^2 \\
&= \frac{4}{m^2}\sum_{j=1}^m \mathbb{E}\|Y_j\|_2^2 \\
&\le \frac{4\max_{\ell\in\{1,\ldots,M\}}\|u_\ell\|_2^2}{m} \\
&\le \varepsilon^2
\end{aligned} \tag{19}$$

where

$$\mathbb{E}\|Y_j\|_2^2 \le \sup\|Y_j\|_2^2 \le \max_{\ell\in\{1,\ldots,M\}}\|u_\ell\|_2^2$$

and the last inequality in (19) follows from the definition of m.

Fix a realization $(Y_1, \ldots, Y_m) = (u_{k_1}, \ldots, u_{k_m})$ for which

$$\left\|y - \frac{1}{m}\sum_{j=1}^m Y_j\right\|_2 \le \varepsilon.$$

The vector $\frac{1}{m}\sum_{j=1}^m Y_j$ belongs to the convex hull of $\{u_\ell : \ell \in L\}$, where L is the set of different elements from the sequence $k_1, \ldots, k_m$. Obviously $|L| \le m$ and the lemma is proved. ■

For each vector $x \in \mathbb{R}^p$, let T_0 denote the locations of the s_0 largest coefficients of x in absolute values. Any vector $x \in \mathcal{C}(s_0, k_0) \cap S^{p-1}$ satisfies:

$$\begin{aligned}
\left\|x_{T_0^c}\right\|_\infty \le \|x_{T_0}\|_1 / s_0 &\le \frac{\|x_{T_0}\|_2}{\sqrt{s_0}} \\
\left\|x_{T_0^c}\right\|_1 \le k_0\sqrt{s_0}\,\|x_{T_0}\|_2 &\le k_0\sqrt{s_0} \\
\text{and } \left\|x_{T_0^c}\right\|_2 &\le 1.
\end{aligned} \tag{20}$$

The next elementary estimate will be used in conjunction with the RE condition.

Lemma 12: For each vector $\upsilon \in \mathcal{C}(s_0, k_0)$, let T_0 denotes the locations of the s_0 largest coefficients of υ in absolute values. Then

$$\|\upsilon_{T_0}\|_2 \ge \frac{\|\upsilon\|_2}{\sqrt{1+k_0}}. \tag{21}$$

Proof: By definition of $\mathcal{C}(s_0, k_0)$, by (20)

$$\begin{aligned}
\left\|\upsilon_{T_0^c}\right\|_2^2 &\le \left\|\upsilon_{T_0^c}\right\|_1 \left\|\upsilon_{T_0^c}\right\|_\infty \\
&\le k_0\|\upsilon_{T_0}\|_1 \cdot \|\upsilon_{T_0}\|_1 / s_0 \\
&\le k_0\|\upsilon_{T_0}\|_2^2 .
\end{aligned}$$

Therefore $\|\upsilon\|_2^2 = \left\|\upsilon_{T_0^c}\right\|_2^2 + \|\upsilon_{T_0}\|_2^2 \le (k_0+1)\|\upsilon_{T_0}\|_2^2$. ■

The next lemma concerns the extremum of a linear functional on a big circle of a q-dimensional sphere. We consider a line passing through the extreme point, and show that the value of the functional on a point of the line, which is relatively close to the extreme point, provides a good bound for the extremum.

Lemma 13: let $u, \theta, x \in \mathbb{R}^q$ be vectors such that

1) $\|\theta\|_2 = 1$.
2) $\langle x, \theta \rangle \neq 0$.
3) Vector u is not parallel to x.

Define $\phi : \mathbb{R} \to \mathbb{R}$ by:

$$\phi(\lambda) = \frac{\langle x + \lambda u, \theta \rangle}{\|x + \lambda u\|_2}. \tag{22}$$

Assume $\phi(\lambda)$ has a local maximum at 0, then

$$\frac{\langle x + u, \theta \rangle}{\langle x, \theta \rangle} \geq 1 - \frac{\|u\|_2}{\|x\|_2}.$$

Proof: Let $v = \frac{x}{\|x\|_2}$. Also let

$$\begin{aligned} \theta &= \beta v + \gamma t, \text{ where } t \perp v,\ \|t\|_2 = 1 \\ &\quad \text{and } \beta^2 + \gamma^2 = 1, \beta \neq 0 \\ \text{and } u &= \eta v + \mu t + s \text{ where } s \perp v \text{ and } s \perp t. \end{aligned}$$

Define $f : \mathbb{R} \to \mathbb{R}$ by:

$$f(\lambda) = \frac{\lambda}{\|x\|_2 + \lambda\eta},\ \lambda \neq -\frac{\eta}{\|x\|_2}. \tag{23}$$

Then

$$\begin{aligned} \phi(\lambda) &= \frac{\langle x + \lambda u, \theta \rangle}{\|x + \lambda u\|_2} \\ &= \frac{\langle (\|x\|_2 + \lambda\eta)v + \lambda\mu t + \lambda s, \beta v + \gamma t \rangle}{\|(\|x\|_2 + \lambda\eta)v + \lambda\mu t + \lambda s\|_2} \\ &= \frac{\beta(\|x\|_2 + \lambda\eta) + \lambda\mu\gamma}{\sqrt{(\|x\|_2 + \lambda\eta)^2 + (\lambda\mu)^2 + \lambda^2 \|s\|_2^2}} \\ &= \frac{\beta + \mu\gamma f(\lambda)}{\sqrt{1 + (\mu^2 + \|s\|_2^2) f^2(\lambda)}} \end{aligned}$$

Since $f(\lambda) = \frac{\lambda}{\|x\|_2} + O(\lambda^2)$ we have

$$\phi(\lambda) = \beta + \mu\gamma \frac{\lambda}{\|x\|_2} + O(\lambda^2)$$

in the neighborhood of 0, Hence, in order to for $\phi(\lambda)$ to have a local maximum at 0, μ or γ must be 0. Consider these cases separately.

- First suppose $\gamma = 0$, then $\beta^2 = 1$ and $|\langle x, \theta \rangle| = \|x\|_2$. Hence,

  $$\frac{\langle x + u, \theta \rangle}{\langle x, \theta \rangle} = 1 + \frac{\langle u, \theta \rangle}{\langle x, \theta \rangle} \geq 1 - \frac{|\langle u, \theta \rangle|}{|\langle x, \theta \rangle|} \geq 1 - \frac{\|u\|_2}{\|x\|_2}$$

  where $|\langle u, \theta \rangle| \leq \|u\|_2$.
- Otherwise, suppose that $\mu = 0$. Then we have $|\eta| = |\langle u, v \rangle| \leq \|u\|_2$ and

  $$\begin{aligned} \frac{\langle x + u, \theta \rangle}{\langle x, \theta \rangle} &= 1 + \frac{\langle \eta v + s, \beta v + \gamma t \rangle}{\langle v \|x\|_2, \beta v + \gamma t \rangle} \\ &= 1 + \frac{\eta\beta}{\|x\|_2 \beta} = 1 + \frac{\eta}{\|x\|_2} \\ &\geq 1 - \frac{\|u\|_2}{\|x\|_2} \end{aligned}$$

  where we used the fact that $\beta \neq 0$ given $\langle x, \theta \rangle \neq 0$.

■

B. Convex hull of sparse vectors

For a set $J \subset \{1, \ldots, p\}$, denote $E_J = \text{span}\{e_j : j \in J\}$. In order to prove the restricted isometry property of Ψ over the set of vectors in $A\Big(\mathcal{C}(s_0, k_0)\Big) \cap S^{q-1}$, we first show that this set is contained in the convex hull of the images of the sparse vectors with norms not exceeding $(1 - \delta)^{-1}$. More precisely, we state the following lemma.

Lemma 14: Let $1 > \delta > 0$. Let $0 < s_0 < p$ and $k_0 > 0$. Let A be a $q \times p$ matrix such that $\mathsf{RE}(s_0, k_0, A)$ condition holds for $0 < K(s_0, k_0, A) < \infty$. Define

$$\begin{aligned} d = d(k_0, A) = \\ s_0 + s_0 \max_j \|Ae_j\|_2^2 \left(\frac{16K^2(s_0, k_0, A) k_0^2 (k_0 + 1)}{\delta^2} \right). \end{aligned} \tag{24}$$

Then

$$\begin{aligned} A\Big(\mathcal{C}(s_0, k_0)\Big) \cap S^{q-1} \subset \\ (1 - \delta)^{-1} \text{conv} \left(\bigcup_{|J| \leq d} AE_J \cap S^{q-1} \right) \end{aligned} \tag{25}$$

where for $d \geq p$, E_J is understood to be $\mathbb{R}^p$.

Proof: Without loss of generality, assume that $d(k_0, A) < p$, otherwise the lemma is vacuously true. For each vector $x \in \mathbb{R}^p$, let T_0 denote the locations of the s_0 largest coefficients of x in absolute values. Decompose a vector $x \in \mathcal{C}(s_0, k_0) \cap S^{p-1}$ as

$$x = x_{T_0} + x_{T_0^c} \in x_{T_0} + k_0 \|x_{T_0}\|_1 \text{absconv}(e_j \mid j \in T_0^c)$$

where $\|x_{T_0}\|_2 \geq \frac{1}{\sqrt{k_0 + 1}}$ by (21) and hence

$$Ax \in Ax_{T_0} + k_0 \|x_{T_0}\|_1 \text{absconv}(Ae_j \mid j \in T_0^c).$$

Since the set $A\mathcal{C}(s_0, k_0) \cap S^{q-1}$ is not easy to analyze, we introduce set of a simpler structure instead. Define V as the set

$$\left\{ x_{T_0} + k_0 \|x_{T_0}\|_1 \text{absconv}(e_j \mid j \in T_0^c) | x \in \mathcal{C}(s_0, k_0) \cap S^{p-1} \right\}.$$

For a given $x \in \mathcal{C}(s_0, k_0) \cap S^{p-1}$, if T_0 is not uniquely defined, we include all possible sets of T_0 in the definition of V. Clearly $V \subset \mathcal{C}(s_0, k_0)$ is a compact set. Moreover, V contains a base of $\mathcal{C}(s_0, k_0)$, that is, for any $y \in \mathcal{C}(s_0, k_0) \setminus \{0\}$ there exists $\lambda > 0$ such that $\lambda y \in V$.

For any $v \in \mathbb{R}^p$ such that $\|Av\|_2 \neq 0$, define

$$F(v) = \frac{Av}{\|Av\|_2}.$$

By condition $\mathsf{RE}(s_0, k_0, A)$, the function F is well-defined and continuous on $\mathcal{C}(s_0, k_0) \setminus \{0\}$, and, in particular, on V. Hence,

$$A\mathcal{C}(s_0, k_0) \cap S^{q-1} = F\big(\mathcal{C}(s_0, k_0) \setminus \{0\}\big) = F(V).$$

By duality, inclusion (25) can be derived from the fact that the supremum of any linear functional over the left side of (25) does not exceed the supremum over the right side of it. By the equality above, it is enough to show that for any $\theta \in S^{q-1}$,

there exists $z' \in \mathbb{R}^p \setminus \{0\}$ such that $|\operatorname{supp}(z')| \le d$ and $F(z')$ is well defined, which satisfies

$$\max_{v\in V} \langle F(v), \theta \rangle \quad \le \quad (1-\delta)^{-1} \langle F(z'), \theta \rangle. \tag{26}$$

For a given θ, we construct a d-sparse vector z' which satisfies (26). Let

$$z := \arg\max_{v\in V} \langle F(v), \theta \rangle.$$

By definition of V there exists $I \subset \{1,\dots,p\}$ such that $|I| = s_0$, and for some $\varepsilon_j \in \{1,-1\}$,

$$z = z_I + \|z_I\|_1 k_0 \sum_{j\in I^c} \alpha_j \varepsilon_j e_j \tag{27}$$

where

$$\alpha_j \in [0,1], \sum_{j\in I^c} \alpha_j \le 1, \text{ and } 1 \ge \|z_I\|_2 \ge \frac{1}{\sqrt{k_0+1}}.$$

Note if $\alpha_i = 1$ for some $i \in I^c$, then z is a sparse vector itself, and we can set $z' = z$ in order for (26) to hold. We proceed assuming $\alpha_i \in [0,1)$ for all $i \in I^c$ in (27) from now on, in which case, we construct a required sparse vector z' via Lemma 11. To satisfy the assumptions of this lemma, denote $e_{p+1} = \vec{0}$, $\varepsilon_{p+1} = 1$ and set

$$\alpha_{p+1} = 1 - \sum_{j\in I^c} \alpha_j, \quad \text{hence} \quad \alpha_{p+1} \in [0,1].$$

Let

$$\begin{aligned} y \quad &:= \quad Az_{I^c} = \|z_I\|_1 k_0 \sum_{j\in I^c} \alpha_j \varepsilon_j A e_j \\ &= \quad \|z_I\|_1 k_0 \sum_{j\in I^c\cup\{p+1\}} \alpha_j \varepsilon_j A e_j \end{aligned}$$

and denote $\mathcal{M} := \{j \in I^c \cup \{p+1\} : \alpha_j > 0\}$. Let $\varepsilon > 0$ be specified later. Applying Lemma 11 with vectors $u_j = k_0 \|z_I\|_1 \varepsilon_j A e_j$ for $j \in \mathcal{M}$, construct a set $J' \subset \mathcal{M}$ satisfying

$$\begin{aligned} |J'| \le m \quad &:= \quad \frac{4 \max_{j\in I^c} k_0^2 \|z_I\|_1^2 \|Ae_j\|_2^2}{\varepsilon^2} \\ &\le \quad \frac{4k_0^2 s_0 \max_{j\in I^c} \|Ae_j\|_2^2}{\varepsilon^2} \end{aligned} \tag{28}$$

and a vector

$$y' = k_0 \|z_I\|_1 \sum_{j\in J'} \beta_j \varepsilon_j A e_j$$

where for $J' \subset \mathcal{M}$,

$$\beta_j \in [0,1] \text{ and } \sum_{j\in J'} \beta_j = 1$$

such that $\|y' - y\|_2 \le \varepsilon$.

Set $u := k_0 \|z_I\|_1 \sum_{j\in J'} \beta_j \varepsilon_j e_j$ and let

$$z' \quad = \quad z_I + u.$$

By construction, $Az' \in AE_J$, where $J := (I \cup J') \cap \{1,\dots,p\}$ and

$$|J| \le |I| + |J'| \le s_0 + m. \tag{29}$$

Furthermore, we have

$$\|Az - Az'\|_2 = \|A(z_{I^c} - u)\|_2 = \|y - y'\|_2 \le \varepsilon$$

For $\{\beta_j, j \in J'\}$ as above, we extend it to

$$\{\beta_j, j \in I^c \cup \{p+1\}\}$$

setting $\beta_j = 0$ for all $j \in I^c \cup \{p+1\} \setminus J'$ and write

$$\begin{aligned} z' \quad &= \quad z_I + k_0 \|z_I\|_1 \sum_{j\in I^c\cup\{p+1\}} \beta_j \varepsilon_j e_j \\ \text{where} \qquad &\beta_j \in [0,1] \text{ and } \sum_{j\in I^c\cup\{p+1\}} \beta_j = 1 \end{aligned}$$

If $z' = z$, we are done. Otherwise, for some λ to be specified, consider the vector

$$z + \lambda(z'-z) = z_I + k_0 \|z_I\|_1 \sum_{j\in I^c\cup\{p+1\}} [(1-\lambda)\alpha_j + \lambda\beta_j] \varepsilon_j e_j.$$

We have $\sum_{j\in I^c\cup\{p+1\}} [(1-\lambda)\alpha_j + \lambda\beta_j] = 1$ and

$$\begin{aligned} &\exists\, \delta_0 > 0 \text{ s. t. } \forall j \in I^c \cup \{p+1\}, \\ &(1-\lambda)\alpha_j + \lambda\beta_j \in [0,1] \text{ if } |\lambda| < \delta_0. \end{aligned}$$

To see this, we note that

- This condition holds by continuity for all j such that $\alpha_i \in (0,1)$.
- If $\alpha_j = 0$ for some j, then $\beta_j = 0$ by construction.

Thus $\sum_{j\in I^c} [(1-\lambda)\alpha_j + \lambda\beta_j] \le 1$ and $z + \lambda(z'-z) = z_I + k_0 \|z_I\|_1 \sum_{j\in I^c} [(1-\lambda)\alpha_j + \lambda\beta_j] \varepsilon_j e_j \in V$ whenever $|\lambda| < \delta_0$.

Consider now a function $\phi : (-\delta_0, \delta_0) \to \mathbb{R}$,

$$\phi(\lambda) := \langle F(z + \lambda(z'-z)), \theta \rangle = \frac{\langle Az + \lambda(Az' - Az), \theta \rangle}{\|Az + \lambda(Az' - Az)\|_2}$$

Since z maximizes $\langle F(v), \theta \rangle$ for all $v \in V$, $\phi(\lambda)$ attains the local maximum at 0. Then by Lemma 13, we have

$$\begin{aligned} \frac{\langle Az', \theta \rangle}{\langle Az, \theta \rangle} \quad &= \quad \frac{\langle Az + (Az' - Az), \theta \rangle}{\langle Az, \theta \rangle} \\ &\ge \quad 1 - \frac{\|(Az' - Az)\|_2}{\|Az\|_2} \\ &= \quad \frac{\|Az\|_2 - \|(Az' - Az)\|_2}{\|Az\|_2} \end{aligned}$$

hence

$$\begin{aligned} \frac{\langle F(z'), \theta \rangle}{\langle F(z), \theta \rangle} &= \frac{\langle Az'/\|Az'\|_2, \theta \rangle}{\langle Az/\|Az\|_2, \theta \rangle} \\ &= \quad \frac{\|Az\|_2}{\|Az'\|_2} \times \frac{\langle Az', \theta \rangle}{\langle Az, \theta \rangle} \\ &\ge \quad \frac{\|Az\|_2}{\|Az\|_2 + \|(Az' - Az)\|_2} \times \frac{\|Az\|_2 - \|(Az' - Az)\|_2}{\|Az\|_2} \\ &= \quad \frac{\|Az\|_2 - \|(Az' - Az)\|_2}{\|Az\|_2 + \|(Az' - Az)\|_2} = \frac{\|Az\|_2 - \varepsilon}{\|Az\|_2 + \varepsilon} \\ &= \quad 1 - \frac{2\varepsilon}{\|Az\|_2 + \varepsilon}. \end{aligned}$$

By definition, $z \in \mathcal{C}(s_0, k_0)$. Hence we apply $\mathsf{RE}(k_0, s_0, A)$ condition and (27) to obtain

$$\|Az\|_2 \geq \frac{\|z_I\|_2}{K(s_0, k_0, A)} \geq \frac{1}{\sqrt{1+k_0}K(s_0, k_0, A)}.$$

Now we can set $\varepsilon = \frac{\delta}{2\sqrt{1+k_0}K(s_0,k_0,A)}$ which yields

$$\frac{\langle F(z'), \theta \rangle}{\langle F(z), \theta \rangle} \geq 1 - \delta \tag{30}$$

and thus (26) holds. Finally, by (28), we have

$$m \leq s_0 \max_{j \in I^c} \|Ae_j\|_2^2 \left(\frac{16K^2(s_0, k_0, A)k_0^2(k_0+1)}{\delta^2} \right)$$

and hence the inclusion (25) holds in view of (29) and (30). ∎

C. Proof of the reduction principle

To prove the restricted isomorphism condition (18), we apply Lemma 14 with k_0 being replaced by $3k_0$. The upper bound in (18) follows immediately from the lemma. To prove the lower bound, we consider a vector $x \in \mathcal{C}(s_0, k_0)$ as an endpoint of an interval, whose midpoint is a sparse vector from the same cone. Then the other endpoint of the interval will be contained in the larger cone $\mathcal{C}(s_0, 3k_0)$. Comparison between the upper estimate for the norm of the image of this endpoint with the lower estimate for the midpoint will yield the required lower estimate for the point x.

Proof of Theorem 10. Let $v \in \mathcal{C}(s_0, 3k_0) \setminus \{0\}$, and so $\|Av\|_2 > 0$ by $\mathsf{RE}(s_0, 3k_0, A)$ condition. Let $d(3k_0, A)$ be defined as in (24). As in the proof of Lemma 14, we may assume that $d(3k_0, A) < p$. By Lemma 14, applied with k_0 replaced with $3k_0$, we have

$$\begin{aligned} \frac{Av}{\|Av\|_2} &\in A\Big(\mathcal{C}(s_0, 3k_0)\Big) \cap S^{q-1} \\ &\subset (1-\delta)^{-1} \operatorname{conv} \left(\bigcup_{|J| = d(3k_0, A)} AE_J \cap S^{q-1} \right) \end{aligned}$$

and

$$\begin{aligned} \left\| \frac{\tilde{\Psi}Av}{\|Av\|_2} \right\|_2 &\leq \frac{1}{1-\delta} \max_{u \in \operatorname{conv}(AE \cap S^{q-1})} \left\| \tilde{\Psi}u \right\|_2 \\ &= \frac{1}{1-\delta} \max_{u \in AE \cap S^{q-1}} \left\| \tilde{\Psi}u \right\|_2. \end{aligned}$$

The last equality holds, since the maximum of $\|\tilde{\Psi}u\|_2$ occurs at an extreme point of the set $\operatorname{conv}(AE \cap S^{q-1})$, because of convexity of the function $f(x) = \|\tilde{\Psi}x\|_2$. Hence, by (17)

$$\begin{aligned} \forall x &\in A\Big(\mathcal{C}(s_0, 3k_0)\Big) \cap S^{q-1}, \\ \left\| \tilde{\Psi}x \right\|_2 &\leq (1+\delta)(1-\delta)^{-1} \leq 1 + 3\delta \end{aligned} \tag{31}$$

where the last inequality is satisfied once $\delta < 1/3$, which proves the upper estimate in (18).

We have to prove the opposite inequality. Let $x = x_I + x_{I^c} \in \mathcal{C}(s_0, k_0) \cap S^{p-1}$, where the set I contains the locations of the s_0 largest coefficients of x in absolute values. We have

$$x = x_I + \|x_{I^c}\|_1 \sum_{j \in I^c} \frac{|x_j|}{\|x_{I^c}\|_1} \operatorname{sgn}(x_j) e_j \tag{32}$$

where by (21).

$$1 \geq \|x_I\|_2 \geq \frac{1}{\sqrt{k_0+1}}.$$

Let $\varepsilon > 0$ be specified later. We now construct a $d(3k_0, A)$-sparse vector $y = x_I + u \in \mathcal{C}(s_0, k_0)$, where u is supported on I^c which satisfies

$$\begin{aligned} \|u\|_1 &= \|y_{I^c}\|_1 = \|x_{I^c}\|_1 \\ \text{and } \|Ax - Ay\|_2 &= \|A(x_{I^c} - y_{I^c})\|_2 \leq \varepsilon \end{aligned} \tag{33}$$

To do so, set

$$w := Ax_{I^c} = \|x_{I^c}\|_1 \sum_{j \in I^c} \frac{|x_j|}{\|x_{I^c}\|_1} \operatorname{sgn}(x_j) Ae_j.$$

Let $\mathcal{M} := \{j \in I^c : x_j \neq 0\}$. Applying Lemma 11 with vectors $u_j = \|x_{I^c}\|_1 \operatorname{sgn}(x_j) Ae_j$ for $j \in \mathcal{M}$, construct a set $J' \subset \mathcal{M}$ satisfying

$$\begin{aligned} |J'| \leq m &:= \frac{4 \max_{j \in \mathcal{M}} \|x_{I^c}\|_1^2 \|Ae_j\|_2^2}{\varepsilon^2} \\ &\leq \frac{4k_0^2 s_0 \max_{j \in \mathcal{M}} \|Ae_j\|_2^2}{\varepsilon^2} \end{aligned} \tag{34}$$

and a vector

$$w' = \|x_{I^c}\|_1 \sum_{j \in J'} \beta_j \operatorname{sgn}(x_j) Ae_j,$$

where for $J' \subset \mathcal{M}$

$$\beta_j \in [0,1] \text{ and } \sum_{j \in J'} \beta_j = 1$$

such that $\|Ax - Ay\|_2 = \|w' - w\|_2 \leq \varepsilon$. Set

$$u := \|x_{I^c}\|_1 \sum_{j \in J'} \beta_j \operatorname{sgn}(x_j) e_j$$

and let

$$\begin{aligned} y &= x_I + u = x_I + \|x_{I^c}\|_1 \sum_{j \in J'} \beta_j \operatorname{sgn}(x_j) e_j \\ \text{where} \quad & \beta_j \in [0,1] \text{ and } \sum_{j \in J'} \beta_j = 1. \end{aligned}$$

By construction, $y \in \mathcal{C}(s_0, k_0) \cap E_J$, where $J := I \cup J'$ and

$$|J| = |I| + |J'| \leq s_0 + m. \tag{35}$$

This, in particular, implies that $\|Ay\|_2 > 0$. Assume that ε is chosen so that $s_0 + m \leq d(3k_0, A)$, and so by (17)

$$\left\| \frac{\tilde{\Psi}Ay}{\|Ay\|_2} \right\|_2 \geq 1 - \delta.$$

Set

$$v = x_I + 2y_{I^c} - x_{I^c} = y + (y_{I^c} - x_{I^c}). \tag{36}$$

Then (33) implies

$$\|Av\|_2 \le \|Ay\|_2 + \|A(y_{I^c} - x_{I^c})\| \le \|Ay\|_2 + \varepsilon, \qquad (37)$$

and $v \in \mathcal{C}(s_0, 3k_0)$ as

$$\begin{aligned} \|v_{I^c}\|_1 &\le 2\|y_{I^c}\|_1 + \|x_{I^c}\|_1 = 3\|x_{I^c}\|_1 \\ &\le 3k_0\|x_I\|_1 = 3k_0\|v_I\|_1 \end{aligned}$$

where we use the fact that $\|x_{I^c}\|_1 = \|y_{I^c}\|_1$. Hence, by the upper estimate (31), we have

$$\left\| \frac{\tilde\Psi A v}{\|Av\|_2} \right\|_2 \le (1+\delta)(1-\delta)^{-1} \qquad (38)$$

Since $y = \frac{1}{2}(x+v)$, where $y_I = x_I$, we have by the lower bound in (17) and the triangle inequality,

$$\begin{aligned} 1-\delta &\le \left\| \frac{\tilde\Psi A y}{\|Ay\|_2} \right\|_2 \\ &\le \frac{1}{2}\left(\left\| \frac{\tilde\Psi A x}{\|Ay\|_2} \right\|_2 + \left\| \frac{\tilde\Psi A v}{\|Ay\|_2} \right\|_2 \right) \\ &\le \frac{1}{2}\left(\left\| \frac{\tilde\Psi A x}{\|Ax\|_2} \right\|_2 + \left\| \frac{\tilde\Psi A v}{\|Av\|_2} \right\|_2 \right) \cdot \frac{\|Ay\|_2 + \varepsilon}{\|Ay\|_2} \\ &\le \frac{1}{2}\left(\left\| \frac{\tilde\Psi A x}{\|Ax\|_2} \right\|_2 + \frac{1+\delta}{1-\delta} \right) \cdot \big(1 + \delta/6\big) \end{aligned}$$

where in the second line, we apply (37) and (33), and in the third line, (38). By the $\mathsf{RE}(s_0, k_0, A)$ condition and (32) we have

$$\begin{aligned} \|Ay\|_2 &\ge \frac{\|y_I\|_2}{K(s_0,k_0,A)} = \frac{\|x_I\|_2}{K(s_0,k_0,A)} \\ &\ge \frac{1}{K(s_0,k_0,A)\cdot\sqrt{k_0+1}}. \end{aligned}$$

Set

$$\varepsilon = \frac{\delta}{6\sqrt{1+k_0}K(s_0,k_0,A)}$$

$$\text{so that} \qquad \frac{\|Ay\|_2 + \varepsilon}{\|Ay\|_2} \le 1 + \frac{\delta}{6}.$$

Then for $\delta < 1/5$

$$\left\| \frac{\tilde\Psi A x}{\|Ax\|_2} \right\|_2 \ge 2\frac{1-\delta}{1+\delta/6} - (1+\delta)(1-\delta)^{-1} \ge 1 - 5\delta.$$

This verifies the lower estimate. It remains to check the bound for the cardinality of J. By (34) and (35), we have for $k_0 > 0$,

$$\begin{aligned} |J| &\le s_0 + m \le s_0 + \\ &\quad s_0 \max_{j\in\mathcal{M}} \|Ae_j\|_2^2 \left(\frac{16K^2(s_0,k_0,a)(3k_0)^2(k_0+1)}{\delta^2} \right) \\ &< d(3k_0, A) \end{aligned}$$

as desired. This completes the proof of Theorem 10. □

Remark 15: Let $\varepsilon > 0$. Instead of v defined in (36), one can consider the vector

$$v_\varepsilon = x_I + y - \varepsilon(x-y) \in \mathcal{C}\big(s_0, (1+\varepsilon)k_0\big).$$

Then replacing v by v_ε throughout the proof, we can establish Theorem 10 under the assumption $\mathsf{RE}(s_0, (1+\varepsilon)k_0, A)$ instead of $\mathsf{RE}(s_0, 3k_0, A)$, if we increase the dimension $d(3k_0)$ by a factor depending on ε.

III. Subgaussian random design

Theorem 6 can be reformulated as an almost isometry condition for the matrix $X = \Psi A$ acting on the set $\mathcal{C}(s_0, k_0)$. Recall that

$$d(3k_0, A) = s_0 + s_0 \max_j \|Ae_j\|_2^2 \left(\frac{16K^2(s_0,3k_0,A)(3k_0)^2(3k_0+1)}{\delta^2} \right).$$

Theorem 16: Set $0 < \delta < 1$, $0 < s_0 < p$, and $k_0 > 0$. Let A be a $q \times p$ matrix satisfying $\mathsf{RE}(s_0, 3k_0, A)$ condition as in Definition 1. Let $m = \min(d(3k_0, A), p) < p$. Let Ψ be an $n \times q$ matrix whose rows are independent isotropic ψ_2 random vectors in $\mathbb{R}^q$ with constant α. Assume that the sample size satisfies

$$n \ge \frac{2000 m\alpha^4}{\delta^2} \log\left(\frac{60ep}{m\delta}\right). \qquad (39)$$

Then with probability at least $1 - 2\exp(\delta^2 n/2000\alpha^4)$, for all $v \in \mathcal{C}(s_0, k_0)$ such that $v \ne 0$,

$$1-\delta \le \frac{1}{\sqrt{n}} \frac{\|\Psi A v\|_2}{\|Av\|_2} \le 1+\delta. \qquad (40)$$

Theorem 6 follows immediately from Theorem 16. Indeed, by (40), $\forall u \in \mathcal{C}(s_0, k_0)$ s.t. $u \ne 0$, we have

$$\left\| \frac{1}{\sqrt{n}} \Psi A u \right\|_2 \ge (1-\delta)\|Au\|_2 \ge (1-\delta)\frac{\|u_{T_0}\|_2}{K(s_0,k_0,A)} > 0.$$

To derive Theorem 16 from Theorem 10 we need a lower estimate for the norm of the image of a sparse vector. Such estimate relies on the standard ε-net argument similarly to [12, Section 3]. A complete proof of Theorem 16 appears in Section III-A.

Theorem 17: Set $0 < \delta < 1$. Let A be a $q \times p$ matrix, and let Ψ be an $n \times q$, matrix whose rows are independent isotropic ψ_2 random vectors in $\mathbb{R}^q$ with constant α. For $m \le p$, assume that

$$n \ge \frac{80 m\alpha^4}{\tau^2} \log\left(\frac{12ep}{m\tau}\right). \qquad (41)$$

Then with probability at least $1 - 2\exp(-\tau^2 n/80\alpha^4)$, for all m-sparse vectors u in $\mathbb{R}^p$,

$$(1-\tau)\|Au\|_2 \le \frac{1}{\sqrt{n}}\|\Psi A u\|_2 \le (1+\tau)\|Au\|_2. \qquad (42)$$

We note that Theorem 17 does not require the RE condition to hold. No particular upper bound on $\rho_{\max}(m, A)$ is imposed here either.

We now state a large deviation bound for m-sparse eigenvalues $\rho_{\min}(m, \tilde X)$ and $\rho_{\max}(m, \tilde X)$ for random design $\tilde X = n^{-1/2}\Psi A$ which follows from Theorem 17 directly.

Corollary 18: Suppose conditions in Theorem 17 hold. Then with probability $\ge 1 - 2\exp(-\tau^2 n/80\alpha^4)$,

$$\begin{aligned} (1-\tau)\sqrt{\rho_{\min}(m,A)} &\le \sqrt{\rho_{\min}(m,\tilde X)} \\ &\le \sqrt{\rho_{\max}(m,\tilde X)} \le (1+\tau)\sqrt{\rho_{\max}(m,A)}. \end{aligned}$$

A. Proof of Theorem 16

For n as bounded in (39), where $m = \min(d(3k_0, A), p)$, we have (41) holds with $\tau = \delta/5$. Then by Theorem 17, we have with probability at least $1 - 2\exp\left(-n\delta^2/(2000\alpha^4)\right)$,

$$\forall m\text{-sparse vectors } u,$$
$$\left(1 - \frac{\delta}{5}\right)\|Au\|_2 \le \frac{1}{\sqrt{n}}\left\|\tilde{\Psi}Au\right\|_2 \le \left(1 + \frac{\delta}{5}\right)\|Au\|_2 .$$

The proof finishes by application of Theorem 10. □

B. Proof of Theorem 17

We start with a definition.

Definition 19: Given a subset $U \subset \mathbb{R}^p$ and a number $\varepsilon > 0$, an ε-net Π of U with respect to the Euclidean metric is a subset of points of U such that ε-balls centered at Π covers U:

$$U \subset \bigcup_{x\in\Pi}(x + \varepsilon B_2^p),$$

where $A + B := \{a + b : a \in A, b \in B\}$ is the Minkowski sum of the sets A and B. The covering number $\mathcal{N}(U, \varepsilon)$ is the smallest cardinality of an ε-net of U.

The proof of Theorem 17 uses two well-known results. The first one is the *volumetric estimate*; see e.g. [31].

Lemma 20: Given $m \ge 1$ and $\varepsilon > 0$. There exists an ε-net $\Pi \subset B_2^m$ of B_2^m with respect to the Euclidean metric such that $B_2^m \subset (1-\varepsilon)^{-1}\operatorname{conv}\Pi$ and $|\Pi| \le (1+2/\varepsilon)^m$. Similarly, there exists an ε-net of the sphere S^{m-1}, $\Pi' \subset S^{m-1}$ such that $|\Pi'| \le (1 + 2/\varepsilon)^m$.

The second lemma with a worse constant can be derived from Bernstein's inequality for subexponential random variables. Since we are interested in the numerical value of the constant, we provide a proof below.

Lemma 21: Let $Y_1, \dots, Y_n$ be independent random variables such that $\mathbb{E}Y_j^2 = 1$ and $\|Y_j\|_{\psi_2} \le \alpha$ for all $j = 1, \dots, n$. Then for any $\theta \in (0, 1)$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n Y_j^2 - 1\right| > \theta\right) \le 2\exp\left(-\frac{\theta^2 n}{10\alpha^4}\right).$$

For a set $J \subset \{1, \dots, p\}$, denote $E_J = \operatorname{span}\{e_j : j \in J\}$, and set $F_J = AE_J$. For each subset $F_J \cap S^{q-1}$, construct an ε-net Π_J, which satisfies

$$\Pi_J \subset F_J \cap S^{q-1} \text{ and } |\Pi_J| \le (1 + 2/\varepsilon)^m.$$

The existence of such Π_J is guaranteed by Lemma 20. If

$$\Pi = \bigcup_{|J|=m}\Pi_J,$$

then the previous estimate implies

$$|\Pi| = (3/\varepsilon)^m\binom{p}{m} \le \left(\frac{3ep}{m\varepsilon}\right)^m = \exp\left(m\log\left(\frac{3ep}{m\varepsilon}\right)\right)$$

For $y \in S^{q-1} \cap F_J \subset F$, let $\pi(y)$ be one of the closest point in the ε-cover Π_J. Then

$$\frac{y - \pi(y)}{\|y - \pi(y)\|_2} \in F_J \cap S^{q-1} \text{ where } \|y - \pi(y)\|_2 \le \varepsilon.$$

Denote by $\Psi_1, \dots, \Psi_n$ the rows of the matrix Ψ, and set $\Gamma = n^{-1/2}\Psi$. Let $x \in S^{q-1}$. Applying Lemma 21 to the random variables $\langle \Psi_1, x\rangle^2, \dots, \langle \Psi_n, x\rangle^2$, we have that for every $\theta < 1$

$$\begin{aligned}\mathbb{P}\left(\left|\|\Gamma x\|_2^2 - 1\right| > \theta\right) &= \\ &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\langle \Psi_i, x\rangle^2 - 1\right| > \theta\right) \\ &\le 2\exp\left(-\frac{n\theta^2}{10\alpha^4}\right). \end{aligned} \tag{43}$$

For

$$n \ge \frac{20m\alpha^4}{\theta^2}\log\left(\frac{3ep}{m\varepsilon}\right),$$

the union bound implies

$$\begin{aligned}&\mathbb{P}\left(\exists x \in \Pi \text{ s. t. } \left|\|\Gamma x\|_2^2 - 1\right| > \theta\right) \\ &\le 2|\Pi|\exp\left(-\frac{n\theta^2}{10\alpha^4}\right) \\ &\le 2\exp\left(-\frac{n\theta^2}{20\alpha^4}\right)\end{aligned}$$

Then for all $y_0 \in \Pi$

$$\begin{aligned}&1 - \theta \le \|\Gamma y_0\|_2^2 \le 1 + \theta \text{ and so} \\ &1 - \theta \le \|\Gamma y_0\|_2 \le 1 + \frac{\theta}{2}\end{aligned}$$

with probability at least $1 - 2\exp\left(-\frac{n\theta^2}{20\alpha^4}\right)$, The bound over the entire $S^{q-1} \cap F_J$ is obtained by approximation. We have

$$\begin{aligned}\|\Gamma\pi(y)\|_2 - \|\Gamma(y - \pi(y))\|_2 &\le \|\Gamma y\|_2 \\ &\le \|\Gamma\pi(y)\|_2 + \|\Gamma(y - \pi(y))\|_2\end{aligned} \tag{44}$$

Define

$$\|\Gamma\|_{2,F_J} := \sup_{y\in S^{q-1}\cap F_J}\|\Gamma y\|_2 .$$

The RHS of (44) is upper bounded by $1 + \frac{\theta}{2} + \varepsilon\|\Gamma\|_{2,F_J}$. By taking the supremum over all $y \in S^{q-1} \cap F_J$, we have

$$\begin{aligned}\|\Gamma\|_{2,F_J} &\le 1 + \frac{\theta}{2} + \varepsilon\|\Gamma\|_{2,F_J} \\ \text{and hence } \|\Gamma\|_{2,F_J} &\le \frac{1 + \theta/2}{1 - \varepsilon}.\end{aligned}$$

The LHS of (44) is lower bounded by $1 - \theta - \varepsilon\|\Gamma\|_{2,F_J}$, and hence for all $y \in S^{q-1} \cap F_J$

$$\|\Gamma y\|_2 \ge 1 - \theta - \varepsilon\|\Gamma\|_{2,F_J} \ge 1 - \theta - \frac{\varepsilon(1 + \theta/2)}{1 - \varepsilon}$$

Putting these together, we have for all $y \in S^{q-1} \cap F_J$

$$1 - \theta - \frac{\varepsilon(1 + \theta/2)}{1 - \varepsilon} \le \|\Gamma y\|_2 \le \frac{1 + \theta/2}{1 - \varepsilon}$$

which holds for all sets J. Thus for $\theta < 1/2$ and $\varepsilon = \frac{\theta}{1+2\theta}$,

$$1 - 2\theta < \|\Gamma y\|_2 < 1 + 2\theta.$$

For any m-sparse vector $u \in S^{p-1}$

$$\frac{Au}{\|Au\|_2} \in F_J \quad \text{for } J = \operatorname{supp}(u),$$

and so

$$(1-2\theta)\left\|Au\right\|_2 \le \left\|\Gamma Au\right\|_2 \le (1+2\theta)\left\|Au\right\|_2.$$

Taking $\tau = \theta/2$ finishes the proof for Theorem 17.

C. Proof of Lemma 21

Note that $\alpha \ge \left\|Y_1\right\|_{\psi_2} \ge \left\|Y_1\right\|_2 = 1$. Using the elementary inequality $t^k \le k! s^k e^{t/s}$, which holds for all $t, s > 0$, we obtain

$$\begin{aligned} |\mathbb{E}(Y_j^2-1)^k| &\le \max(\mathbb{E}Y_j^{2k}, 1) \\ &\le \max(k!\alpha^{2k}\cdot \mathbb{E}e^{Y_j^2/\alpha^2}, 1) \\ &\le 2k!\alpha^{2k} \end{aligned}$$

for any $k \ge 2$. Since for any j $\mathbb{E}Y_j^2 = 1$, for any $\tau \in \mathbb{R}$ with $|\tau|\alpha^2 < 1$

$$\begin{aligned} \mathbb{E}\exp\left[\tau(Y_j^2-1)\right] &\le 1 + \sum_{k=2}\frac{1}{k!}|\tau|^k \cdot |\mathbb{E}(Y_j^2-1)^k| \\ &\le 1 + \sum_{k=2}|\tau|^k \cdot 2\alpha^{2k} \\ &\le 1 + \frac{2\tau^2\alpha^4}{1-|\tau|\alpha^2} \\ &\le \exp\left(\frac{2\tau^2\alpha^4}{1-|\tau|\alpha^2}\right). \end{aligned}$$

By Markov's inequality, for $\tau \in (0, \alpha^{-2})$

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n Y_j^2 - 1 > \theta\right) \\ &\le \mathbb{E}\exp\left(\tau\sum_{j=1}^n (Y_j^2-1) - \tau\theta n\right) \\ &= e^{-\tau\theta n}\cdot\left(\mathbb{E}\exp\left[\tau(Y^2-1)\right]\right)^n \\ &\le \exp\left(-\tau\theta n + \frac{2\tau^2\alpha^4 n}{1-|\tau|\alpha^2}\right). \end{aligned}$$

Set $\tau = \frac{\theta}{5\alpha^4}$, so $\tau\alpha^2 \le 1/5$. Then the previous inequality implies

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n Y_j^2 - 1 > \theta\right) \le \exp\left(-\frac{\theta^2 n}{10\alpha^4}\right).$$

Similarly, considering $\tau < 0$, we obtain

$$\mathbb{P}\left(1 - \frac{1}{n}\sum_{j=1}^n Y_j^2 > \theta\right) \le \exp\left(-\frac{\theta^2 n}{10\alpha^4}\right).$$

□

IV. RE condition for random matrices with bounded entries

We next consider the case of design matrix X consisting of independent identically distributed rows with bounded entries. As in the previous section, we reformulate Theorem 8 in the form of an almost isometry condition.

Theorem 22: Let $0 < \delta < 1$and $0 < s_0 < p$. Let $Y \in \mathbb{R}^p$ be a random vector such that $\left\|Y\right\|_\infty \le M$ a.s., and denote $\Sigma = \mathbb{E}YY^T$. Let X be an $n\times p$ matrix, whose rows $X_1, \dots, X_n$ are independent copies of Y. Let Σ satisfy the $\mathsf{RE}(s_0, 3k_0, \Sigma^{1/2})$ condition as in Definition 1. Set

$$d = d(3k_0, \Sigma^{1/2}) = s_0 + s_0 \max_j \left\|\Sigma^{1/2}e_j\right\|_2^2 \times \left(\frac{16K^2(s_0, 3k_0, \Sigma^{1/2})(3k_0)^2(3k_0+1)}{\delta^2}\right).$$

Assume that $d \le p$ and $\rho = \rho_{\min}(d, \Sigma^{1/2}) > 0$. If for some absolute constant C

$$n \ge \frac{CM^2 d\cdot\log p}{\rho\delta^2}\cdot\log^3\left(\frac{CM^2 d\cdot\log p}{\rho\delta^2}\right),$$

then with probability at least $1 - \exp\left(-\delta\rho n/(6M^2 d)\right)$ all vectors $u \in \mathcal{C}(s_0, k_0)$ satisfy

$$(1-\delta)\left\|u\right\|_2 \le \frac{\left\|Xu\right\|_2}{\sqrt{n}} \le (1+\delta)\left\|u\right\|_2.$$

Similarly to Theorem 16, Theorem 22 can be derived from Theorem 10, and the corresponding bound for d-sparse vector, which is proved below.

Theorem 23: Let $Y \in \mathbb{R}^p$ be a random vector such that $\left\|Y\right\|_\infty \le M$ a.s., and denote $\Sigma = \mathbb{E}YY^T$. Let X be an $n\times p$ matrix, whose rows $X_1, \dots, X_n$ are independent copies of Y. Let $0 < m \le p$. If $\rho = \rho_{\min}(m, \Sigma^{1/2}) > 0$ and

$$n \ge \frac{CM^2 m\cdot\log p}{\rho\delta^2}\cdot\log^3\left(\frac{CM^2 m\cdot\log p}{\rho\delta^2}\right), \tag{45}$$

then with probability at least $1 - 2\exp\left(-\frac{\varepsilon\rho n}{6M^2 m}\right)$ all m-sparse vectors u satisfy

$$1-\delta \le \frac{1}{\sqrt{n}}\cdot\left\|\frac{Xu}{\left\|\Sigma^{1/2}u\right\|_2}\right\|_2 \le 1+\delta.$$

To prove Theorem 23 we consider random variables $Z_u = \left\|Xu\right\|_2/(\sqrt{n}\left\|\Sigma^{1/2}u\right\|_2) - 1$, and estimate the expectation of the supremum of Z_u over the set of sparse vectors using Dudley's entropy integral. The proof of this part closely follows [9], so we will only sketch it. To derive the large deviation estimate from the bound on the expectation we use Talagrand's measure concentration theorem for empirical processes [29], which provides a sharper estimate, than the method used in [9].

Proof of Theorem 23. For $J \subset \{1, \dots, p\}$, let E_J be the coordinate subspace spanned by the vectors e_j, $j \in J$. Set

$$F = \bigcup_{|J|=m} \Sigma^{1/2}E_J \cap S^{p-1}.$$

Denote $\Psi = \Sigma^{-1/2}X$ so $\mathbb{E}\Psi\Psi^T = id$, and let $\Psi_1, \dots, \Psi_n$ be independent copies of Ψ. It is enough to show that with probability at least $1 - \exp\left(-\frac{\varepsilon\rho n}{6M^2 m}\right)$ for any $y \in F$

$$\left|1 - \frac{1}{n}\sum_{j=1}^n \langle\Psi_j, y\rangle^2\right| \le \delta.$$

To this end we estimate

$$\Delta := \mathbb{E}\sup_{y\in F}\left|1-\frac{1}{n}\sum_{j=1}^{n}\langle\Psi_j,y\rangle^2\right|.$$

The standard symmetrization argument implies that

$$\mathbb{E}\sup_{y\in F}\left|1-\frac{1}{n}\sum_{j=1}^{n}\langle\Psi_j,y\rangle^2\right| \le \frac{2}{n}\mathbb{E}\sup_{y\in F}\left|\sum_{j=1}^{n}\varepsilon_j\langle\Psi_j,y\rangle^2\right|,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are independent Bernoulli random variables taking values ±1 with probability $1/2$. The estimate of the last quantity is based on the following Lemma, which is similar to Lemma 3.6 [9].

Lemma 24: Let F be as above, and let $\psi_1,\dots,\psi_n\in\mathbb{R}^p$. Set

$$Q=\max_{j=1,\dots,n}\left\|\Sigma^{1/2}\psi_j\right\|_\infty.$$

Then

$$\mathbb{E}\sup_{y\in F}\left|\sum_{j=1}^{n}\varepsilon_j\langle\psi_j,y\rangle^2\right| \le \sqrt{\frac{CmQ^2\cdot\log n\cdot\log p}{\rho}}\times \log\left(\frac{CmQ^2}{\rho}\right)\cdot\sup_{y\in F}\left(\sum_{j=1}^{n}\langle\psi_j,y\rangle^2\right)^{1/2}.$$

Assuming Lemma 24, we finish the proof of the Theorem. First, note that by the definition of Ψ_j,

$$\max_{j=1,\dots,n}\left\|\Sigma^{1/2}\Psi_j\right\|_\infty \le M \text{ a.s.}$$

Hence, conditioning on $\Psi_1,\dots,\Psi_n$ and applying Lemma 24, we obtain

$$\Delta \le \frac{2}{n}\cdot\sqrt{\frac{CmM^2\cdot\log n\cdot\log p}{\rho}}\times \log\left(\frac{CmM^2}{\rho}\right)\cdot\mathbb{E}\sup_{y\in F}\left(\sum_{j=1}^{n}\langle\Psi_j,y\rangle^2\right)^{1/2},$$

and by Cauchy–Schwartz inequality,

$$\mathbb{E}\sup_{y\in F}\left(\sum_{j=1}^{n}\langle\Psi_j,y\rangle^2\right)^{1/2} \le \left(\mathbb{E}\sup_{y\in F}\sum_{j=1}^{n}\langle\Psi_j,y\rangle^2\right)^{1/2},$$

so

$$\Delta \le \frac{2}{\sqrt{n}}\cdot\sqrt{\frac{CmM^2\cdot\log n\cdot\log p}{\rho}}\times \log\left(\frac{CmM^2}{\rho}\right)\cdot(\Delta+1)^{1/2}.$$

If n satisfies (45), then

$$\Delta\le\delta\cdot(\Delta+1)^{1/2}, \text{and thus } \Delta\le 2\delta.$$

For $y\in F$ define a random variable $f(y)=\langle\Psi,y\rangle^2-1$. Then $|f(y)|\le\langle X,\Sigma^{-1/2}y\rangle^2+1\le M^2\rho^{-1}m+1:=a$ a.s., because $\Sigma^{-1/2}y$ is an m-sparse vector, whose norm does not exceed $\rho^{-1/2}$. Set

$$Z=\sup_{y\in F}\sum_{j=1}^{n}f_j(y),$$

where $f_1(y),\dots,f_n(y)$ are independent copies of $f(y)$. The argument above shows that $\mathbb{E}Z\le 2\delta n$. Then Talagrand's concentration inequality for empirical processes [29], [32] reads

$$\mathbb{P}(Z\ge t)\le\exp\left(-\frac{t}{6a}\right)\le\exp\left(-\frac{t\rho}{6M^2m}\right)$$

for all $t\ge 2\mathbb{E}Z$. Setting $t=4\delta n$, we have

$$\mathbb{P}(\sup_{y\in F}\sum_{j=1}^{n}\left(\langle\Psi_j,y\rangle^2-1\right)\ge 4\delta n)\le\exp\left(-\frac{4\delta n\rho}{6M^2m}\right).$$

Similarly, considering random variables $g(y)=1-\langle\Psi,y\rangle^2$, we show that

$$\mathbb{P}(\sup_{y\in F}\sum_{j=1}^{n}\left(1-\langle\Psi_j,y\rangle^2\right)\ge 4\delta n)\le\exp\left(-\frac{4\delta n\rho}{6M^2m}\right),$$

which completes the proof of the theorem. □

It remains to prove Lemma 24.

Proof of Lemma 24. By Dudley's inequality, see e.g. [33, Eq. (1.18)]

$$\mathbb{E}\sup_{y\in F}\left|\sum_{j=1}^{n}\varepsilon_j\langle\psi_j,y\rangle^2\right|\le C\int_0^\infty\log^{1/2}N(F,d,u)\,du$$

where $N(F,d,u)$ is the covering number, which is the minimal number of balls of radius u in the metric d covering the set F. Here d is the natural metric of the related Gaussian process defined as

$$\begin{aligned} d(x,y) &= \left[\sum_{j=1}^{n}\left(\langle\psi_j,x\rangle^2-\langle\psi_j,y\rangle^2\right)^2\right]^{1/2} \\ &\le \left[\sum_{j=1}^{n}\left(\langle\psi_j,x\rangle+\langle\psi_j,y\rangle\right)^2\right]^{1/2}\cdot\max_{j=1,\dots,n}|\langle\psi_j,x-y\rangle| \\ &\le 2R\cdot\|x-y\|_Y, \end{aligned}$$

where

$$\begin{aligned} R &= \sup_{y\in F}\left(\sum_{j=1}^{n}\langle\psi_j,y\rangle^2\right)^{1/2}, \\ \text{and } \|z\|_Y &= \max_{j=1,\dots,n}|\langle\psi_j,z\rangle|. \end{aligned}$$

The inclusion $\sqrt{m}B_1^p\supset\bigcup_{|J|=m}E_J\cap S^{p-1}$ implies

$$\sqrt{m}\Sigma^{1/2}B_1^p\supset\Sigma^{1/2}\operatorname{conv}\Big(\bigcup_{|J|=m}E_J\cap S^{p-1}\Big)\supset\rho^{1/2}F.$$

Hence, for any $y\in F$

$$\begin{aligned} \|z\|_Y &\le \rho^{-1/2}\sqrt{m}\max_{j=1,\dots,n}\left\|\Sigma^{1/2}\psi_j\right\|_\infty \\ &= \rho^{-1/2}\sqrt{m}Q. \end{aligned} \tag{46}$$

Replacing the metric d with the norm $\|\cdot\|_Y$, we obtain

$$\mathbb{E}\sup_{y\in F}\left|\sum_{j=1}^n \varepsilon_j\langle\psi_j,y\rangle^2\right| \leq CR\int_0^{\rho^{-1/2}\sqrt{m}Q}\log^{1/2} N(F,\|\cdot\|_Y,u)\,du.$$

The upper limit of integration is greater or equal than the diameter of F in the norm $\|\cdot\|_Y$, so for $u>\rho^{-1/2}\sqrt{m}Q$ the integrand is 0. Arguing as in Lemma 3.7 [9], we can show that

$$\begin{aligned} N(F,\|\cdot\|_Y,u) &\leq N(\rho^{-1/2}\sqrt{m}\Sigma^{1/2}B_1^p,\|\cdot\|_Y,u) \\ &\leq (2p)^l, \end{aligned} \tag{47}$$

where

$$\begin{aligned} l &= C\rho^{-1}m\times \\ &\frac{\left(\max_{i=1,\dots,p}\max_{j=1,\dots,n}|\langle\Sigma^{1/2}e_i,\psi_j\rangle|\right)^2}{u^2}\cdot\log n \\ &= \frac{CmQ^2\cdot\log n}{\rho u^2} \end{aligned}$$

Also, since F consists of the union $\binom{p}{m}$ Euclidean spheres, the inclusion (46) and the volumetric estimate yield

$$\begin{aligned} N(F,\|\cdot\|_Y,u) &\leq \binom{p}{m}\cdot\left(1+\frac{2\rho^{-1/2}\sqrt{m}Q}{u}\right)^m \\ &\leq \left(\frac{ep}{m}\right)^m\cdot\left(1+\frac{2\rho^{-1/2}\sqrt{m}Q}{u}\right)^m. \end{aligned} \tag{48}$$

Estimating the covering number of F as in (47) for $u\geq 1$, and as in (48) for $0<u<1$, we obtain

$$\begin{aligned} \mathbb{E}\sup_{y\in F}\left|\sum_{j=1}^n \varepsilon_j\langle\psi_j,y\rangle^2\right| &\leq CR\times \\ &\int_0^1\sqrt{m}\cdot\sqrt{\log\left(\frac{ep}{m}\right)+\log\left(1+\frac{2\rho^{-1/2}\sqrt{m}Q}{u}\right)}\,du \\ &+CR\int_1^{\rho^{-1/2}\sqrt{m}Q}\sqrt{\frac{CmQ^2\cdot\log n}{\rho u^2}}\cdot\sqrt{\log 2p}\,du \\ \leq\ & CR\sqrt{\frac{mQ^2\cdot\log n\cdot\log p}{\rho}}\cdot\log\left(\frac{CmQ^2}{\rho}\right). \end{aligned}$$

□

Remark 25: Note that unlike the case of a random matrix with subgaussian marginals, the estimate of Theorem 23 contains the minimal sparse singular value ρ. This is, however, necessary, as the following example shows.

Let $m=2^l$, and assume that $p=k\cdot m$, for some $k\in\mathbb{N}$. For $j=1,\dots,k$ let D_j be the $m\times m$ Walsh matrix. Let A be a $p\times p$ block-diagonal matrix with blocks $D_1,\dots,D_k$ on the diagonal, and let $Y\in\mathbb{R}^p$ be a random vector, whose values are the rows of the matrix A taken with probabilities $1/p$. Then $\|Y\|_\infty=1$ and $\mathbb{E}YY^T=(m/p)\cdot id$, so $\rho=m/p$. Hence, the right-hand side of (45) reduces to

$$\frac{Cp\cdot\log p}{\delta^2}\cdot\log^3\left(\frac{Cp\cdot\log p}{\delta^2}\right)$$

From the other side, if the matrix X satisfies the conditions of Theorem 23 with, say, $\delta=1/2$, then all rows of the matrix A should be present among the rows of the matrix X. An elementary calculation shows that in this case it is necessary to assume that $n\geq Cp\log p$, so the estimate (45) is exact up to a power of the logarithm.

Unlike the matrix Σ, the matrix A is not symmetric. However, the example above can be easily modified by considering a $2p\times 2p$ matrix

$$\tilde{A}=\begin{pmatrix}0 & A\\ A^T & 0\end{pmatrix}.$$

This shows that the estimate (45) is tight under the symmetry assumption as well.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their very careful reading and helpful comments.

REFERENCES

[1] E. Candès and T. Tao, "Decoding by Linear Programming," *IEEE Trans. Info. Theory*, vol. 51, pp. 4203–4215, 2005.
[2] ——, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5406–5425, December 2006.
[3] ——, "The Dantzig selector: statistical estimation when p is much larger than n." *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.
[4] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific and Statistical Computing*, vol. 20, pp. 33–61, 1998.
[5] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
[6] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications in Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, August 2006.
[7] D. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Communications in Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
[8] M. Rudelson and R. Vershynin, "Sparse reconstruction by convex relaxation: Fourier and gaussian measurements," in *40th Annual Conference on Information Sciences and Systems (CISS 2006)*, March 2006, pp. 207–212.
[9] ——, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, pp. 1025–1045, 2008.
[10] ——, "Geometric approach to error correcting codes and reconstruction of signals," *International Mathematical Research Notices*, vol. 64, pp. 4019–4041, 2005.
[11] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. B. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.
[12] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008.
[13] R. Adamczak, A. E. Litvak, , A. Pajor, and N. Tomczak-Jaegermann, "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," 2009, 0904.4723v1.
[14] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.
[15] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[16] S. van de Geer and P. Buhlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.
[17] G. Raskutti, M. Wainwright, and B. Yu, "Restricted nullspace and eigenvalue properties for correlated gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.
[18] S. Zhou, "Restricted eigenvalue conditions on subgaussian random matrices," 2009, unpublished Manuscript. Available at http://arxiv.org/pdf/0912.4045v2.pdf.

[19] Y. Gordon, "Some inequalities for gaussian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.
[20] S. Zhou, S. van de Geer, and P. Bühlmann, "Adaptive Lasso for high dimensional regression and gaussian graphical modeling," 2009, arXiv:0903.2515.
[21] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and subgaussian operators in asymptotic geometric analysis," *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.
[22] S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann, "High-dimensional covariance estimation based on Gaussian graphical models," *Journal of Machine Learning Research*, vol. 12, pp. 2975–3026, 2011.
[23] S. Zhou, J. Lafferty, and L. Wasserman, "Compressed and privacy sensitive sparse regression," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 846–866, 2009.
[24] G. Duncan and R. Pearson, "Enhancing access to microdata while protecting confidentiality: Prospects for the future," *Statistical Science*, vol. 6, no. 3, pp. 219–232, August 1991.
[25] R. Vershynin, "Approximating the moments of marginals of high dimensional distributions," *Annals of Probability*, vol. 39, pp. 1591–1606, 2011.
[26] ——, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability, to appear*, 2011.
[27] R. Adamczak, R. Latala, A. E. Litvak, , A. Pajor, and N. Tomczak-Jaegermann, "Geometry of log-concave ensembles of random matrices and approximate reconstruction," 2011, 1103.0401v1.
[28] G. Pisier, "Remarques sur un résultat non publié de b. Maurey," *Seminar on Functional Analysis, École Polytech., Palaiseau*, vol. 5, 1981.
[29] M. Talagrand, "New concentration inequalities in product spaces," *Invent. Math.*, vol. 126, pp. 505–563, 1996.
[30] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and processes*. Springer, 1991.
[31] V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Mathematics 1200*. Springer, 1986.
[32] M. Ledoux, *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, 89. American Mathematical Society, 2001.
[33] M. Talagrand, *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer, 2000.

Mark Rudelson received Ph.D. in Mathematics from the Hebrew University of Jerusalem, Israel, in 1997. He is a Professor in the Department of Mathematics at the University of Michigan, Ann Arbor, since 2010. His research interests include random matrix theory and geometric functional analysis, as well as applications to computer science.

Shuheng Zhou is currently an assistant Professor in the Department of Statistics, with a courtesy appointment with the Department of Electrical Engineering and Computer Sciences at the University of Michigan, Ann Arbor. She received Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania. Her research interests include statistical machine learning, high dimensional statistics, and theoretical computer science, in particular, convex optimization, privacy, approximation and randomized algorithms, and network and combinatorial optimization.