Sparse recovery under weak moment assumptions

Guillaume Lecué[1,3] Shahar Mendelson[2,4,5]

June 4, 2014

Abstract

We prove that iid random vectors that satisfy a rather weak moment assumption can be used as measurement vectors in Compressed Sensing, and the number of measurements required for exact reconstruction is the same as the best possible estimate – exhibited by a random gaussian matrix. We also prove that this moment condition is necessary, up to a $\log\log$ factor. Applications to the Compatibility Condition and the Restricted Eigenvalue Condition in the noisy setup and to properties of neighbourly random polytopes are also discussed.

1 Introduction and main results

In Compressed Sensing (see, e.g., [8] and [13]), one observes linear measurements $y_i = \langle X_i, x_0 \rangle$, $i = 1, ..., N$ of an unknown vector $x_0 \in \mathbb{R}^n$, and the goal is to identify x_0 using those measurements.

Let $(e_1, \ldots, e_N)$ be the canonical basis of $\mathbb{R}^N$. Given the measurements matrix $\Gamma = N^{-1/2} \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$, a possible recovery procedure is the basis pursuit algorithm, defined by

$$\hat{x} \in \operatorname{argmin}\big(\|t\|_1 : \Gamma t = \Gamma x_0 \big).$$

A well known question is to identify conditions on the vectors $X_1, ..., X_N$ that ensure that if x_0 is s-sparse, that is, if it is supported on at most s coordinates, the unique minimizer of the basis pursuit algorithm is x_0 itself.

[1]CNRS, CMAP, Ecole Polytechnique, 91120 Palaiseau, France.
[2]Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.
[3]Email: guillaume.lecue@cmap.polytechnique.fr
[4]Email: shahar@tx.technion.ac.il
[5]Supported by the Mathematical Sciences Institute – The Australian National University and by ISF grant 900/10.

Definition 1.1 *A matrix $\Gamma \in \mathbb{R}^{N\times n}$ satisfies the* ***exact reconstruction property of order*** *s if for every $x_0 \in \Sigma_s$, the set of all $s-$sparse vectors in $\mathbb{R}^n$,*

$$\operatorname{argmin}\big(\|t\|_1 : \Gamma t = \Gamma x_0\big) = \{x_0\}. \tag{ER(s)}$$

It follows from Proposition 2.2.18 in [11] that if Γ satisfies ER(s) then $N \gtrsim s\log\big(en/s\big)$. And, there are constructions of random matrices that satisfy ER(s) with high probability and with the optimal number of measurements (rows) $N \sim s\log\big(en/s\big)$.

A typical example of such a matrix is the gaussian matrix, that has independent standard normal random variables as entries. Other examples are measurements matrices Γ when $X_1, ..., X_N$ are independent, isotropic and L-subgaussian random vectors.

Definition 1.2 *A random vector X in $\mathbb{R}^n$ is isotropic if for every $t \in \mathbb{R}^n$, $\mathbb{E}\big\langle X,t\big\rangle^2 = \|t\|_2^2$, and it is L-subgaussian if for every $t \in \mathbb{R}^n$ and every $p \geq 2$, $\|\big\langle X,t\big\rangle\|_{L_p} \leq L\sqrt{p}\|\big\langle X,t\big\rangle\|_{L_2}$.*

The reason behind this result, and many others like it, is that isotropic subgaussian matrices act on Σ_s in an isomorphic way with high probability, when $N \gtrsim s\log(en/s)$. In Compressed Sensing literature this isomorphic behaviour is called the *Restricted isometry property* (RIP) (see, for example [7, 9, 25]). A matrix Γ satisfies the RIP in Σ_s if for every $t \in \Sigma_s$,

$$(1-\delta)\|t\|_2 \leq \|\Gamma t\|_2 \leq (1+\delta)\|t\|_2, \tag{1.1}$$

for some fixed $0 < \delta < 1$. It is straightforward to show that if RIP holds in Σ_s then ER($c_0 s$) is satisfied where c_0 is an absolute constant (cf. [8, 9, 6]).

Proving the RIP for subgaussian matrices uses the fact that tails of linear functionals $\big\langle X,t\big\rangle$ decay faster than the corresponding gaussian variable. Thus, it seemed natural to ask whether the same type of estimates hold in cases in which linear functionals exhibit a slower decay – for example, when X is sub-exponential, and the linear functionals satisfy that $\|\big\langle X,t\big\rangle\|_{L_p} \leq Lp\|\big\langle X,t\big\rangle\|_{L_2}$ for every $t \in \mathbb{R}^n$ and every $p \geq 2$.

Proving the RIP for a sub-exponential random ensemble is a much harder task than for subgaussian ensembles (see, e.g. [1]). Moreover, the RIP does not exhibit the same behaviour as in the gaussian case. Indeed, one may show that for sub-exponential ensembles, the RIP holds with high probability only when $N \gtrsim s\log^2(en/s)$, and this estimate is optimal as can be seen when X has independent, symmetric exponential random variables as coordinates [1].

On the other hand, the result in [18] (see Chapter 7 there) shows that exact reconstruction can still be achieved by isotropic sub-exponential measurement vectors when $N \gtrsim s\log(en/s)$ – the same number of measurements needed for the gaussian ensemble.

Clearly, this estimate cannot be based on the RIP, and one may ask whether weaker assumptions on the measurement vectors may still lead to exact recovery with the optimal number of measurements. The main result presented here is in this direction. It uses the following condition that was recently used in [23] and [24].

Definition 1.3 *A random vector X in $\mathbb{R}^n$ satisfies the* ***small ball property in the set*** *Σ_s if there exist $u, \beta > 0$ for which, for any s-sparse vector $t \in \mathbb{R}^n$, $P\big(|\langle X, t\rangle| > u\,\|t\|_2\big) \geq \beta$.*

The small ball property is a rather minimal assumption on the measurement vector and is satisfied in fairly general situations. For example, if one of the following conditions holds then X satisfies the small ball property with constants that depend only on κ_0 (and on ε for the first condition) (see [23]):

1. X is isotropic and for every $t \in \Sigma_s$, $\big\|\langle X, t\rangle\big\|_{L_{2+\varepsilon}} \leq \kappa_0 \big\|\langle X, t\rangle\big\|_{L_2}$, for some $\varepsilon > 0$;
2. X is isotropic and for every $t \in \Sigma_s$, $\big\|\langle X, t\rangle\big\|_{L_2} \leq \kappa_0 \big\|\langle X, t\rangle\big\|_{L_1}$;
3. $x_1, \ldots, x_n$ are n independent, real valued random variables that are absolutely continuous with respect to the Lebesgue measure and with almost surely bounded densities by κ_0 and $X = (x_1, \ldots, x_n)$.

Our first result shows that the small ball condition of Definition 1.3 and a weak moment assumption suffices to ensure the exact reconstruction property with the optimal number of measurements.

Theorem A. *There exist absolute constants c_0, c_1 and c_2 and for every $\alpha \geq 1/2$ there exists a constant $c_3(\alpha)$ that depends only on α for which the following holds. Let $X = (x_i)_{i=1}^n$ be a random vector on $\mathbb{R}^n$ such that*

1. *There are $\kappa_1, \kappa_2, w > 1$ that satisfy that for every $1 \leq j \leq n$, $\|x_j\|_{L_2} = 1$ and, for $p = \kappa_2 \log(wn)$, $\|x_j\|_{L_p} \leq \kappa_1 p^\alpha$.*
2. *The small ball property in Σ_s is satisfied by X.*

If $s_1 = c_2 u^2 \beta s$,

$$N \geq c_0 \max\left\{ s \log\left(\frac{en}{s}\right), (c_3(\alpha)\kappa_1)^2 (\kappa_2 \log(wn))^{\max\{2\alpha-1,1\}} \right\}$$

and $X_1, ..., X_N$ *are independent copies of* X*, then, with probability at least* $1-2\exp(-c_1\beta^2 N)-1/w^{\kappa_2} n^{\kappa_2-1}$, $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$ *satisfies the exact reconstruction property in* Σ_{s_1}.

For example, it follows from Theorem A that a random matrix with iid centered entries that have variance 1 and an L_p moment bounded by p for $p = 2\log n$ can be used as a measurement matrix, and just as in the gaussian case, requires only $N \gtrsim s\log(en/s)$ measurements.

Another straightforward application of Theorem A is for measurement vectors that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$ with a bounded density and that have almost surely bounded coordinates. The first condition implies the small ball assumption while the second implies the moment condition. And, it is relatively simple to derive many other results of a similar flavour of random ensembles that allow exact reconstruction with the optimal number of measurements.

Just as noted for sub-exponential ensembles, Theorem A cannot be proved using the RIP, and its proof must take a different path. A key ingredient in the proof is the following observation:

Theorem B. *Let* $\Gamma : \mathbb{R}^n \mapsto \mathbb{R}^N$ *and denote by* $(e_1, \ldots, e_n)$ *the canonical basis of* $\mathbb{R}^n$*. Assume that*

a) for every s*-sparse vector* x, $\|\Gamma x\|_2 \geq c_0 \|x\|_2$, *and*

b) for every $i \in \{1, \ldots, n\}$, $\|\Gamma e_i\|_2 \leq c_1$.

Then, for $s_1 = \lfloor (c_0^2(s-1))/(4c_1^2) \rfloor - 1$, Γ *satisfies the Exact Reconstruction property in* Σ_{s_1}.

Compared with the RIP, conditions in Theorem B are weaker, as it suffices to verify the right-hand side of the RIP in (1.1) just for 1-sparse vectors and not for all s-sparse vectors. This is a substantial difference, as the fact that for every $t \in \Sigma_s$, $\|\Gamma t\|_2 \leq (1+\delta)\|t\|_2$ is a costly one, and is the reason for the gap between RIP and the Exact Reconstruction property. While the lower bound holds almost for free (see [23], [24] and the next section) and requires the small-ball assumption, the upper bound is almost

equivalent to having a subgaussian behaviour of moments, at least up to some level. Even the fact that one has to verify the upper bound only for 1-sparse vectors comes at a cost, and as will be explained later, is the reason for a *price to pay* for convex relaxation.

To illustrate this fact, the next question we would like to address is whether the moment condition in Theorem A, that the coordinates of X should have $\log(n)$ moments, is really needed for the ER property. Our second result shows that this condition is almost necessary.

We say that a random matrix Γ is generated by the random variable x if $\Gamma = N^{-1/2}\sum_{i=1}^{N}\langle X_i, \cdot\rangle e_i$, where $X_1, ..., X_N$ are independent copies of the random vector $X = (x_1, ..., x_n)^\top$ whose coordinates are independent copies of x.

Theorem C. *There exist absolute constants c_0, c_1, c_2 and c_3 for which the following holds. Given $n \geq c_0$ and $N \log N \leq c_1 n$, there exists a mean-zero, variance one random variable x that satisfies $\|x\|_{L_4} \leq c_2$, $\|x\|_{L_p} \leq c_2\sqrt{p}$ for $p = c_3(\log n)/(\log N)$, and if Γ is the $N \times n$ matrix generated by x then with probability larger than $1/2$, Γ does not satisfy the exact reconstruction property of order 1.*

Note that if Γ is generated by x that satisfies $\|x\|_{L_2} = 1$, $\|x\|_{L_4} \leq c_2$ and $\|x\|_{L_p} \leq c_2\sqrt{p}$ for $p \sim \log n$, then, it follows from Theorem A that for $N \sim \log n$, Γ satisfies ER(1) with high probability. On the other hand, the random ensemble from Theorem C is generated by x for which $\|x\|_{L_2} = 1$, $\|x\|_{L_4} \leq c_2$ and $\|x\|_{L_p} \leq c_2\sqrt{p}$ for $p \sim (\log n)/\log\log n$, but still does not satisfy ER(1) with probability at least $1/2$ when $N \sim \log n$. Therefore, in the case $s = 1$, a subgaussian moment assumption for $p \sim \log n$ is a sharp condition for exact recovery by the basis pursuit algorithm with an optimal number of measurements (up to a $\log\log n$ factor).

An alternative formulation of Theorem C is the following:

Theorem C′. *If $n \geq c_0$ and $p > 2$, there exists a mean-zero and variance 1 random variable x, for which $\|x\|_{L_4} \leq \kappa$ and $\|x\|_{L_p} \leq \kappa\sqrt{p}$, and with probability at least $1/2$, if $N \lesssim \sqrt{p}n^{1/p}$, Γ does not satisfy the exact reconstruction property of order 1.*

Theorem C implies that Basis Pursuit may perform badly when the coordinates of X do not have enough moments. This is not the case for the ℓ_0-minimization procedure. Recall that ℓ_0-minimization is defined by

$\min\big(\|t\|_0 : \Gamma t = \Gamma x_0\big)$, where $\|t\|_0$ is the cardinality of the support of t. In the following result, we show that ℓ_0-minimization can reconstruct any s-sparse vector from only $N \sim s\log(en/s)$ measurements solely under the small-ball assumption.

Theorem D. *There exists absolute constants c_0, c_1 and c_2 for which the following holds. Let X be a random vector in $\mathbb{R}^n$ that satisfies the small ball property in Σ_s as in Definition 1.3. Let $X_1, \dots, X_N$ be N independent copies of X and set $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$. If $N \geq c_0 s\log(en/s)$ then with probability larger than $1 - c_1\exp(-c_2 N)$, given any $\lfloor s/2\rfloor$-sparse vector x_0, the only $t \in \Sigma_{\lfloor s/2\rfloor}$ for which $\Gamma t = \Gamma x_0$ is x_0 itself.*

Theorem C together with Theorem D exhibit that *a price has to be paid for convex relaxation* (basis pursuit being the convex relaxation of the ℓ_0-minimization procedure). Indeed, no moment assumption is needed for exact reconstruction via ℓ_0-minimization (this follows from Theorem D) whereas at least $(\log n)/(\log N)$ moments are required to get the same property for Basis Pursuit (as indicated in Theorem C).

The price of convex relation can also be seen through the number of measurements needed for exact reconstruction for the matrix Γ constructed in Theorem C. Indeed, under the assumptions of Theorem C$'$ for, say $p = 4$, the random vector $X = (x_1, ..., x_n)^\top$ satisfies the conditions of Theorem D. Therefore, one requires only $N \sim s\log(en/s)$ random measurements using independent copies of X to identify any s-sparse vector using ℓ_0-minimization. In contrast, Basis Pursuit requires at least $\sim n^{1/4}$ measurements to reconstruct 1-sparse vectors.

We end this introduction with a word about notation and organization of the article. In the next section, we prove Theorem A, B and D and in Section 3 we prove Theorem C and C$'$. The final section is devoted to results in a noisy setup that is the natural extension of Compressed Sensing. In particular, we prove the so-called Compatibility Condition and Restricted Eigenvalue Condition for the measurement matrix under weak moment assumptions and study random neighbourly polytopes.

Throughout, absolute constants or constants that depend on other parameters are denoted by c, C, c_1, c_2, etc., (and, of course, we will specify when a constant is absolute and when it depends on other parameters). The values of these constants may change from line to line. The notation $x \sim y$ (resp. $x \lesssim y$) means that there exist absolute constants $0 < c < C$ for which $cy \leq x \leq Cy$ (resp. $x \leq Cy$). If $b > 0$ is a parameter then $x \lesssim_b y$ means that $x \leq C(b)y$ for some constant $C(b)$ that depends only on b.

Let ℓ_p^n be $\mathbb{R}^n$ endowed with the norm $\|x\|_{\ell_p^n} = \left(\sum_j |x_j|^p\right)^{1/p}$. The unit ball there is denoted by B_p^n and the unit Euclidean sphere in $\mathbb{R}^n$ is S^{n-1}. If $A \subset \mathbb{R}^n$ then $\mathbb{1}_A$ denotes the indicator function of A.

2 Proof of Theorem A, B and D

The proof of Theorem A has several components, and although the first of them is rather standard, we present it for the sake of completeness.

Lemma 2.1 *Let $\Gamma : \mathbb{R}^n \to \mathbb{R}^N$ be a given matrix and fix $0 < r < 1$. If $B_1^n \cap rS^{n-1}$ does not intersect the kernel $\ker(\Gamma)$, then Γ satisfies the exact reconstruction property in $\Sigma_{\lfloor (2r)^{-2} \rfloor}$.*

Proof. Observe that if $x \in B_1^n$ and $\|x\|_2 \geq r$ then $y = rx/\|x\|_2 \in B_1^n \cap rS^{n-1}$. Thus, $\Gamma y \neq 0$ implies that $\Gamma x \neq 0$ – and therefore,

$$\sup_{x \in B_1^n \cap \ker(\Gamma)} \|x\|_2 < r.$$

Let $s = \lfloor (2r)^{-2} \rfloor$, fix $x_0 \in \Sigma_s$ and put I to be the set of coordinates on which x_0 is supported. Given a nonzero $h \in \ker(\Gamma)$, let $h = h_I + h_{I^c}$ - the decomposition of h to coordinates in I and in I^c. Since $h/\|h\|_1 \in B_1^n \cap \ker(\Gamma)$ then $\|h\|_2 \leq r\|h\|_1$, and by the choice of s, $2\sqrt{s}\|h\|_2 \leq \|h\|_1$. Therefore,

$$\begin{aligned}\|x_0 + h\|_1 &= \|x_0 + h_I\|_1 + \|h_{I^c}\|_1 \geq \|x_0\|_1 - \|h_I\|_1 + \|h_{I^c}\|_1 \\ &= \|x_0\|_1 - 2\|h_I\|_1 + \|h\|_1 \geq \|x_0\|_1 - 2\sqrt{|I|}\|h_I\|_2 + \|h\|_1 > \|x_0\|_1.\end{aligned}$$

Hence, $\|x_0+h\|_1 > \|x_0\|_1$ and x_0 is the unique minimizer of the basis pursuit algorithm, proving exact reconstruction. ∎

The main component in the proof of Theorem A is a uniform empirical small-ball estimate, following the same lines as the results in [24] and [23].

Definition 2.2 *Let $\mathcal{G}$ be a class of $\{0,1\}$-valued functions defined on a space $\mathcal{X}$. The set $\mathcal{G}$ is a VC-class if there exists an integer V for which, given any points $x_1, ..., x_{V+1} \in \mathcal{X}$,*

$$|\{(g(x_1), ..., g(x_{V+1})) : g \in \mathcal{G}\}| < 2^{V+1}. \tag{2.1}$$

The VC-dimension of $\mathcal{G}$, denoted by $VC(\mathcal{G})$, is the smallest integer V for which (2.1) *holds.*

Lemma 2.3 *There exists absolute constants c_1 and c_2 for which the following holds. Let $\mathcal{F}$ be a class of functions and assume that there is β and u for which*

$$\inf_{f\in\mathcal{F}} P\big(|f|>u\big)\geq\beta.$$

Let $\mathcal{G}_u=\big\{\mathbb{1}_{\{|f|>u\}}:f\in\mathcal{F}\big\}$. If $VC(\mathcal{G}_u)\leq d$ and $N\geq c_1 d/\beta^2$ then with probability at least $1-\exp(-c_2\beta^2 N)$,

$$\inf_{f\in\mathcal{F}}\big|\big\{i\in\{1,\ldots,N\}:|f(X_i)|>u\big\}\big|\geq\frac{\beta N}{2}.$$

Proof. Let $H(X_1,...,X_N)=\sup_{g\in\mathcal{G}_u}|N^{-1}\sum_{i=1}^N g(X_i)-\mathbb{E}g(X)|$. By the bounded differences inequality (see, for example, Theorem 6.2 in [4]), with probability at least $1-\exp(-t)$,

$$H(X_1,...,X_N)\leq\mathbb{E}H(X_1,...,X_N)+c_1\sqrt{\frac{t}{N}}.$$

Since $VC(\mathcal{G}_u)\leq d$, then by standard empirical processes arguments (symmetrization, the fact that Bernoulli processes are subgaussian and entropy estimates - see, for example [33]),

$$\mathbb{E}H(X_1,...,X_N)\leq c_2\sqrt{\frac{d}{N}}\leq\beta/4, \tag{2.2}$$

provided that $N\gtrsim d/\beta^2$. Therefore, taking $t=N\beta^2/16c_1^2$, then with probability at least $1-\exp(-c_3\beta^2 N)$, for every $f\in\mathcal{F}$,

$$\frac{1}{N}\sum_{i=1}^N\mathbb{1}_{\{|f|>u\}}(X_i)\geq P\big(|f|>u\big)-\frac{\beta}{2}\geq\frac{\beta}{2},$$

and on that event, $|\{i:|f(X_i)|>u\}|\geq\beta N/2$ for every $f\in\mathcal{F}$. ■

Corollary 2.4 *There exist absolute constants c_1 and c_2 for which the following holds. Let X be a random vector on $\mathbb{R}^n$.*

1. *If there are $\beta,u>0$ such that $P\big(|\langle t,X\rangle|>u\big)\geq\beta$ for any $t\in S^{n-1}$ and if $N\geq c_1 n/\beta^2$, then with probability at least $1-\exp(-c_2 N\beta^2)$,*

$$\inf_{t\in S^{n-1}}\frac{1}{N}\sum_{i=1}^N\langle X_i,t\rangle^2\geq\frac{u^2\beta}{2}.$$

2. *If there are $\beta, u > 0$ such that $P\big(|\langle t, X\rangle| > u\big) \geq \beta$ for any $t \in \Sigma_s \cap S^{n-1}$ and if $N \geq c_1 s \log(en/s)/\beta^2$, then with probability at least $1 - \exp(-c_2 N\beta^2)$,*

$$\inf_{t\in\Sigma_s\cap S^{n-1}} \frac{1}{N}\sum_{i=1}^{N}\langle X_i, t\rangle^2 \geq \frac{u^2\beta}{2}.$$

Remark 2.5 *Note that the first part of Corollary 2.4 gives an estimate on the smallest singular value of the random matrix $\Gamma = N^{-1/2}\sum_{i=1}^{N}\langle X_i, \cdot\rangle e_i$ along the lines of the estimate from [24], but without any assumption on the covariance structure of X, which is used in [24] or [29]. Note that the covariance matrix does not even have to exist to apply Corollary 2.4. It appears that the only small ball assumption over S^{n-1} is enough to prove that with large probability the smallest singular value of a random matrix is larger than a constant.*

Proof of Corollary 2.4. To prove the first part of the claim, let $\mathcal{F} = \{\langle t, \cdot\rangle : t \in S^{n-1}\}$. Recall that the VC dimension of a class of half-spaces in $\mathbb{R}^n$ is at most n, and thus, one may verify that for every u, the VC dimension of

$$\mathcal{G}_u = \{\mathbb{1}_{\{|f|>u\}} : f \in \mathcal{F}\}$$

is at most $c_1 n$ for a suitable absolute constant c_1 (see, e.g., Chapter 2.6 in [33]). The claim now follows immediately from Lemma 2.3; indeed, for every $t \in S^{n-1}$,

$$\frac{1}{N}\sum_{i=1}^{N}\langle t, X_i\rangle^2 \geq \frac{u^2}{N}|\{i : |\langle X_i, t\rangle| > u\}|.$$

Turning to the second part, note that $\Sigma_s \cap S^{n-1}$ is a union of $\binom{n}{s}$ spheres of dimension s. Applying the first part to each one of the spheres, combined with the union bound, it follows that if $N \geq c_2\beta^{-2}s\log(en/s)$, then with probability at least $1 - \exp(-c_3 N\beta^2)$,

$$\inf_{t\in\Sigma_s\cap S^{n-1}} \frac{1}{N}\sum_{i=1}^{N}\langle X_i, t\rangle^2 \geq \frac{u^2\beta}{2}.$$

■

Corollary 2.4 shows that the weak small-ball assumption for linear functionals implies that Γ 'acts well' on s-sparse vectors. However, according to Lemma 2.1, exact recovery follows if one can show that it also acts well

on the larger set $\sqrt{r}B_1^n \cap S^{n-1}$ for a well chosen r that is proportional to s. Therefore, one has to use information on the way a matrix acts on Σ_s to study the way it acts on the set

$$\sqrt{\kappa_0 s}B_1^n \cap S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_1 \leq \sqrt{\kappa_0 s}, \quad \|x\|_2 = 1\}.$$

In the standard (RIP-based) argument, one proves exact reconstruction by showing that the RIP holds on Σ_s. The fact that each vector in $\sqrt{\kappa_0 s}B_1^n \cap S^{n-1}$ is well approximated by vectors from Σ_s (see, for instance, [11]) allows one to extend the RIP from Σ_s to $\sqrt{\kappa_0 s}B_1^n \cap S^{n-1}$. However, extending the RIP requires *both upper and lower* estimates, and obtaining the upper part of the RIP on Σ_s forces severe restrictions on the random vector X that do not hold under only moment assumptions. Thus, passing from Σ_s to $\sqrt{\kappa_0 s}B_1^n \cap S^{n-1}$, with only a lower bound on $\inf_{t\in\Sigma_s} \|\Gamma t\|_2$ at one's disposal, requires a totally different argument.

The method we present below is based on Maurey's empirical method and has been recently used in [26] where it is called the 'transfer principle'.

Lemma 2.6 *Let $\Gamma : \mathbb{R}^n \to \mathbb{R}^N$ be a matrix and put $1 < s \leq n$. Denote by $(e_1, \ldots, e_n)$ the canonical basis of $\mathbb{R}^n$. Assume that for every $x \in \Sigma_s$, $\|\Gamma x\|_2 \geq \lambda \|x\|_2$. Let $y \in \mathbb{R}^n$ be nonzero, set $\mu_j = |y_j|/\|y\|_1$, then,*

$$\|\Gamma y\|_2^2 \geq \lambda^2 \|y\|_2^2 - \frac{\|y\|_1^2}{s-1}\left(\sum_{j=1}^n \|\Gamma e_j\|_2^2 \mu_j - \lambda^2\right).$$

Proof. Fix $y \in \mathbb{R}^n$, let Y be a random vector in $\mathbb{R}^n$ defined by

$$P(Y = \|y\|_1 \mathrm{sgn}(y_j) e_j) = |y_j|/\|y\|_1,$$

for all $j = 1, \ldots, n$ and observe that $\mathbb{E}Y = y$.

Let $Y_1, ..., Y_s$ be independent copies of Y and set $Z = s^{-1}\sum_{k=1}^s Y_k$; therefore, $Z \in \Sigma_s$ for every realization of $Y_1, ..., Y_s$.

By the assumption, $\|\Gamma Z\|_2^2 \geq \lambda^2 \|Z\|_2^2$, and thus,

$$\mathbb{E}\|\Gamma Z\|_2^2 \geq \lambda^2 \mathbb{E}\|Z\|_2^2. \tag{2.3}$$

It is straightforward to verify that $\mathbb{E}\langle Y, Y\rangle = \|y\|_1^2$, that if $i \neq j$ then $\mathbb{E}\langle \Gamma Y_i, \Gamma Y_j\rangle = \langle \Gamma y, \Gamma y\rangle$ and that for every $1 \leq k \leq s$,

$$\mathbb{E}\langle \Gamma Y_k, \Gamma Y_k\rangle = \|y\|_1 \sum_{j=1}^n |y_j| \, \|\Gamma e_j\|_2^2.$$

Therefore, setting $\mu_j = |y_j|/\|y\|_1$, and $W = \sum_{j=1}^n \|\Gamma e_j\|_2^2 \mu_j$,

$$\mathbb{E}\|\Gamma Z\|_2^2 = \frac{1}{s^2}\sum_{i,j=1}^{s} \mathbb{E}\langle \Gamma Y_i, \Gamma Y_j\rangle = \left(1-\frac{1}{s}\right)\|\Gamma y\|_2^2 + \frac{\|y\|_1}{s}\sum_{j=1}^{n}|y_j|\,\|\Gamma e_j\|_2^2$$
$$= \left(1-\frac{1}{s}\right)\|\Gamma y\|_2^2 + W\frac{\|y\|_1^2}{s}.$$

Using the same argument one can show that

$$\mathbb{E}\|Z\|_2^2 = \left(1-\frac{1}{s}\right)\|y\|_2^2 + \frac{\|y\|_1^2}{s}.$$

Combining these two estimates with (2.3),

$$\left(1-\frac{1}{s}\right)\|\Gamma y\|_2^2 \geq \lambda^2\left(\left(1-\frac{1}{s}\right)\|y\|_2^2 + \frac{\|y\|_1^2}{s}\right) - W\frac{\|y\|_1^2}{s},$$

proving the claim. ∎

Proof of Theorem B: Let Γ satisfying the two conditions of Theorem B. It follows from Lemma 2.6 that for every $y \in B_1^n \cap rS^{n-1}$, one has

$$\|\Gamma y\|_2^2 \geq c_0^2\|y\|_2^2 - \frac{\|y\|_1}{s-1}\sum_{i=1}^{n}\|\Gamma e_i\|_2^2\,|y_i| \geq c_0^2 r^2 - \frac{c_1^2}{s-1} > 0$$

when $s-1 > c_1^2/(c_0^2 r^2)$. Then the result follows from Lemma 2.1. ∎

The matrix Γ in question will be $N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$. Thus, for every $t \in \mathbb{R}^n$, $\|\Gamma t\|_2^2 = N^{-1}\sum_{i=1}^N \langle X_i, t\rangle^2$, and if $X_j = (x_{i,j})_{i=1}^n$ then $\|\Gamma e_j\|_2^2 = N^{-1}\sum_{i=1}^N x_{i,j}^2$, which is an average of N iid random variables (though $\|\Gamma e_1\|_2$, $\ldots$, $\|\Gamma e_n\|_2$ need not be independent).

The next and final component needed for the proof of Theorem A is information on the sum of iid random variables, which will be used to upper bound $\max_{1\leq j\leq n}\|\Gamma e_j\|_2^2$.

Lemma 2.7 *There exists an absolute constant c_0 for which the following holds. Let z be a mean-zero random variable and put $z_1, \ldots, z_N$ to be N independent copies of z. Let $p_0 \geq 2$ and assume that there exists $\kappa > 0$ and $\alpha \geq 1/2$ that satisfy that $\|z\|_{L_p} \leq \kappa_1 p^\alpha$ for every $2 \leq p \leq p_0$. If $N \geq p_0^{\max\{2\alpha-1,1\}}$ then for every $p \leq p_0$,*

$$\|\sum_{i=1}^N z_i\|_{L_p} \leq c_1(\alpha)\kappa_1\sqrt{Np}$$

where $c_1(\alpha) = c_0\exp((2\alpha-1))$.

Lemma 2.7 shows that even under a weak moment assumption, $\|z\|_{L_p} \lesssim p^\alpha$ for $p \leq p_0$ and $\alpha \geq 1/2$ that can be large, a normalized sum of N independent copies of z exhibits a 'subgaussian' moment growth up to the same p_0, provided that N is sufficiently large.

The proof of Proposition 2.7 is based on the following result due to Latała [20].

Theorem 2.8 *If z is a mean-zero random variable and $z_1, ..., z_N$ are independent copies of z, then for any $p \geq 2$,*

$$\|\sum_{i=1}^N z_i\|_{L_p} \sim \sup\left\{\frac{p}{s}\left(\frac{N}{p}\right)^{1/s}\|z\|_{L_s} : \max\{2, p/N\} \leq s \leq p\right\}.$$

Proof of Proposition 2.7. Since $\|z\|_{L_p} \leq \kappa_1 p^\alpha$, it follows from Theorem 2.8 that

$$\|\sum_{i=1}^N z_i\|_{L_p} \leq c\kappa_1 \sup_s p(N/p)^{1/s} s^{-1+\alpha},$$

where the supremum is for $\max\{2, p/N\} \leq s \leq p$. It is straightforward to verify that the function $h(s) = (N/p)^{1/s} s^{-1+\alpha}$ is decreasing when $\alpha \leq 1$ and attains its maximum in $s = \max\{2, p/N\}$ or in $s = p$ when $\alpha > 1$.

Therefore, if $N \geq p$ and $\alpha \leq 1$, then

$$\|\sum_{i=1}^N z_i\|_{L_p} \leq c_1\kappa_1\sqrt{Np},$$

and if $\alpha > 1$,

$$\|\sum_{i=1}^N z_i\|_{L_p} \leq c_1\kappa_1 \max\left\{\sqrt{Np}, N^{1/p}p^\alpha\right\}.$$

Finally, if $N \geq p^{2\alpha-1}$ then $e^{2\alpha-1}\sqrt{Np} \geq N^{1/p}p^\alpha$, which completes the proof. ■

Proof of Theorem A. By Corollary 2.4, if $N \geq c_1 s\log(en/s)/\beta^2$, then with probability at least $1 - \exp(-c_2 N\beta^2)$,

$$\inf_{t\in\Sigma_s\cap S^{n-1}} \frac{1}{N}\sum_{i=1}^N \langle X_i, t\rangle^2 \geq \frac{u^2\beta}{2}.$$

Moreover, by Lemma 2.6 used for the measurements matrix Γ and $\lambda^2 = u^2\beta/2$, it follows that for $r \geq 1$ and on the same event as above,

$$\inf_{t \in \sqrt{r}B_1^n \cap S^{n-1}} \|\Gamma t\|_2^2 \geq \lambda^2 - \frac{2r}{s} \max_{1 \leq j \leq n} \|\Gamma e_j\|_2^2 . \tag{2.4}$$

Finally, fix $w \geq 1$ and consider $z = x_j^2 - 1$ - where x_j is the j-th coordinate of X. Since z is a centred random variable, then by Lemma 2.7 for $p = \kappa_2 \log(wn)$, and setting $c_3(\alpha) \sim \exp((2\alpha - 1))$,

$$\|\frac{1}{N}\sum_{i=1}^N z_i\|_{L_p} \leq c_3(\alpha)\kappa_1 \sqrt{\frac{p}{N}},$$

provided that $N \geq p^{\max\{2\alpha-1,1\}} = (\kappa_2 \log(wn))^{\max\{2\alpha-1,1\}}$. Therefore, if $N \geq (c_3(\alpha)\kappa_1)^2(\kappa_2 \log(wn))^{\max\{2\alpha-1,1\}}$, then, for $V_j = \|\Gamma e_j\|_2^2$,

$$\|V_j\|_{L_p} = \|\frac{1}{N}\sum_{i=1}^N x_{i,j}^2\|_{L_p} \leq 1 + c_3(\alpha)\kappa_1 \sqrt{\frac{\kappa_2 \log(wn)}{N}} \leq 2.$$

Observe that

$$\begin{aligned} P(\max_{1 \leq j \leq n} V_j \geq 2e) \leq & \sum_{j=1}^n P(V_j \geq 2e) \leq n \sum_{j=1}^n \left(\frac{\|V_j\|_{L_p}}{2e}\right)^p \\ \leq & n\left(\frac{1}{e}\right)^p = \frac{1}{w^{\kappa_2} n^{\kappa_2 - 1}}. \end{aligned}$$

Thus, with probability at least $1 - \exp(-c_2 N\beta^2) - 1/(w^{\kappa_2} n^{\kappa_2 - 1})$,

$$\inf_{t \in \sqrt{r}B_1^n \cap S^{n-1}} \|\Gamma t\|_2^2 \geq \lambda^2 - \frac{4er}{s} \geq \lambda^2/2 \tag{2.5}$$

provided that $r \leq s\lambda^2/8e = su^2\beta/16e$.

Combining (2.5) with Lemma 2.1 shows that if

$$N \gtrsim \max\left\{ s\log(en/s), (c_3(\alpha)\kappa_1)^2(\kappa_2 \log(wn))^{\max\{2\alpha-1,1\}} \right\},$$

then on the same event as above, Γ satisfies the exact reconstruction property for vectors that are $c_4(u,\beta)s$-sparse, as claimed. ∎

Proof of Theorem D: It follows from point 2. in Corollary 2.4 that $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \rangle e_i$ is invertible over Σ_s with probability at least $1 - \exp(-c_2 N\beta^2)$ when $N \geq c_1 s \log(en/s)$. Therefore, under the same condition and with the same probability estimate, for any $x_0 \in \Sigma_{\lfloor s/2 \rfloor}$, the only $t \in \mathbb{R}^n$ such that $\Gamma t = \Gamma x_0$ and $\|t\|_0 \leq \|x_0\|_0$ is x_0 itself. That is the ℓ_0-minimization procedure can reconstruct all vectors in $\Sigma_{\lfloor s/2 \rfloor}$. ∎

3 Proof of Theorem C and C′

Let $(e_1, ..., e_n)$ be the standard basis in $\mathbb{R}^n$. Given an $N \times n$ matrix Γ and $J \subset \{1, \ldots, n\}$ set Γ_J to be the restriction of Γ to $\mathrm{span}\{e_j : j \in J\}$. Let B_1^n be the unit ball in ℓ_1^n, and put $B_1^{J^c}$ to be the set of vectors in B_1^n that are supported in J^c – the complement of J in $\{1, ..., n\}$.

Lemma 3.1 *Fix integers $s, N \leq n$. Let $v \in \mathbb{R}^n$ be supported on $J \subset \{1, ..., n\}$ of cardinality at most s, that satisfies $\|v\|_1 = 1$. If $\Gamma v \in \Gamma B_1^{J^c}$ then Γ does not satisfy the exact reconstruction property of order s.*

Proof. By assumption, there is $w \in B_1^{J^c}$ for which $\Gamma v = \Gamma w$. Also, $v \neq w$, otherwise, $v \in B_1^J \cap B_1^{J^c}$ implying that $v = 0$, which is impossible since $\|v\|_1 = 1$.

If one performs the basis pursuit algorithm trying to recover v from Γv, w is at least as good 'candidate' as v (since $\|w\|_1 \leq 1 = \|v\|_1$), and therefore, v cannot be the unique solution to the ℓ_1-minimization problem $\min\big(\|t\|_1 : \Gamma t = \Gamma v\big)$. ∎

It immediately follows from Lemma 3.1 that if one wants to prove that the $N \times n$ matrix

$$\Gamma = \frac{1}{\sqrt{N}}\big(x_{ij}\big) = \frac{1}{\sqrt{N}} \begin{pmatrix} x_{1\cdot}^\top \\ \vdots \\ x_{N\cdot}^\top \end{pmatrix} = \frac{1}{\sqrt{N}}[x_{\cdot 1}, \cdots, x_{\cdot n}]$$

does not satisfy ER(1), it suffices to find $j \in \{1, \ldots, n\}$ for which

$$\Gamma e_j = x_{\cdot j} \in \mathrm{absconv}\big(x_{\cdot k} : k \neq j\big) = \mathrm{absconv}\big(\Gamma e_k : k \neq j\big).$$

To that end, if B_2^N denotes the Euclidean unit ball in $\mathbb{R}^N$,

$$\|x_{\cdot j}\|_2 \leq \sqrt{N} \quad \text{and} \quad \sqrt{N} B_2^N \subset \mathrm{absconv}\big(x_{\cdot k} : k \neq j\big), \tag{3.1}$$

then Γ does not satisfy ER(1).

The proof of Theorem C and of Theorem C′ thus follows from the construction of a matrix for which (3.1) holds with probability larger than $1/2$.

Let η be a selector (a $\{0,1\}$-valued random variable) with mean δ to be named later, and let ε be a symmetric $\{-1,1\}$-valued random variable that is independent of η. Fix $R > 0$ and set $z = \varepsilon(1 + R\eta)$.

Observe that if $p \geq 2$ and $R \geq 1$ then

$$\frac{\|z\|_{L_p}}{\|z\|_{L_2}} = \frac{\big(1 + \big((1+R)^p - 1\big)\delta\big)^{1/p}}{\big(1 + \big((1+R)^2 - 1\big)\delta\big)^{1/2}} \sim \frac{(1 + R^p\delta)^{1/p}}{(1 + R^2\delta)^{1/2}} \sim R\delta^{1/p},$$

and the last equivalence holds when $R^2\delta \lesssim 1$ and $R^p\delta \gtrsim 1$. Set $R = \sqrt{p}(1/\delta)^{1/p}$, and thus $\|z\|_{L_p} / \|z\|_{L_2} \sim \sqrt{p}$ (for the right choice of δ).

One can view $x = z/\|z\|_{L_2}$ as a mean-zero, variance one random variable exhibiting 'subgaussian' moments only up to the level p. Indeed, note that if $q > p$, $\|z\|_{L_q} / \|z\|_{L_2} \sim \sqrt{p}\delta^{1/q-1/p}$; hence, it may be far larger than $\sqrt{q}$ if δ is sufficiently small, as will be the case.

Let $X = (x_1, \ldots, x_n)$ be a vector whose coordinates are independent, distributed as x and let Γ be the measurements matrix generated by x. Note that up to the normalization factor of $\|z\|_{L_2}$, which is of the order of a constant when $R^2\delta \lesssim 1$, $\sqrt{N}\Gamma$ is a perturbation of a Rademacher matrix by a sparse matrix with few random spikes that are either R or $-R$.

Denote by $\mathbb{E}_\eta$ (resp. $\mathbb{E}_\varepsilon$) the expectation with respect to the η-variables (resp. ε-variables). A straightforward application of Khintchine's inequality (see, e.g., p.91 in [21]) shows that for every vector $t \in \mathbb{R}^n$,

$$\mathbb{E}\big\langle X, t\big\rangle^4 \lesssim \mathbb{E}_\eta \mathbb{E}_\varepsilon \Big(\sum_{j=1}^n \varepsilon_j (1 + R\eta_j) t_j \Big)^4 \lesssim \mathbb{E}_\eta \Big(\sum_{j=1}^n (1 + R\eta_j)^2 t_j^2 \Big)^2$$
$$= \mathbb{E}_\eta \sum_{k,\ell} (1 + R\eta_k)^2 t_k^2 (1 + R\eta_\ell)^2 t_\ell^2 \lesssim \|t\|_2^4 = \Big(\mathbb{E}\big\langle X, t\big\rangle^2 \Big)^2$$

provided that $R^4\delta \lesssim 1$.

To show that the Basis Pursuit algorithm performs poorly using random measurements generated by Γ, set $(f_i)_{i=1}^N$ to be the canonical basis of $\mathbb{R}^N$. Let $\eta_{\cdot j} = \sum_{i=1}^N \eta_{ij} f_i$ and observe that conditioned on ε_{ij}'s, for every fixed $1 \le i \le N$,

$$P_\eta \Big(\text{ there exists } j \in \{2, \ldots, n\} : z_{\cdot j} = \varepsilon_{\cdot j} + \varepsilon_{ij} R f_i \Big)$$
$$P_\eta \Big(\text{ there exists } j \in \{2, \ldots, n\} : \eta_{\cdot j} = f_i \Big)$$
$$= 1 - (1 - (1-\delta)^{N-1}\delta)^{n-1} \ge 1 - \frac{1}{4N}$$

provided that

$$\frac{\log N}{n} \lesssim \delta \lesssim \frac{\log\big(en/N\big)}{N}.$$

Hence, by a Fubini argument, with probability at least 3/4 there are (random) $y_1, ..., y_N \in B_\infty^N$ for which

$$\operatorname{absconv}\big(Rf_i + y_i : 1 \le i \le N\big) \subset \operatorname{absconv}\big(z_{\cdot k} : k \ne 1\big).$$

Lemma 3.2 *Using the notation above, if* $v_i = Rf_i + y_i$ *for* $1 \le i \le N$ *and* $y_i \in B_\infty^N$, *then* $\big(R/\sqrt{N} - \sqrt{N}\big)B_2^N \subset \mathrm{absconv}(v_1, ..., v_N) \equiv V$

Proof. A straightforward separation argument may be used to show that if, for every $w \in S^{N-1}$, $\sup_{v\in V} |\langle v, w\rangle| \ge \rho$, then $\rho B_2^N \subset V$ (indeed, otherwise there would be some $x \in \rho B_2^N \backslash V$; but it is impossible to separate x and the convex and symmetric V using any norm one functional).

Now, to complete the proof, observe that for every $w \in S^{N-1}$,

$$\begin{aligned}&\sup_{v\in V} |\langle v, w\rangle| = \max_{1\le i\le N} |\langle Rf_i + y_i, w\rangle| \\ &\ge \max_{1\le i\le N} |\langle w, Rf_i\rangle| - \max_{1\le i\le N} |\langle y_i, w\rangle| \ge R/\sqrt{N} - \sqrt{N}.\end{aligned}$$

■

Applying Lemma 3.2, if $R \ge 2N$ then with probability at least 3/4, $\sqrt{N}B_2^N \subset \mathrm{absconv}\big(z_{\cdot k} : k \ne 1\big)$. On the other hand, if $\delta \lesssim 1/N$ then

$$P\big[\, \|z_{\cdot 1}\|_2 = \sqrt{N}\big] = (1-\delta)^N \ge 3/4.$$

Hence, combining the two observations, with probability at least 1/2,

$$\|z_{\cdot 1}\|_2 \le \sqrt{N} \ \text{ and } \ \sqrt{N}B_2^N \subset \mathrm{absconv}\big(z_{\cdot k} : k \ne 1\big),$$

and thus

$$x_{\cdot 1} \in \mathrm{absconv}\big(x_{\cdot k} : k \ne 1\big). \tag{3.2}$$

Of course, this assertion holds under several conditions on the parameters involved: namely, that $R = \sqrt{p}(1/\delta)^{1/p} \ge 2N$; that $(\log N)/n \lesssim \delta \lesssim \log\big(en/N\big)/N$; that $R^4\delta \lesssim 1$ and that $\delta \lesssim 1/N$.

For instance, one may select $\delta \sim (\log N)/n$ and $p \sim (\log n)/\log N$, in which case all the conditions above are met and with probability at least 1/2, Γ does not satisfy ER(1), proving Theorem C. A similar calculation leads to the proof of Theorem C′. ■

4 Results in the noisy measurements setup

In previous sections, we considered the idealized scenario in which the data was noiseless. In this section, we study properties of the measurement vectors in the noisy setup.

In the noisy framework, one observes N couples $(z_i, X_i)_{i=1}^N$ where the z_i's are noisy observations of the inner products $\langle X_i, x_0 \rangle, i = 1, \dots, N$ that have been corrupted by some noise:

$$z_i = \langle X_i, x_0 \rangle + g_i, \quad i = 1, \dots, N \tag{4.1}$$

where $(g_i)_{i=1}^N$ represents the 'noise-vector' and $x_0 \in \mathbb{R}^n$ is an unknown vector. For simplicity, we will assume that the g_i's are independent gaussian random variables $\mathcal{N}(0, \sigma^2)$ that are also independent of the X_i's. The goal is to obtain as much information as possible on the unknown vector x_0 with only the data $(z_i, X_i)_{i=1}^N$ at one's disposal.

Unlike the noiseless case, there is no hope to reconstruct x_0 from the N noisy observations $z_1, \dots, z_N$. Instead of exact reconstruction, there are three natural questions that arise in the noisy setup. Firstly, the *estimation problem*: given some norm $\|\cdot\|$ in $\mathbb{R}^n$, one would like to construct a procedure $\hat{x}$ for which $\|\hat{x} - x_0\|$ is as small as possible; secondly, the *prediction problem*: given a new (random, independent) "input" $X \in \mathbb{R}^n$, one has to find a good guess $\langle X, \hat{x} \rangle$ of the most likely associated output z, knowing that (z, X) shares the same distribution with the other couples $(z_1, X_1), \dots, (z_N, X_N)$; and finally, the *de-noising problem*: given a norm $\|\cdot\|$ in $\mathbb{R}^N$, one has to construct a procedure $\hat{x}$ for which $\|\Gamma\hat{x} - \Gamma x_0\|$ is small, where $\Gamma = N^{-1/2} \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$ is the measurement matrix.

These three problems are central in modern statistics, and are featured in numerous statistical monographs, particularly in the context of the gaussian regression model (Equation (4.1)).

Recently, all three problems have been recast in the high-dimensional setup: the number of observations N can be much smaller than the ambient dimension n but the target vector x_0 is usually believed to have a short support: $\|x_0\|_0 := s$, and s is assumed to be much smaller than n.

Although one may still consider classical procedures, like Ordinary Least Square (OLS) estimators to deal with the problems in the 'high-dimensional' setup, it turns out that performance is poor, because n is much larger than N. This phenomenon is usually referred to as the *curse of dimensionality* and is a major theme in current-day problems.

What saves the day is the assumption that even though x_0 lives in the high dimensional space $\mathbb{R}^n$ and one has access to a small number of observations, x_0 is believed to have some *low-dimensional structure*: it is s-sparse. The aim is therefore to use this information to design procedures having the ability to perform as if the true dimension were s rather than n.

The heart of this *dimension reduction problem* is the fact that one does not know the support of x_0, other than it is short, and to address it one may

resort to penalization methods, and, in particular, the ℓ_0 and ℓ_1 penalization methods.

Among the first dimension reduction schemes were ℓ_0-minimization procedures, which were sometimes called Model Selection procedures. A detailed analysis in the context of the model (4.1), as well as for other examples, may be found in Chapter 4 of [22] or in [3].

The obvious downside of ℓ_0 minimization is that its practical implementation requires, in some situations, a combinatorial search among far too many subspaces of the ambient space $\mathbb{R}^n$. To overcome this barrier, convex relaxations have been introduced, among which is ℓ_1 minimization, and the idea of a convex relaxation is why the ℓ_1-norm plays such a central role in the noisy setup.

In this context, a classical ℓ_1-minimization procedure is the *Lasso* (cf. [30]):

$$\hat{x}_\lambda \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \Big(\frac{1}{N}\sum_{i=1}^{N} \big(z_i - \langle X_i, x\rangle\big)^2 + \lambda \|x\|_1 \Big), \tag{4.2}$$

and $\lambda > 0$ is called the regularization parameter. Another ℓ_1-based procedure is the Dantzig selector (see [10]); estimation and de-noising properties of both these estimators have been obtained under several conditions (cf. [5, 2, 31] or Chapter 7 and 8 in [19]).

In this section, we shall focus on two conditions. The first, called the Compatibility Condition, was introduced in [31] (see also Definition 2.1 in [32]); the second, the Restricted Eigenvalue Condition, was introduced in [2].

Definition 4.1 *Let Γ be an $N \times n$ matrix. For $L > 0$ and a set $S \subset \{1, \dots, n\}$, put*

$$\delta(L, S) = \min \Big(\|\Gamma\zeta_S - \Gamma\zeta_{S^c}\|_2 : \|\zeta_S\|_1 = 1, \|\zeta_{S^c}\|_1 \le L \Big) \tag{4.3}$$

where ζ_S (resp. ζ_{S^c}) denotes a vector that is supported on S (resp. S^c).

The function $\phi(L, S) = \sqrt{|S|}\delta(L, S)$ is called ***compatibility constant*** *associated with L and S.*

Given an absolute constant $c_0 > 0$, Γ satisfies the ***Compatibility Condition for the set*** *S_0 if there exists $L > 1$ for which $\phi(L, S_0) \ge c_0$ and we say that Γ satisfies the* ***uniform Compatibility Condition (CC) of order*** *s if $\min_{|S|\le s} \phi(L, S) \ge c_0$.*

A typical result for the Lasso, in the Gaussian model (4.1) and when Γ satisfies the Compatibility Condition is Theorem 6.1 in [5]: if S_0 is the sup-

port of x_0 and $\lambda \sim \sigma\sqrt{(\log n)/N}$, then with high probability (with respect to the noise),

$$\|\Gamma\hat{x}_\lambda - \Gamma x_0\|_2^2 \lesssim \sigma^2 \frac{\|x_0\|_0 \log n}{N\phi(3, S_0)} \text{ and } \|\hat{x}_\lambda - x_0\|_1 \lesssim \frac{\sigma \|x_0\|_0}{\phi(3, S_0)} \sqrt{\frac{\log n}{N}} \quad (4.4)$$

Note that, up to a logarithmic factor, the rates obtained in (4.4) are the same as the Ordinary Least Square estimator would have given, had the support of x_0 been known. Thus, under the Compatibility Condition for S_0, the Lasso yields dimension reduction, and the extra $\log n$ factor seems a rather small price to pay for not knowing the support of x_0 in advance (see Corollary 3 in [34] for the necessity of the $\log n$ factor). Moreover, compared with ℓ_0-minimization procedures, the Lasso can be efficiently implemented using, for instance the LARS algorithm from [16, 17].

Naturally, although the Compatibility Condition in S_0, the support of x_0, suffices to prove that the Lasso is an effective procedure, the fact that S_0 is unknown makes this condition hard to verify. Thus, showing that Γ satisfies the uniform Compatibility Condition is a safer requirement – and the one we shall explore below.

Another uniform condition of a similar flavour is the Restricted Eigenvalue Condition from [2]. To define it, let us introduce the following notation: for $x \in \mathbb{R}^n$ and a set $S_0 \subset \{1, \dots, n\}$ of cardinality $|S_0| \leq s$, let S_1 be the subset of indexes of the m largest coordinates of $(|x_i|)_{i=1}^n$ that are outside S_0. Let $x_{S_{01}}$ be the restriction of x to the set $S_{01} = S_0 \cup S_1$.

Definition 4.2 *Let Γ be an $N \times n$ matrix. For a constant $c_0 \geq 1$ and an integer $1 \leq s \leq m \leq n$ for which $m + s \leq n$, let the* ***restricted eigenvalue constant*** *be*

$$\kappa(s, m, c_0) = \min\Big(\frac{\|\Gamma x\|_2}{\|x_{S_{01}}\|_2} : S_0 \subset \{1, \dots, n\}, |S_0| \leq s, \big\|x_{S_0^c}\big\|_1 \leq c_0 \|x_{S_0}\|_1\Big).$$

The matrix Γ satisfies the ***Restricted Eigenvalue Condition (REC) of order s with a constant*** *c if $\kappa(s, s, 3) \geq c$.*

Estimation and de-noising results follow from Theorem 6.1 (for the Dantzig selector) and Theorem 6.2 (for the Lasso) in [2] when the measurement matrix Γ, normalized so that the diagonal elements of $\Gamma^\top\Gamma$ equal 1, satisfies the REC of the appropriate order and with a constant that is independent of the dimension. We also refer to Lemma 6.10 in [5] for similar results that do not require normalization.

Observe that the REC of order s is a stronger condition than the uniform CC of order s since $\kappa(s,m,c_0) \leq \min_{|S|\leq s} \phi(c_0,S)$ for every $c_0 > 0$ and $1 \leq s,m \leq n$. Having said that, it also yields more information [2], since under the REC the Lasso has better estimation properties: for $\lambda \sim \sigma\sqrt{(\log n)/N}$ and with high probability (with respect to the noise), simultaneously for all $1 < p \leq 2$,

$$\|\hat{x}_\lambda - x_0\|_p^p \lesssim \|x_0\|_0 \left(\frac{\sigma}{\kappa(s,s,3)}\sqrt{\frac{\log n}{N}}\right)^p. \tag{4.5}$$

One question that comes to mind is whether there are matrices that satisfy the uniform CC or the REC. Just as in Compressed Sensing, the only matrices that are known to satisfy the Uniform Compatibility Condition or the Restricted Eigenvalue Condition for the optimal number of measurements (rows) are some random matrices.

For example, results in [27] show that 'typical' measurement matrices $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$ with independent gaussian random measurement vectors $X_1,\ldots,X_N$, selected according to the centred gaussian measure with a covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, satisfy the REC of order s with the optimal number of observations $N \gtrsim s\log(en/s)$ when Σ satisfies the REC of order s. This result was extended to general measurement matrices with subgaussian rows in [28]. A recent result [26] shows that if $\tilde{D} = \mathrm{Diag}(\|\Gamma e_1\|_2,\ldots,\|\Gamma e_n\|_2)$ where $(e_1,\ldots,e_n)$ is the canonical basis of $\mathbb{R}^n$, the normalized measurement matrices $\tilde{\Gamma} = \Gamma\tilde{D}^{-1}$ satisfy a non-uniform REC (that is a REC that holds only one set S_0) under a weak L_4-moment assumption: $\|\langle X,t\rangle\|_{L_4} \leq \kappa_0 \|\langle X,t\rangle\|_{L_2}$ for all $t \in \Sigma_N$.

Our aim in this final section is to extend the results presented in previous sections to the noisy setup, by identifying almost necessary and sufficient moment assumptions for the CC and REC. This turns out to be a straightforward: on one hand, the proof of Theorem A actually provides a stronger quantitative version of the Exact Reconstruction property; on the other, the uniform Compatibility Condition can be viewed as a *quantitative version* of a condition on the polytope ΓB_1^n that characterizes the Exact Reconstruction property of Γ. A similar observation is true for the Restricted Eigenvalue Condition: it can be viewed as a quantitative version of the Null Space Property (see [14, 15]) which is also equivalent to the Exact Reconstruction property.

Let us recall the definitions of the two properties in question.

Definition 4.3 *Let $1 \leq s \leq N$. A centrally symmetric polytope $P \subset \mathbb{R}^N$ is **s-neighbourly** if every set of s of its vertices, containing no antipodal pair, is the set of all vertices of some face of P.*

A well known result in Compressed Sensing obtained in [12] shows that Γ satisfies ER(s) (see Definition 1.1) if and only if ΓB_1^n has $2n$ vertices and ΓB_1^n is a centrally symmetric s-neighbourly polytope. It turns out that this property is characterized by the uniform CC.

Proposition A. *Let Γ be an $N \times n$ matrix. The following are equivalent:*

1. *ΓB_1^n has $2n$ vertices and ΓB_1^n is s-neighborly,*
2. *$\min\big(\phi(1,S) : S \subset \{1,\dots,n\}, |S| \leq s\big) > 0$.*

In particular, $\min_{|S|\leq s} \phi(L,S)$ for some $L \geq 1$ is a quantitative measure of the s-neighbourly property of ΓB_1^n. Indeed, if ΓB_1^n is s-neighbourly and has $2n$ vertices then the two sets

$$\big\{\Gamma\zeta_S : \|\zeta_S\|_1 = 1\big\} \text{ and } \big\{\Gamma\zeta_{S^c} : \|\zeta_{S^c}\|_1 \leq 1\big\} \tag{4.6}$$

are disjoint for every $|S| \leq s$. However, $\min_{|S|\leq s} \phi(1,S)$ measures how far the two sets are from one another, uniformly over all S or cardinality $|S| \leq s$.

Proof. Let $C_1,\dots,C_n$ be the n columns of Γ. It follows from Proposition 2.2.13 and Proposition 2.2.16 in [11] that ΓB_1^n has $2n$ vertices and is a centrally symmetric s-neighbourly polytope if and only if for every $S \subset \{1,\dots,n\}$ of cardinality $|S| \leq s$ and every choice of signs $(\varepsilon_i) \in \{-1,1\}^S$,

$$\mathrm{conv}\big(\{\varepsilon_i C_i : i \in S\}\big) \cap \mathrm{conv}\big(\{\theta_j C_j : j \notin S, \theta_j = \pm 1\}\big) = \emptyset \tag{4.7}$$

It is straightforward to verify that

$$\bigcup_{(\varepsilon_i)\in\{\pm1\}^S} \mathrm{conv}\big(\{\varepsilon_i C_i : i \in S\}\big) = \Big\{\Gamma\zeta_S : \|\zeta_S\|_1 = 1\Big\}$$

and

$$\mathrm{conv}\big(\{\theta_i C_i : i \notin S, \theta_i = \pm 1\}\big) = \Big\{\Gamma\zeta_{S^c} : \|\zeta_{S^c}\|_1 \leq 1\Big\}.$$

As a consequence, (4.7) holds for every $S \subset \{1,\dots,n\}$ of cardinality at most s if and only if $\min\big(\phi(1,S) : S \subset \{1,\dots,n\}, |S| \leq s\big) > 0$. ∎

An observation of a similar nature is true for the Restricted Eigenvalue Condition, which can be viewed as a quantitative measure of the Null Space Property.

Definition 4.4 *Let Γ be an $N \times n$ matrix. Γ satisfies the **Null Space property of order** s if it is invertible in the cone*

$$\big\{x \in \mathbb{R}^n : \text{ there exists } S \subset \{1,\dots,n\}, |S| \leq s \text{ and } \|x_{S^c}\|_1 \leq \|x_S\|_1\big\}. \tag{4.8}$$

In [14, 15], the authors prove that Γ satisfies ER(s) if and only if it has the Null Space Property of order s.

A natural way to quantify the invertibility of Γ in the cone is to consider its smallest singular value, restricted to this cone, which is simply $\kappa(s, n-s, 1)$. Unfortunately, statistical properties of the Lasso and the Dantzig selector are not known under the assumption that $\kappa(s, n-s, 1)$ is of the order of an absolute constant. On the other hand, (optimal) statistical properties for the Lasso from [2] were obtained under the assumption that $\kappa(s, s, 3)$ is of the order of a constant.

Since the definition of $\kappa(s, s, 3)$ involves a larger cone and vectors of a different support size than the one considered in (4.8), $\kappa(s, s, 3)$ is not directly comparable with $\kappa(s, n-s, 1)$. Despite that, it may still serve as a quantitative version of the Null Space property. In fact, all the parameters $\kappa(s, m, c)$ can be used to quantify the Null Space property, although only $k(s, m, c) \geq c_0$ for $m \geq s$ and $c \geq 3$ are known to be sufficient conditions to guarantee (optimal) statistical properties for the LASSO (and for the Dantzig selector).

The main result of this section identifies the measurement vectors for which the associated measurement matrix satisfies the two uniform conditions in question.

Theorem E. *Let $L > 0$, $1 \leq s \leq n$ and $c_0 > 0$. Under the same assumptions as in Theorem A and with the same probability estimate, $\Gamma = N^{-1/2} \sum_{i=1}^{N} \langle X_i, \cdot \rangle e_i$ satisfies:*

1. *A uniform Compatibility Condition of order $c_1 s$, namely that*

$$\min_{|S| \leq c_1 s} \phi(L, S) \geq u^2 \beta / 4$$

   for $c_1 = u^2 \beta / (16e(1+L)^2)$.

2. *A Restricted Eigenvalue Condition of order $c_2 s$, with*

$$\kappa(c_2 s, m, c_0) \geq u^2 \beta / 4$$

   for any $1 \leq m \leq n$, as long as $(1+c_0)^2 c_2 \leq u^2 \beta / (16e)$.

On the other hand, if Γ is the matrix considered in Theorem C, then with probability at least $1/2$, $\phi(1, \{e_1\}) = 0$ and $\kappa(1, m, 1) = 0$ for any $1 \leq m \leq n$.

Just like Theorem A and Theorem C, Theorem E shows that the requirement that the coordinates of the measurement vector have $\log n$ moments

is almost a necessary and sufficient condition for the uniform Compatibility Condition and the Restricted Eigenvalue Condition to hold. Moreover, it shows that the role of the small ball property (Definition 1.3) of the measurement vector plays a major role in the context of these two conditions in the noisy setup.

It also follows from Theorem E and Proposition A (or from Theorem A and the main result in [12]) that under the same conditions as in Theorem A, ΓB_1^n has $2n$ vertices and is s-neighbourly with high probability for $N \sim s\log(en/s)$. In particular, this improves Theorem 4.3 in [1] for matrices generated by sub-exponential variables by a logarithmic factor.

Note that the counter-example constructed to prove Theorem C (or for the second part of Theorem E) and for which Γ does not satisfy ER(1), does not necessarily generate ΓB_1^n that is not s-neighbourly. Indeed, an inspection of the construction shows that the reason ER(1) fails is that with probability at least $1/2$, ΓB_1^n has less than $2n-2$ vertices, rather than that ΓB_1^n is not s-neighbourly. Thus, the question of whether a moment condition is necessary for the random polytope ΓB_1^n to be s-neighbourly with probability at least $1/2$ is still unresolved.

Proof of Theorem E: Fix a constant c_1 to be named later and let $S \subset \{1,\ldots,n\}$ of cardinality $|S| \leq c_1 s$. Let $\zeta_S \in \mathbb{R}^n$ be a vector supported on S with $\|\zeta_S\|_1 = 1$ and let $\zeta_{S^c} \in \mathbb{R}^n$ be supported on S^c with $\|\zeta_{S^c}\|_1 \leq L$. Consider $\gamma = (\zeta_S - \zeta_{S^c})/\|\zeta_S - \zeta_{S^c}\|_2$. Since $\|\zeta_S - \zeta_{S^c}\|_2 \geq \|\zeta_S\|_2 \geq \|\zeta_S\|_1/\sqrt{|S|} = 1/\sqrt{|S|}$, it follows that $\gamma \in \big((1+L)\sqrt{|S|}\big)B_1^n \cap S^{n-1}$.

Recall that by (2.5), if $r = (1+L)^2 c_1 s$ is smaller than $su^2\beta/(16e)$, then $\|\Gamma\gamma\|_2 \geq (u^2\beta)/4$. Therefore,

$$\|\Gamma\zeta_S - \Gamma\zeta_{S^c}\|_2 \geq \frac{u^2\beta}{4}\|\zeta_S - \zeta_{S^c}\|_2 \geq \frac{u^2\beta}{4}\|\zeta_S\|_2 \geq \frac{u^2\beta\,\|\zeta_S\|_1}{4\sqrt{|S|}} = \frac{u^2\beta}{4\sqrt{|S|}},$$

and thus $\min_{|S|\leq c_1 s}\phi(L,S) \geq u^2\beta/4$ for $c_1 = u^2\beta/\big(16e(1+L)^2\big)$.

Turning to the REC, fix a constant c_2 to be named later, let $x \in \mathbb{R}^n$, put $S_0 \subset \{1,\ldots,n\}$ of cardinality $|S_0| \leq c_2 s$ and for which $\big\|x_{S_0^c}\big\|_1 \leq c_0\|x_{S_0}\|_1$. Let $S_1 \subset \{1,\ldots,n\}$ be the set of indexes of the m largest coordinates of $(|x_i|)_{i=1}^n$ outside S_0 and put $S_{01} = S_0 \cup S_1$.

Observe that $\|x\|_1 \leq (1+c_0)\|x_{S_0}\|_1 \leq (1+c_0)\sqrt{|S_0|}\,\|x\|_2$; hence $x/\|x\|_2 \in \big((1+c_0)\sqrt{|S_0|}\big)B_1^n \cap S^{n-1}$. Applying (2.5) again, if $(1+c_0)^2 c_2 s \leq su^2\beta/(16e)$, then $\|\Gamma x\|_2 \geq \big((u^2\beta)/4\big)\|x\|_2$. Thus,

$$\frac{\|\Gamma x\|_2}{\|x_{S_{01}}\|_2} \geq \frac{\|\Gamma x\|_2}{\|x\|_2} \geq \frac{u^2\beta}{4}$$

and $\kappa(c_2 s, m, c_0) \geq u^2\beta/4$ for any $1 \leq m \leq n$, as long as $(1+c_0)^2 c_2 \leq u^2\beta/(16e)$.

The proof of the second part of Theorem E is an immediate corollary of the construction used in Theorem C. According to (3.2), $x_{\cdot 1} \in \mathrm{absconv}(x_{\cdot k} : k \neq 1)$ with probability at least $1/2$. Therefore, on that event, $\|\Gamma e_1 - \Gamma\zeta\|_2 = 0$ for $\zeta = \sum_{j=2}^n \lambda_j e_j$ for some well chosen weights $\lambda_2, \cdots, \lambda_n$ satisfying $\sum_{i=2}^n |\lambda_i| \leq 1$. Hence $\phi(1, \{e_1\}) = 0$ and $\kappa(1, m, 1) = 0$ for any $1 \leq m \leq n$, as claimed. ∎

Remark 4.5 *Results obtained in Theorem A and E are also valid under the same assumptions and with the same probability estimate for the normalized measurement matrix:*

$$\tilde{\Gamma} := \Gamma \tilde{D}^{-1} \text{ where } \tilde{D} = \mathrm{diag}\big(\|\Gamma e_1\|_2, \ldots, \|\Gamma e_n\|_2 \big)$$

where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. This follows from the same argument as for the matrix Γ even though $\tilde{\Gamma}$ does not have independent rows vectors because of the normalization. We do not provide more details in this case for the sake of shortness.

References

[1] Radosław Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constr. Approx.*, 34(1):61–88, 2011.

[2] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[3] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[5] Peter Bühlmann and Sara A. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[6] Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

[7] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[8] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[9] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[10] Emmanuel J. Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[11] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012.

[12] David L. Donoho. Neighborly polytopes and sparse solutions of under-determined linear equations. Technical report, Department of Statistics, Standford University, 2005.

[13] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[14] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202 (electronic), 2003.

[15] David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[16] Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642, 2004. With comments and a rejoinder by the author.

[17] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[18] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[19] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[20] Rafał Latała. Estimation of moments of sums of independent real random variables. *Ann. Probab.*, 25(3):1502–1513, 1997.

[21] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[22] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[23] Shahar Mendelson. Learning without concentration. Technical report, Technion, 2013. arXiv:1401.0304.

[24] Shahar Mendelson and Vladimir Koltchinskii. Bounding the smallest singular value of a random matrix without concentration. Technical report, Technion and Georgia Tech, 2013. arXiv:1312.3580.

[25] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2008.

[26] Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. Technical report, IMPA, 2013. arXiv:1312.2903.

[27] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

[28] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59(6):3434–3447, 2013.

[29] Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *Ann. Probab.*, 41(5):3081–3111, 2013.

[30] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[31] Sara A. van de Geer. The deterministic lasso. *In JSM proceedings. American Statistical Association*, 140, 2007.

[32] Sara A. van de Geer. Weakly decomposable regularization penalties and structured sparsity. Technical report, ETH Zürich, 2013.

[33] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[34] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.