A REMARK ON THE LASSO AND THE DANTZIG SELECTOR

YOHANN DE CASTRO

Abstract. This article investigates a new parameter for the high-dimensional regression with noise: *the distortion*. This latter has attracted a lot of attention recently with the appearance of new deterministic constructions of "almost"-Euclidean sections of the L1-ball. It measures how far is the intersection between the kernel of the design matrix and the unit L1-ball from an L2-ball. We show that the distortion holds enough information to derive *oracle inequalities* (i.e. a comparison to an ideal situation where one knows the s largest coefficients of the target) for the lasso and the Dantzig selector.

1. Introduction

In the past decade much emphasis has been put on recovering a large number of unknown variables from few noisy observations. Consider the high-dimensional linear model where one observes a vector $y \in \mathbb{R}^n$ such that

$$y = X\beta^\star + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times p}$ is called the design matrix (known from the experimenter), $\beta^\star \in \mathbb{R}^p$ is an unknown target vector one would like to recover, and $\varepsilon \in \mathbb{R}^n$ is a stochastic error term that contains all the perturbations of the experiment.

A standard hypothesis in high-dimensional regression [HTF09] requires that one can provide a constant $\lambda_n^0 \in \mathbb{R}$, as small as possible, such that

$$\big\| X^\top \varepsilon \big\|_{\ell_\infty} \leq \lambda_n^0, \tag{1}$$

with an overwhelming probability, where $X^\top \in \mathbb{R}^{p\times n}$ denotes the transpose matrix of X. In the case of n-multivariate Gaussian distribution, it is known that $\lambda_n^0 = \mathcal{O}(\sigma_n\sqrt{\log p})$, where $\sigma_n > 0$ denotes the standard deviation of the noise; see Lemma A.1.

Suppose that you have far less observation variables y_i than the unknown variables $\beta_i^\star$. For instance, let us mention the *Compressed Sensing* problem [Don06, CRT06] where one would like to simultaneously acquire and compress a signal using few (non-adaptive) linear measurements, i.e. $n \ll p$. In general terms, we are interested in accurately estimating the target vector $\beta^\star$ and the response $X\beta^\star$ from few and corrupted observations. During the past decade, this challenging issue has attracted a lot of attention among the statistical society. In 1996, R. Tibshirani introduced the lasso [Tib96]:

$$\beta^\ell \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2}\big\| X\beta - y \big\|_{\ell_2}^2 + \lambda_\ell \big\|\beta\big\|_{\ell_1} \Big\}, \tag{2}$$

where $\lambda_\ell > 0$ denotes a tuning parameter. Two decades later, this estimator continues to play a key role in our understanding of high-dimensional inverse problems. Its popularity might be due to the fact that this estimator is computationally tractable. Indeed, the lasso can be recasted

Date: September 28, 2012.
Key words and phrases. Lasso; Dantzig selector; Oracle inequality; Almost-Euclidean section; Distortion.

in a Second Order Cone Program (SOCP) that can be solved using an interior point method. Recently, E.J. Candès and T. Tao [CT07] have introduced the Dantzig selector as

$$\beta^d \in \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_{\ell_1} \quad \text{s.t.} \quad \|X^\top(y - X\beta)\|_{\ell_\infty} \leq \lambda_d\,, \tag{3}$$

where $\lambda_d > 0$ is a tuning parameter. It is known that it can be recasted as a linear program. Hence, it is also computationally tractable. A great statistical challenge is then to find efficiently verifiable conditions on X ensuring that the lasso (2) or the Dantzig selector (3) would recover "most of the information" about the target vector $\beta^\star$.

1.1. **Our goal.** What do we precisely mean by "most of the information" about the target? What is the amount of information one could recover from few observations? These are two of the important questions raised by Compressed Sensing. Suppose that you want to find an s-sparse vector (i.e. a vector with at most s non-zero coefficients) that represents the target, then you would probably want that it contains the s largest (in magnitude) coefficients $\beta_i^\star$. More precisely, denote by $\mathcal{S}_\star \subseteq \{1,\ldots,p\}$ the set of the indices of the s largest coefficients. The s-best term approximation vector is $\beta^\star_{\mathcal{S}_\star} \in \mathbb{R}^p$ where $(\beta^\star_{\mathcal{S}_\star})_i = \beta^\star_i$ if $i \in \mathcal{S}_\star$ and 0 otherwise. Observe that it is the s-sparse projection in respect to any ℓ_q-norm for $1 \leq q < +\infty$ (i.e. it minimizes the ℓ_q-distance to $\beta^\star$ among all the s-sparse vectors), and then the most natural approximation by an s-sparse vector.

Suppose that someone gives you all the keys to recover $\beta^\star_{\mathcal{S}_\star}$. More precisely, imagine that you know the subset $\mathcal{S}_\star$ in advance and that you observe $y^{oracle} = X\beta^\star_{\mathcal{S}_\star} + \varepsilon$. This is an ideal situation referred as the oracle case. Assume that the noise ε is a Gaussian white noise of standard deviation σ_n, i.e. $\varepsilon \sim \mathcal{N}_n(0, \sigma_n^2\,\mathrm{Id}_n)$ where $\mathcal{N}_n$ denotes the n-multivariate Gaussian distribution. Then the optimal estimator is the ordinary least square $\beta^{ideal} \in \mathbb{R}^p$ on the subset $\mathcal{S}_\star$, namely:

$$\beta^{ideal} \in \arg\min_{\substack{\beta\in\mathbb{R}^p \\ \mathrm{Supp}(\beta)\subseteq\mathcal{S}_\star}} \big\|X\beta - y^{oracle}\big\|_{\ell_2}^2,$$

where $\mathrm{Supp}(\beta) \subseteq \{1,\ldots,p\}$ denotes the support (i.e. the set of the indices of the non-zero coefficients) of the vector β. It holds

$$\big\|\beta^{ideal} - \beta^\star\big\|_{\ell_1} = \big\|\beta^{ideal} - \beta^\star_{\mathcal{S}_\star}\big\|_{\ell_1} + \big\|\beta^\star_{\mathcal{S}_\star^c}\big\|_{\ell_1} \leq \sqrt{s}\big\|\beta^{ideal} - \beta^\star_{\mathcal{S}_\star}\big\|_{\ell_2} + \big\|\beta^\star_{\mathcal{S}_\star^c}\big\|_{\ell_1}\,,$$

where $\beta^\star_{\mathcal{S}_\star^c} = \beta^\star - \beta^\star_{\mathcal{S}_\star}$ denotes the error vector of the s-best term approximation. A calculation of the solution of the least square estimator shows that:

$$\begin{aligned}
\mathbb{E}\big\|\beta^{ideal} - \beta^\star_{\mathcal{S}_\star}\big\|_{\ell_2}^2 &= \mathbb{E}\big\|\big(X_{\mathcal{S}_\star}^\top X_{\mathcal{S}_\star}\big)^{-1} X_{\mathcal{S}_\star}^\top y^{oracle} - \beta^\star_{\mathcal{S}_\star}\big\|_{\ell_2}^2\,,\\
&= \mathbb{E}\big\|\big(X_{\mathcal{S}_\star}^\top X_{\mathcal{S}_\star}\big)^{-1} X_{\mathcal{S}_\star}^\top \varepsilon\big\|_{\ell_2}^2 = \mathrm{Trace}\big(\big(X_{\mathcal{S}_\star}^\top X_{\mathcal{S}_\star}\big)^{-1}\big)\cdot\sigma_n^2\,,\\
&\geq \Big(\frac{1}{\rho_1}\Big)^2\cdot\sigma_n^2\cdot s\,,
\end{aligned}$$

where $X_{\mathcal{S}_\star} \in \mathbb{R}^{n\times s}$ denotes the matrix composed by the columns $X_i \in \mathbb{R}^n$ of the matrix X such that $i \in \mathcal{S}_\star$, and ρ_1 is the largest singular value of X. It yields that

$$\Big[\mathbb{E}\big\|\beta^{ideal} - \beta^\star_{\mathcal{S}_\star}\big\|_{\ell_2}^2\Big]^{1/2} \geq \frac{1}{\rho_1}\cdot\sigma_n\cdot\sqrt{s}.$$

In a nutshell, the ℓ_1-distance between the target $\beta^\star$ and the optimal estimator β^{ideal} can be reasonably said of the order of

$$\frac{1}{\rho_1}\cdot\sigma_n\cdot s + \big\|\beta^\star_{\mathcal{S}_\star^c}\big\|_{\ell_1}. \tag{4}$$

In this article, we say that the lasso satisfies a *variable selection oracle inequality of order s* if and only if its ℓ_1-distance to the target, namely $\|\beta^\ell - \beta^\star\|_{\ell_1}$, is bounded by (4) up to a "satisfactory" multiplicative factor.

In some situations it could be interesting to have a good approximation of $X\beta^\star$. In the oracle case, we have

$$\begin{aligned}\|X\beta^{ideal} - X\beta^\star\|_{\ell_2} &\leq \|X\beta^{ideal} - X\beta^\star_{\mathcal{S}_\star}\|_{\ell_2} + \|X\beta^\star_{\mathcal{S}_\star^c}\|_{\ell_2}\,,\\ &\leq \|X\beta^{ideal} - X\beta^\star_{\mathcal{S}_\star}\|_{\ell_2} + \rho_1\|\beta^\star_{\mathcal{S}_\star^c}\|_{\ell_1}\,.\end{aligned}$$

where ρ_1 denotes the largest singular value of X. An easy calculation gives that

$$\mathbb{E}\|X\beta^{ideal} - X\beta^\star_{\mathcal{S}_\star}\|_{\ell_2}^2 = \mathrm{Trace}\big(X_{\mathcal{S}_\star}\big(X_{\mathcal{S}_\star}^\top X_{\mathcal{S}_\star}\big)^{-1}X_{\mathcal{S}_\star}^\top\big)\cdot\sigma_n^2 = \sigma_n^2\cdot s.$$

Hence a tolerable upper bound is given by

$$\sigma_n\cdot\sqrt{s} + \rho_1\|\beta^\star_{\mathcal{S}_\star^c}\|_{\ell_1}. \tag{5}$$

We say that the lasso satisfies an *error prediction oracle inequality of order s* if and only if its prediction error is upper bounded by (5) up to a "satisfactory" multiplicative factor (say logarithmic in p).

1.2. **Framework.** In this article, we investigate designs with known distortion. We begin with the definition of this latter:

Definition 1 — *A subspace $\Gamma \subset \mathbb{R}^p$ has a distortion $1 \leq \delta \leq \sqrt{p}$ if and only if*

$$\forall x \in \Gamma,\quad \|x\|_{\ell_1} \leq \sqrt{p}\,\|x\|_{\ell_2} \leq \delta\|x\|_{\ell_1}.$$

A long standing issue in approximation theory in Banach spaces is to find "almost"-Euclidean sections of the unit ℓ_1-ball, i.e. subspaces with a distortion δ close to 1 and a dimension close to p. In particular, we recall that it has been established [Kas77] that, with an overwhelming probability, a random subspace of dimension $p-n$ (with respect to the Haar measure on the Grassmannian) satisfies

$$\delta \leq C\left(\frac{p(1+\log(p/n))}{n}\right)^{1/2} \tag{6}$$

where $C > 0$ is a universal constant. In other words, it was shown that, for all $n \leq p$, there exists a subspace Γ_n of dimension $p-n$ such that, for all $x \in \Gamma_n$,

$$\|x\|_{\ell_2} \leq C\left(\frac{1+\log(p/n)}{n}\right)^{1/2}\|x\|_{\ell_1}.$$

Remark. Hence, our framework deals also with unitary invariant random matrices. For instance, the matrices with i.i.d. Gaussian entries. Observe that their distortion satisfies (6).

Recently, new **deterministic** constructions of "almost"-Euclidean sections of the ℓ_1-ball have been given. Most of them can be viewed as related to the context of error-correcting codes. Indeed, the construction of [Ind07] is based on amplifying the minimum distance of a code using expanders. While the construction of [GLR08] is based on Low-Density Parity Check (LDPC) codes. Finally, the construction of [IS10] is related to the tensor product of error-correcting codes. The main reason of this surprising fact is that the vectors of a subspace of low distortion must be "well-spread", i.e. a small subset of its coordinates cannot contain most of its ℓ_2-norm (cf [Ind07, GLR08]). This property is required from a good error-correcting code, where the weight (i.e. the ℓ_0-norm) of each codeword cannot be concentrated on a small subset of its coordinates. Similarly, this property was intensively studied in Compressed Sensing; see for instance the Nullspace Property in [CDD09].

Remark. The main point of this article is that all of these deterministic constructions give efficient designs for the lasso and the Dantzig selector.

1.3. **The Universal Distortion Property.** In the past decade, numerous conditions have been given to prove oracle inequalities for the lasso and the Dantzig selector. An overview of important conditions can be found in [vdGB09]. We introduce a new condition, the Universal Distortion Property (UDP).

Definition 2 (UDP(S_0, κ_0, Δ)) — *Given $1 \leq S_0 \leq p$ and $0 < \kappa_0 < 1/2$, we say that a matrix $X \in \mathbb{R}^{n\times p}$ satisfies the universal distortion condition of order S_0, magnitude κ_0 and parameter Δ* **if and only if** *for all $\gamma \in \mathbb{R}^p$, for all integers $s \in \{1, \ldots, S_0\}$, for all subsets $\mathcal{S} \subseteq \{1, \ldots, p\}$ such that $|\mathcal{S}| = s$, it holds*

$$\|\gamma_{\mathcal{S}}\|_{\ell_1} \leq \Delta\sqrt{s}\,\|X\gamma\|_{\ell_2} + \kappa_0\|\gamma\|_{\ell_1}. \tag{7}$$

Remark. – Observe that the design X is not normalized. Equation (8) in Theorem 1.2 shows that Δ can depend on the inverse of the smallest singular value of X. Hence the quantity $\Delta\|X\gamma\|_{\ell_2}$ is scalar invariant.
– The UDP condition is similar to the Magic Condition [BLPR11] and the Compatibility Condition [vdGB09].

The main point of this article is that UDP is verifiable *as soon as one can give an upper bound on the distortion of the kernel of the design matrix*; see Theorem 1.2. Hence, instead of proving that a sufficient condition (such as RIP [CRT06], REC [BRT09], Compatibility [vdGB09], ...) holds it is sufficient to compute the distortion and the largest singular value of the design. Especially as these conditions can be hard to prove for a given matrix. We recall that an open problem is to find a computationally efficient algorithm that can tell if a given matrix satisfies the RIP condition [CRT06] or not.

We call the property "Universal Distortion" because it is satisfied by all the full rank matrices (Universal) and the parameters S_0 and Δ can be expressed in terms of the distortion of the kernel Γ of X:

Theorem 1.1 — *Let $X \in \mathbb{R}^{n\times p}$ be a full rank matrix. Denote by δ the distortion of its kernel:*

$$\delta = \sup_{\gamma\in\ker(X)} \frac{\|\gamma\|_{\ell_1}}{\sqrt{p}\|\gamma\|_{\ell_2}},$$

and ρ_n its smallest singular value. Then, for all $\gamma \in \mathbb{R}^p$,

$$\|\gamma\|_{\ell_2} \leq \frac{\delta}{\sqrt{p}}\|\gamma\|_{\ell_1} + \frac{2\delta}{\rho_n}\|X\gamma\|_{\ell_2}.$$

Equivalently, we have $\mathcal{B} := \{\gamma \in \mathbb{R}^p \mid (\delta/\sqrt{p})\|\gamma\|_{\ell_1} + (2\delta/\rho_n)\|X\gamma\|_{\ell_2} \leq 1\} \subset B_2^p$, where B_2^p denotes the Euclidean unit ball in $\mathbb{R}^p$.

This result implies that every full rank matrix satisfies UDP with parameters described as follows.

Theorem 1.2 — *Let $X \in \mathbb{R}^{n\times p}$ be a full rank matrix. Denote by δ the distortion of its kernel and ρ_n its smallest singular value. Let $0 < \kappa_0 < 1/2$ then X satisfies* UDP(S_0, κ_0, Δ) *where*

$$S_0 = \Big(\frac{\kappa_0}{\delta}\Big)^2 p \quad \text{and} \quad \Delta = \frac{2\delta}{\rho_n}. \tag{8}$$

This theorem is sharp in the following sense. The parameter S_0 represents (see Theorem 2.1) the maximum number of coefficients that can be recovered using lasso, we call it the *sparsity* *level*. It is known [CDD09] that the best bound one could expect is $S_{opt} \approx n/\log(p/n)$, up to a multiplicative constant. In the case where (6) holds, the sparsity level satisfies

$$S_0 \approx \kappa_0^2\, S_{opt}. \tag{9}$$

It shows that any design matrix with low distortion satisfies UDP with an optimal sparsity level.

2. Oracle inequalities

The results presented here fold into two parts. In the first part we assume only that UDP holds. In particular, it is not excluded that one can get better upper bounds on the parameters than Theorem 1.2. As a matter of fact, the smaller Δ is, the sharper the oracle inequalities are. Then, we give oracle inequalities in terms of only the distortion of the design.

Theorem 2.1 — *Let $X \in \mathbb{R}^{n\times p}$ be a full column rank matrix. Assume that X satisfies* $\mathrm{UDP}(S_0,\kappa_0,\Delta)$ *and that* (1) *holds. Then for any*

$$\lambda_\ell > \lambda_n^0/(1-2\kappa_0), \tag{10}$$

it holds

$$\left\|\beta^\ell-\beta^\star\right\|_{\ell_1} \leq \frac{2}{\left(1-\frac{\lambda_n^0}{\lambda_\ell}\right)-2\kappa_0} \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}} \Big(\lambda_\ell\,\Delta^2\, s + \left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\Big). \tag{11}$$

— *For every full column rank matrix $X \in \mathbb{R}^{n\times p}$, for all $0<\kappa_0<1/2$ and λ_ℓ satisfying* (10), *we have*

$$\left\|\beta^\ell-\beta^\star\right\|_{\ell_1} \leq \frac{2}{\left(1-\frac{\lambda_n^0}{\lambda_\ell}\right)-2\kappa_0} \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq(\kappa_0/\delta)^2 p.}} \Big(\lambda_\ell\cdot\frac{4\,\delta^2}{\rho_n^2}\cdot s + \left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\Big), \tag{12}$$

where ρ_n denotes the smallest singular value of X and δ the distortion of its kernel.

✧ Consider the case where the noise satisfies the hypothesis of Lemma A.1 and take $\lambda_n^0=\lambda_n^0(1)$. Assume that κ_0 is constant (say $\kappa_0=1/3$) and take $\lambda_\ell=3\lambda_n^0$; then (11) becomes

$$\left\|\beta^\ell-\beta^\star\right\|_{\ell_1} \leq 12 \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}} \Big(6\left\|X\right\|_{\ell_2,\infty}\cdot\Delta^2\sqrt{\log p}\cdot\sigma_n\, s + \left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\Big),$$

which is an oracle inequality up to a multiplicative factor $\Delta^2\sqrt{\log p}$. In the same way, (12) becomes

$$\left\|\beta^\ell-\beta^\star\right\|_{\ell_1} \leq 12 \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq p/9\delta^2.}} \Big(24\left\|X\right\|_{\ell_2,\infty}\cdot\frac{\delta^2\sqrt{\log p}}{\rho_n}\cdot\frac{1}{\rho_n}\,\sigma_n\, s + \left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\Big),$$

which is an oracle inequality up to a multiplicative factor $C_{mult} := (\delta^2\sqrt{\log p})/\rho_n$.
✧ In the optimal case (6), this latter becomes:

$$C_{mult} = C\cdot\frac{p\,(1+\log(p/n))\,\sqrt{\log p}}{n\,\rho_n}, \tag{13}$$

where $C>0$ is the same universal constant as in (6). Roughly speaking, up to a factor of the order of (13), the lasso is as good as the oracle that knows the S_0-best term approximation of the target. Moreover, as mentioned in (9), S_0 is an optimal sparsity level. However, this multiplicative constant takes small values for a restrictive range of the parameter n. As a matter of fact, it is meaningful when n is a constant fraction of p.

Similarly, we shows oracle inequalities in error prediction in terms of the distortion of the kernel of the design.

Theorem 2.2 — *Let $X \in \mathbb{R}^{n\times p}$ be a full column rank matrix. Assume that X satisfies* UDP(S_0, κ_0, Δ) *and that* (1) *holds. Then for any*

$$\lambda_\ell > \lambda_n^0/(1-2\kappa_0), \tag{10}$$

it holds

$$\left\|X\beta^\ell - X\beta^\star\right\|_{\ell_2} \leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}} \left[4\lambda_\ell\,\Delta\sqrt{s} + \frac{\left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}}{\Delta\sqrt{s}}\right]. \tag{14}$$

— *For every full column rank matrix $X \in \mathbb{R}^{n\times p}$, for all $0<\kappa_0<1/2$ and λ_ℓ satisfying* (10)*, we have*

$$\left\|X\beta^\ell - X\beta^\star\right\|_{\ell_2} \leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq(\kappa_0/\delta)^2 p.}} \left[4\lambda_\ell\cdot\frac{2\,\delta}{\rho_n}\cdot\sqrt{s} + \frac{1}{2\delta\sqrt{s}}\cdot\rho_n\left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\right], \tag{15}$$

where ρ_n denotes the smallest singular value of X and δ the distortion of its kernel.

✧ Consider the case where the noise satisfies the hypothesis of Lemma A.1 and take $\lambda_n^0 = \lambda_n^0(1)$. Assume that κ_0 is constant (say $\kappa_0 = 1/3$) and take $\lambda_\ell = 3\lambda_n^0$ then (14) becomes

$$\left\|X\beta^\ell - X\beta^\star\right\|_{\ell_2} \leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}} \left[24\left\|X\right\|_{\ell_2,\infty}\cdot\Delta\sqrt{\log p}\cdot\sigma_n\sqrt{s} + \frac{\left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}}{\Delta\sqrt{s}}\right],$$

which is not an oracle inequality *stricto sensu* because of $1/(\Delta\sqrt{s})$ in the second term. As a matter of fact, it tends to lower the s-best term approximation term $\left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}$. Nevertheless, it is "almost" an oracle inequality up to a multiplicative factor of the order of $\Delta\sqrt{\log p}$. In the same way, (15) becomes

$$\left\|X\beta^\ell - X\beta^\star\right\|_{\ell_2} \leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq p/9\delta^2.}} \left[48\left\|X\right\|_{\ell_2,\infty}\cdot\frac{\delta\sqrt{\log p}}{\rho_n}\cdot\frac{1}{\rho_n}\sigma_n\sqrt{s} + \frac{1}{2\delta\sqrt{s}}\cdot\rho_n\left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\right],$$

which is an oracle inequality up to a multiplicative factor $C'_{mult} := (\delta\sqrt{\log p})/\rho_n$.
✧ In the optimal case (6), this latter becomes:

$$C'_{mult} = C\cdot\frac{(p\log p\,(1+\log(p/n)))^{1/2}}{\rho_n\sqrt{n}}, \tag{16}$$

where $C>0$ is the same universal constant as in (6).

2.1. **Results for the Dantzig selector.** Similarly, we derive the same results for the Dantzig selector. The only difference is that the parameter κ_0 must be less than $1/4$. Here again the results fold into two parts. In the first one, we only assume that UDP holds. In the second, we invoke Theorem 1.2 to derive results in terms of the distortion of the design.

Theorem 2.3 — *Let $X \in \mathbb{R}^{n\times p}$ be a full column rank matrix. Assume that X satisfies* UDP(S_0, κ_0, Δ) *with $\kappa_0 < 1/4$ and that* (1) *holds. Then for any*

$$\lambda_d > \lambda^0/(1-4\kappa_0), \tag{17}$$

it holds

$$\left\|\beta^d - \beta^\star\right\|_{\ell_1} \leq \frac{4}{(1-\frac{\lambda^0}{\lambda_d})-4\kappa_0}\min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}}\left(\lambda_d\,\Delta^2\,s + \left\|\beta^\star_{\mathcal{S}^c}\right\|_{\ell_1}\right). \tag{18}$$

— *For every full column rank matrix $X \in \mathbb{R}^{n\times p}$, for all $0<\kappa_0<1/4$ and λ_d satisfying* (17)*, we have*

$$\|\beta^d-\beta^\star\|_{\ell_1} \leq \frac{4}{\big(1-\frac{\lambda^0}{\lambda_d}\big)-4\kappa_0} \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq(\kappa_0/\delta)^2p.}} \Big(\lambda_d\cdot\frac{4\,\delta^2}{\rho_n^2}\cdot s+\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}\Big), \tag{19}$$

where ρ_n denotes the smallest singular value of X and δ the distortion of its kernel.

The prediction error is given by the following theorem.

Theorem 2.4 — *Let $X\in\mathbb{R}^{n\times p}$ be a full column rank matrix. Assume that X satisfies* UDP(S_0,κ_0,Δ) *with $\kappa_0<1/4$ and that* (1) *holds. Then for any*

$$\lambda_d>\lambda^0/(1-4\kappa_0), \tag{17}$$

it holds

$$\|X\beta^d-X\beta^\star\|_{\ell_2}\leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\ s\leq S_0.}} \left[4\lambda_d\,\Delta\sqrt{s}+\frac{\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}}{\Delta\sqrt{s}}\right]. \tag{20}$$

— *For every full column rank matrix $X\in\mathbb{R}^{n\times p}$, for all $0<\kappa_0<1/4$ and λ_d satisfying* (10)*, we have*

$$\|X\beta^d-X\beta^\star\|_{\ell_2}\leq \min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\},\\ |\mathcal{S}|=s,\\ s\leq(\kappa_0/\delta)^2p.}} \left[4\lambda_d\cdot\frac{2\,\delta}{\rho_n}\cdot\sqrt{s}+\frac{1}{2\delta\sqrt{s}}\cdot\rho_n\,\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}\right], \tag{21}$$

where ρ_n denotes the smallest singular value of X and δ the distortion of its kernel.

Observe that the same comments as in the lasso case (e.g. (13), (16)) hold. Eventually, every deterministic construction of almost-Euclidean sections gives design that satisfies the oracle inequalities above.

3. An overview of the standard results

Oracle inequalities for the lasso and the Dantzig selector have been established under a variety of different conditions on the design. In this section, we show that the UDP condition is comparable to the standard conditions (RIP, REC and Compatibility) and that our results are relevant in the literature on the high-dimensional regression.

3.1. The standard conditions.

We recall some sufficient conditions here. For all $s\in\{1,\dots,p\}$, we denote by $\Sigma_s\subseteq\mathbb{R}^p$ the set of all the s-sparse vectors.

- ✦ **Restricted Isoperimetric Property:** A matrix $X\in\mathbb{R}^{n\times p}$ satisfies $RIP(\theta_S)$ if and only if there exists $0<\theta_S<1$ (as small as possible) such that for all $s\in\{1,\dots,S\}$, for all $\forall\gamma\in\Sigma_s$, it holds
$$(1-\theta_S)\|\gamma\|_{\ell_2}^2\leq\|X\gamma\|_{\ell_2}^2\leq(1+\theta_S)\|\gamma\|_{\ell_2}^2.$$
The constant θ_S is called the S-restricted isometry constant.
- ✦ **Restricted Eigenvalue Assumption** [BRT09]**:** A matrix X satisfies $RE(S,c_0)$ if and only if
$$\kappa(S,c_0)=\min_{\substack{\mathcal{S}\subseteq\{1,\dots,p\}\\ |\mathcal{S}|\leq S}}\ \min_{\substack{\gamma\neq0\\ \|\gamma_{\mathcal{S}^c}\|_{\ell_1}\leq c_0\|\gamma_{\mathcal{S}}\|_{\ell_1}}}\frac{\|X\gamma\|_{\ell_2}}{\|\gamma_{\mathcal{S}}\|_{\ell_2}}>0\,.$$
The constant $\kappa(S,c_0)$ is called the (S,c_0)-restricted ℓ_2-eigenvalue.

- **Compatibility Condition** [vdGB09]**:** We say that a matrix $X \in \mathbb{R}^{n\times p}$ satisfies the condition *Compatibility*(S, c_0) if and only if

$$\phi(S,c_0)=\min_{\substack{\mathcal{S}\subseteq\{1,\ldots,p\}\\ |\mathcal{S}|\leq S}}\ \min_{\substack{\gamma\neq 0\\ \|\gamma_{\mathcal{S}^c}\|_{\ell_1}\leq c_0\|\gamma_{\mathcal{S}}\|_{\ell_1}}}\frac{\sqrt{|\mathcal{S}|}\|X\gamma\|_{\ell_2}}{\|\gamma_{\mathcal{S}}\|_{\ell_1}}>0\,.$$

The constant $\phi(S, c_0)$ is called the (S, c_0)-restricted ℓ_1-eigenvalue.
- $\mathbf{H}_{S,1}$ **Condition** [JN10]**:** $X \in \mathbb{R}^{n\times p}$ satisfies the $\mathbf{H}_{S,1}(\kappa)$ condition (with $\kappa < 1/2$) if and only if for all $\gamma \in \mathbb{R}^p$ and for all $\mathcal{S} \subseteq \{1, \ldots , p\}$ such that $|\mathcal{S}| \leq S$, it holds

$$\big\|\gamma_{\mathcal{S}}\big\|_{\ell_1}\leq\hat{\lambda}\,S\,\|X\gamma\|_{\ell_2}+\kappa\|\gamma\|_{\ell_1}\,, \tag{22}$$

where $\hat{\lambda}$ denotes the maximum of the ℓ_2-norms of the columns in X.

Remark. The first term of the right hand side (i.e. $s\,\|X\gamma\|_{\ell_2}$) is greater than the first term of the right hand side of the UDP condition (i.e. $\sqrt{s}\,\|X\gamma\|_{\ell_2}$). Hence the $\mathbf{H}_{S,1}$ condition is weaker than the UDP condition. Nevertheless, the authors [JN10] established limits of performance on their conditions: the condition $\mathbf{H}_{s,\infty}(1/3)$ (that implies $\mathbf{H}_{s,1}(1/3)$) is feasible only in a severe restricted range of the sparsity parameter s. Notice that this is not the case of the UDP condition, the equality (9) shows that it is feasible for a large range of the sparsity parameter s. Moreover, a comparison of the two approaches is given in Table 1.

Let us emphasize that the above description is not meant to be exhaustive. In particular we do not mention the irrepresentable condition [ZY06] which ensures exact recovery of the support.

The next proposition shows that the UDP condition is weaker than the RIP, RE and Compatibility conditions.

Proposition 3.1 — *Let $X \in \mathbb{R}^{n\times p}$ be a full column rank matrix; then the following is true:*
- *The* RIP(θ_{5S}) *condition with* $\theta_{5S} < \sqrt{2} - 1$ *implies* UDP(S, κ_0, Δ) *for all pairs* (κ_0, Δ) *such that*

$$\Big[1+2\Big[\frac{1-\theta_{5S}}{1+\theta_{5S}}\Big]^{\frac12}\Big]^{-1}<\kappa_0<\frac{1}{2}\,,\ \text{and}\ \Delta\geq\Big[\sqrt{1-\theta_{5S}}+\frac{\kappa_0-1}{2\kappa_0}\sqrt{1+\theta_{5S}}\Big]^{-1}. \tag{23}$$

- *The RE(S, c_0) condition implies* UDP$(S, c_0, \kappa(S, c_0)^{-1})$.
- *The Compatibility(S, c_0) condition implies* UDP$(S, c_0, \phi(S, c_0)^{-1})$.

Remark. The point here is to show that the UDP condition is similar to the standard conditions of the high-dimensional regression. For the sake of simplicity, we do not study that if the converse of Proposition 3.1 is true. As a matter of fact, the UDP, RE and Compatibility conditions are expressions with the same flavor: they aim at controlling the eigenvalues of X on a cone:

$$\{\gamma\in\mathbb{R}^p\quad|\quad\forall\,S\in\{1,\ldots,p\}\,,\ \text{s.t. }|S|\leq s\,,\ \big\|\gamma_{S^c}\big\|_{\ell_1}\leq c\big\|\gamma_S\big\|_{\ell_1}\}\,,$$

where $c > 0$ is a tunning parameter.

3.2. **The results.** Table 1 shows that our results are similar to standard results in the literature.

Appendix A. Appendix

The appendix is devoted to the proof of the different results of this paper.

Lemma A.1 — *Suppose that $\varepsilon = (\varepsilon_i)_{i=1}^n$ is such that the ε_i's are i.i.d with respect to a Gaussian distribution with mean zero and variance σ_n^2. Choose $t \geq 1$ and set*

$$\lambda_n^0(t)=(1+t)\cdot\big\|X\big\|_{\ell_{2,\infty}}\cdot\sigma_n\cdot\sqrt{\log p},$$

Reference	Condition	Risk	Prediction
[CP09]	RIP	$\ell_2^2 \lesssim \sigma^2 s \log p + \ell_2^2$	$L_2^2 \lesssim \sigma^2 s \log p + L_2^2$
[BRT09]	REC	$\ell_1 \lesssim \sigma s \sqrt{\log p}$ (*)	$L_2 \lesssim \sigma \sqrt{s \log p}$ (*)
[vdGB09]	Comp.	$\ell_1 \lesssim \sigma s \sqrt{\log p}$ (*)	$L_2 \lesssim \sigma \sqrt{s \log p}$ (*)
[JN10]	$\mathbf{H}_{S,1}$	$\ell_1 \lesssim \sigma s^2 \sqrt{\log p} + \ell_1$	No
This article	UDP	$\ell_1 \lesssim \sigma s \sqrt{\log p} + \ell_1$	$L_2 \lesssim \sigma \sqrt{s \log p} + \ell_1/\sqrt{s}$

where notations are given by:

Location	ℓ_1	ℓ_2	L_2
Right hand side	$\lVert\beta^\star_{\mathcal{S}^c}\rVert_{\ell_1}$	$\lVert\beta^\star_{\mathcal{S}^c}\rVert_{\ell_2}$	$\lVert X\beta^{ideal} - X\beta^\star\rVert_{\ell_2}$
Left hand side	$\lVert\hat\beta - \beta^\star\rVert_{\ell_1}$	$\lVert\hat\beta - \beta^\star\rVert_{\ell_2}$	$\lVert X\hat\beta - X\beta^\star\rVert_{\ell_2}$

TABLE 1. Comparison of results in risk and prediction for the Lasso and the Dantzig selector. Observe that all the inequalities are satisfied with an overwhelming probability. The $\lesssim$ notation means that the inequality holds up to a multiplicative factor that may depends on the parameters of the condition. The (*) notation means that the result is given for s-sparse targets. The $\hat\beta$ notation represents the estimator (i.e. the lasso or the Dantzig selector). The parameters σ and p represent respectively the standard deviation of the noise and the dimension of the target vector $\beta^\star$.

where $\lVert X\rVert_{\ell_2,\infty}$ *denotes the maximum* ℓ_2*-norm of the columns of* X*. Then,*

$$\mathbb{P}\big(\lVert X^\top\varepsilon\rVert_{\ell_\infty} \leq \lambda_n^0(t)\big) \geq 1 - \sqrt{2}/\Big[(1+t)\sqrt{\pi \log p}\; p^{\frac{(1+t)^2}{2}-1}\Big]\,.$$

Proof of Lemma A.1 — Observe that $X^\top\varepsilon \sim \mathcal{N}_p(0, \sigma_n^2\, X^\top X)$. Hence, $\forall j = 1, \ldots, p, \quad X_j^\top\varepsilon \sim \mathcal{N}\big(0, \sigma_n^2 \lVert X_j\rVert_{\ell_2}^2\big)$. Using Šidák's inequality [Sid68], it yields

$$\mathbb{P}\big(\lVert X^\top\varepsilon\rVert_{\ell_\infty} \leq \lambda_n^0\big) \geq \mathbb{P}\big(\lVert\widetilde{\varepsilon}\rVert_{\ell_\infty} \leq \lambda_n^0\big) = \prod_{i=1}^{p} \mathbb{P}\left(|\widetilde{\varepsilon}_i| \leq \lambda_n^0\right)\,,$$

where the $\widetilde{\varepsilon}_i$'s are i.i.d. with respect to $\mathcal{N}\big(0, \sigma_n^2 \lVert X\rVert_{\ell_2,\infty}^2\big)$. Denote by Φ and φ respectively the cumulative distribution function and the probability density function of the standard normal. Set $\theta = (1+t)\sqrt{\log p}$. It holds

$$\prod_{i=1}^{p} \mathbb{P}\left(|\widetilde{\varepsilon}_i| \leq \lambda_n^0\right) = \mathbb{P}\left(|\varepsilon_1| \leq \lambda_n^0\right)^p = (2\Phi(\theta) - 1)^p > \big(1 - 2\varphi(\theta)/\theta\big)^p\,,$$

using an integration by parts to get $1 - \Phi(\theta) < \varphi(\theta)/\theta$. It yields that

$$\mathbb{P}\big(\lVert X^\top\varepsilon\rVert_{\ell_\infty} \leq \lambda_n^0\big) \geq \big(1 - 2\varphi(\theta)/\theta\big)^p \geq 1 - 2p\frac{\varphi(\theta)}{\theta} = 1 - \frac{\sqrt{2}}{(1+t)\sqrt{\pi \log p}\; p^{\frac{(1+t)^2}{2}-1}}\,.$$

This concludes the proof. □

Proof of Theorem 1.2 — Consider the following singular value decomposition $X = U^\top D A$ where

✧ $U \in \mathbb{R}^{n\times n}$ is such that $UU^\top = \mathrm{Id}_n$,

- ✧ $D = \mathrm{Diag}(\rho_1, \ldots, \rho_n)$ is a diagonal matrix where $\rho_1 \geq \cdots \geq \rho_n > 0$ are the singular values of X,
- ✧ and $A \in \mathbb{R}^{n\times p}$ is such that $AA^\top = \mathrm{Id}_n$.

We recall that the only assumption on the design is that it has full column rank which yields that $\rho_n > 0$. Let δ be the distortion of the kernel Γ of the design. Denote by π_Γ (resp. $\pi_{\Gamma^\perp}$) the ℓ_2-projection onto Γ (resp. $\Gamma^\perp$). Let $\gamma \in \mathbb{R}^p$; then $\gamma = \pi_\Gamma(\gamma) + \pi_{\Gamma^\perp}(\gamma)$. An easy calculation shows that $\pi_{\Gamma^\perp}(\gamma) = A^\top A\gamma$. Let $s \in \{1, \ldots, S\}$ and let $\mathcal{S} \subseteq \{1, \ldots, p\}$ be such that $|\mathcal{S}| = s$. It holds,

$$\begin{aligned}
\|\gamma_{\mathcal{S}}\|_{\ell_1} &\leq \sqrt{s}\|\gamma\|_{\ell_2} = \sqrt{s}\,\|\pi_\Gamma(\gamma)\|_{\ell_2} + \sqrt{s}\,\|\pi_{\Gamma^\perp}(\gamma)\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\|\pi_\Gamma(\gamma)\|_{\ell_1} + \sqrt{s}\,\|A^\top A\gamma\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\big(\|\gamma\|_{\ell_1} + \|(\pi_{\Gamma^\perp}(\gamma))\|_{\ell_1}\big) + \sqrt{s}\,\|A\gamma\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\|\gamma\|_{\ell_1} + \delta\,\sqrt{s}\,\|A^\top A\gamma\|_{\ell_2} + \sqrt{s}\,\|A\gamma\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\|\gamma\|_{\ell_1} + (1+\delta)\,\sqrt{s}\,\|A\gamma\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\|\gamma\|_{\ell_1} + \frac{1+\delta}{\rho_n}\,\sqrt{s}\,\|X\gamma\|_{\ell_2}, \\
&\leq \frac{\sqrt{s}}{\sqrt{p}}\,\delta\,\|\gamma\|_{\ell_1} + \frac{2\delta}{\rho_n}\,\sqrt{s}\,\|X\gamma\|_{\ell_2},
\end{aligned}$$

using the triangular inequality and the distortion of the kernel Γ. Eventually, set $\kappa_0 = (\sqrt{S}/\sqrt{p})\,\delta$ and $\Delta = 2\delta/\rho_n$. This ends the proof. □

Proof of Theorem 2.1 — We recall that λ_n^0 denotes an upper bound on the amplification of the noise; see (1). We begin with a standard result.

Lemma A.2 — *Let $h = \beta^\ell - \beta^\star \in \mathbb{R}^p$ and $\lambda_\ell \geq \lambda_n^0$. Then, for all subsets $\mathcal{S} \subseteq \{1, \ldots, p\}$, it holds,*

$$\frac{1}{2\lambda_\ell}\Big[\frac{1}{2}\|Xh\|_{\ell_2}^2 + (\lambda_\ell - \lambda_n^0)\|h\|_{\ell_1}\Big] \leq \|h_{\mathcal{S}}\|_{\ell_1} + \|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}. \tag{A.1}$$

Proof. By optimality, we have

$$\frac{1}{2}\|X\beta^\ell - y\|_{\ell_2}^2 + \lambda_\ell\|\beta^\ell\|_{\ell_1} \leq \frac{1}{2}\|X\beta^\star - y\|_{\ell_2}^2 + \lambda_\ell\|\beta^\star\|_{\ell_1}.$$

It yields

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 - \langle X^\top\varepsilon, h\rangle + \lambda_\ell\|\beta^\ell\|_{\ell_1} \leq \lambda_\ell\|\beta^\star\|_{\ell_1}.$$

Let $\mathcal{S} \subseteq \{1, \ldots, p\}$; we have

$$\begin{aligned}
\frac{1}{2}\|Xh\|_{\ell_2}^2 + \lambda_\ell\|\beta^\ell_{\mathcal{S}^c}\|_{\ell_1} &\leq \lambda_\ell\big(\|\beta^\star_{\mathcal{S}}\|_{\ell_1} - \|\beta^\ell_{\mathcal{S}}\|_{\ell_1}\big) + \lambda_\ell\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1} + \langle X^\top\varepsilon, h\rangle, \\
&\leq \lambda_\ell\|h_{\mathcal{S}}\|_{\ell_1} + \lambda_\ell\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1} + \lambda_n^0\|h\|_{\ell_1},
\end{aligned}$$

using (1). Adding $\lambda_\ell\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}$ on both sides, it holds

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 + (\lambda_\ell - \lambda_n^0)\|h_{\mathcal{S}^c}\|_{\ell_1} \leq (\lambda_\ell + \lambda_n^0)\|h_{\mathcal{S}}\|_{\ell_1} + 2\lambda_\ell\|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}.$$

Adding $(\lambda_\ell - \lambda_n^0)\big\|h_{\mathcal{S}}\big\|_{\ell_1}$ on both sides, we conclude the proof. $\square$

Using (7) and (A.1), it follows that

$$\text{(A.2)} \qquad \frac{1}{2\lambda_\ell}\Big[\frac{1}{2}\big\|Xh\big\|_{\ell_2}^2 + (\lambda_\ell - \lambda_n^0)\big\|h\big\|_{\ell_1}\Big] \leq \Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2} + \kappa_0\big\|h\big\|_{\ell_1} + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

It yields,

$$\begin{aligned}\Big[\frac{1}{2}\Big(1 - \frac{\lambda_n^0}{\lambda_\ell}\Big) - \kappa_0\Big]\big\|h\big\|_{\ell_1} &\leq \Big(-\frac{1}{4\lambda_\ell}\big\|Xh\big\|_{\ell_2}^2 + \Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2}\Big) + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1},\\ &\leq \lambda_\ell\,\Delta^2\,s + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1},\end{aligned}$$

using the fact that the polynomial $x \mapsto -(1/4\lambda_\ell)\,x^2 + \Delta\sqrt{s}\,x$ is not greater than $\lambda_\ell\,\Delta^2\,s$. This concludes the proof. $\square$

Proof of Theorem 2.3 — We begin with a standard result.

Lemma A.3 — *Let $h = \beta^\ell - \beta^\star \in \mathbb{R}^p$ and $\lambda_\ell \geq \lambda_n^0$. Then, for all subsets $\mathcal{S} \subseteq \{1, \ldots, p\}$, it holds,*

$$\text{(A.3)} \qquad \frac{1}{4\lambda_d}\Big[\big\|Xh\big\|_{\ell_2}^2 + (\lambda_d - \lambda_n^0)\big\|h\big\|_{\ell_1}\Big] \leq \big\|h_{\mathcal{S}}\big\|_{\ell_1} + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

Proof. Set $h = \beta^\star - \beta^d$. Recall that $\|X^\top\varepsilon\|_{\ell_\infty} \leq \lambda_n^0$, it yields

$$\begin{aligned}\|Xh\|_{\ell_2}^2 &\leq \|X^\top Xh\|_{\ell_\infty}\|h\|_{\ell_1} = \|X^\top(y - X\beta^d) + X^\top(X\beta^\star - y)\|_{\ell_\infty}\big\|h\big\|_{\ell_1}\\ &\leq (\lambda_d + \lambda_n^0)\big\|h\big\|_{\ell_1}.\end{aligned}$$

Hence we get

$$\text{(A.4)} \qquad \|Xh\|_{\ell_2}^2 - (\lambda_d + \lambda_n^0)\big\|h_{\mathcal{S}^c}\big\|_{\ell_1} \leq (\lambda_d + \lambda_n^0)\big\|h_{\mathcal{S}}\big\|_{\ell_1}.$$

Since $\beta^\star$ is feasible, it yields $\big\|\beta^d\big\|_{\ell_1} \leq \big\|\beta^\star\big\|_{\ell_1}$. Thus,

$$\|\beta^d_{\mathcal{S}^c}\|_{\ell_1} \leq \big(\|\beta^\star_{\mathcal{S}}\|_{\ell_1} - \|\beta^d_{\mathcal{S}}\|_{\ell_1}\big) + \|\beta^\star_{\mathcal{S}^c}\|_{\ell_1} \leq \|h_{\mathcal{S}}\|_{\ell_1} + \|\beta^\star_{\mathcal{S}^c}\|_{\ell_1}.$$

Since $\big\|h_{\mathcal{S}^c}\big\|_{\ell_1} \leq \big\|\beta^d_{\mathcal{S}^c}\big\|_{\ell_1} + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}$, it yields

$$\text{(A.5)} \qquad \big\|h_{\mathcal{S}^c}\big\|_{\ell_1} \leq \big\| + \big\|_{\ell_1 h_{\mathcal{S}}} 2\big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

Combining (A.4) $+ 2\lambda_d \cdot$ (A.5), we get

$$\|Xh\|_{\ell_2}^2 + (\lambda_d - \lambda_n^0)\big\|h_{\mathcal{S}^c}\big\|_{\ell_1} \leq (3\lambda_d + \lambda_n^0)\big\|h_{\mathcal{S}}\big\|_{\ell_1} + 4\lambda_d\big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

Adding $(\lambda_d - \lambda_n^0)\big\|h_{\mathcal{S}}\big\|_{\ell_1}$ on both sides, we conclude the proof. $\square$

Using (7) and (A.3), it follows that

$$\text{(A.6)} \qquad \frac{1}{4\lambda_\ell}\Big[\big\|Xh\big\|_{\ell_2}^2 + (\lambda_\ell - \lambda_n^0)\big\|h\big\|_{\ell_1}\Big] \leq \Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2} + \kappa_0\big\|h\big\|_{\ell_1} + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

It yields,

$$\begin{aligned}\Big[\frac{1}{4}\Big(1 - \frac{\lambda_n^0}{\lambda_\ell}\Big) - \kappa_0\Big]\big\|h\big\|_{\ell_1} &\leq \Big(-\frac{1}{4\lambda_\ell}\big\|Xh\big\|_{\ell_2}^2 + \Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2}\Big) + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1},\\ &\leq \lambda_\ell\,\Delta^2\,s + \big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1},\end{aligned}$$

using the fact that the polynomial $x \mapsto -(1/4\lambda_\ell)\,x^2 + \Delta\sqrt{s}\,x$ is not greater than $\lambda_\ell\,\Delta^2\,s$. This concludes the proof. $\square$

Proof of Theorem 2.2 and Theorem 2.4 — Using (A.2), we know that

$$\frac{1}{2\lambda_\ell}\Big[\frac{1}{2}\big\|Xh\big\|_{\ell_2}^2+(\lambda_\ell-\lambda_n^0)\big\|h\big\|_{\ell_1}\Big]\leq\Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2}+\kappa_0\big\|h\big\|_{\ell_1}+\big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

It follows that

$$\big\|Xh\big\|_{\ell_2}^2-4\lambda_\ell\,\Delta\sqrt{s}\,\big\|Xh\big\|_{\ell_2}\leq 4\lambda_\ell\,\big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}.$$

This latter is of the form $x^2-bx\leq c$ which implies that $x\leq b+c/b$. Hence,

$$\big\|Xh\big\|_{\ell_2}\leq 4\lambda_\ell\,\Delta\sqrt{s}+\frac{\big\|\beta^\star_{\mathcal{S}^c}\big\|_{\ell_1}}{\Delta\sqrt{s}}.$$

The same analysis holds for Theorem 2.4. □

Proof of Proposition 3.1 — One can check that $RE(S,c_0)$ implies $\mathrm{UDP}(S,c_0,\kappa(S,c_0)^{-1})$, and that $\mathit{Compatibility}(S,c_0)$ implies $\mathrm{UDP}(S,c_0,\phi(S,c_0)^{-1})$.

Assume that X satisfies $\mathrm{RIP}(\theta_{5S})$. Let $\gamma\in\mathbb{R}^p$, $s\in\{1,\ldots,S_0\}$, and $T_0\subseteq\{1,\ldots,p\}$ such that $|T_0|=s$. Choose a pair (κ_0,Δ) as in (23).

✧ If $\big\|\gamma_{T_0}\big\|_{\ell_1}\leq\kappa_0\big\|\gamma\big\|_{\ell_1}$ then $\big\|\gamma_{T_0}\big\|_{\ell_1}\leq\Delta\sqrt{s}\big\|X\gamma\big\|_{\ell_2}+\kappa_0\big\|\gamma\big\|_{\ell_1}$.

✧ Suppose that $\big\|\gamma_{T_0}\big\|_{\ell_1}>\kappa_0\big\|\gamma\big\|_{\ell_1}$ then

$$\big\|\gamma_{T_0^c}\big\|_{\ell_1}<\frac{1-\kappa_0}{\kappa_0}\big\|\gamma_{T_0}\big\|_{\ell_1}. \tag{A.7}$$

Denote by T_1 the set of the indices of the $4s$ largest coefficients (in absolute value) in T_0^c, denote by T_2 the set of the indices of the $4s$ largest coefficients in $(T_0\cup T_1)^c$, etc... Hence we decompose T_0^c into disjoint sets $T_0^c=T_1\cup T_2\cup\ldots\cup T_l$. Using (A.7), it yields

$$\sum_{i\geq 2}\big\|\gamma_{T_i}\big\|_{\ell_2}\leq(4s)^{-1/2}\sum_{i\geq 1}\big\|\gamma_{T_i}\big\|_{\ell_1}=(4s)^{-1/2}\big\|\gamma_{T_0^c}\big\|_{\ell_1}\leq\frac{1-\kappa_0}{2\kappa_0\sqrt{s}}\big\|\gamma_{T_0}\big\|_{\ell_1} \tag{A.8}$$

Using $\mathrm{RIP}(\theta_{5S})$ and (A.8), it follows that

$$\begin{aligned}
\big\|X\gamma\big\|_{\ell_2}&\geq\big\|X(\gamma_{(T_0\cup T_1)})\big\|_{\ell_2}-\sum_{i\geq 2}\big\|X(\gamma_{T_i})\big\|_{\ell_2},\\
&\geq\sqrt{1-\theta_{5S}}\,\big\|\gamma_{(T_0\cup T_1)}\big\|_{\ell_2}-\sqrt{1+\theta_{5S}}\sum_{i\geq 2}\big\|\gamma_{T_i}\big\|_{\ell_2},\\
&\geq\sqrt{1-\theta_{5S}}\,\big\|\gamma_{T_0}\big\|_{\ell_2}-\sqrt{1+\theta_{5S}}\,\frac{1-\kappa_0}{2\kappa_0}\,\frac{\big\|\gamma_{T_0}\big\|_{\ell_1}}{\sqrt{s}},\\
&\geq\Big[\sqrt{1-\theta_{5S}}+\frac{\kappa_0-1}{2\kappa_0}\sqrt{1+\theta_{5S}}\Big]\,\frac{\big\|\gamma_{T_0}\big\|_{\ell_1}}{\sqrt{s}},\\
&=\frac{\sqrt{1+\theta_{5S}}}{2\kappa_0}\Big[1+2\Big(\frac{1-\theta_{5S}}{1+\theta_{5S}}\Big)^{\frac{1}{2}}\Big]\Big[\kappa_0-\Big[1+2\Big(\frac{1-\theta_{5S}}{1+\theta_{5S}}\Big)^{\frac{1}{2}}\Big]^{-1}\Big]\,\frac{\big\|\gamma_{T_0}\big\|_{\ell_1}}{\sqrt{s}}.
\end{aligned}$$

The lower bound on κ_0 shows that the right hand side is positive. Observe that we took Δ such that this latter is exactly $\big\|\gamma_{T_0}\big\|_{\ell_1}/(\Delta\sqrt{s})$. Eventually, we get

$$\big\|\gamma_{T_0}\big\|_{\ell_1}\leq\Delta\sqrt{s}\big\|X\gamma\big\|_{\ell_2}\leq\Delta\sqrt{s}\big\|X\gamma\big\|_{\ell_2}+\kappa_0\big\|\gamma\big\|_{\ell_1}.$$

This ends the proofs. □

Acknowledgments — The author would like to thank Jean-Marc Azaïs and Franck Barthe for their support. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[BLPR11] K. Bertin, E. Le Pennec, and V. Rivoirard, *Adaptive Dantzig density estimation*, Ann. Inst. Henri Poincaré Probab. Stat. **47** (2011), no. 1, 43–74. MR 2779396 (2012a:62107)

[BRT09] P.J. Bickel, Y. Ritov, and A.B. Tsybakov, *Simultaneous analysis of lasso and Dantzig selector*, Ann. Statist. **37** (2009), no. 4, 1705–1732. MR 2533469 (2010j:62118)

[CDD09] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), no. 1, 211–231. MR 2449058 (2010d:94024)

[CP09] E.J. Candès and Y. Plan, *Near-ideal model selection by ℓ_1 minimization*, Ann. Statist. **37** (2009), no. 5A, 2145–2177. MR 2543688 (2010j:62017)

[CRT06] E.J. Candès, J.K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59** (2006), no. 8, 1207–1223. MR 2230846 (2007f:94007)

[CT07] E.J. Candès and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35** (2007), no. 6, 2313–2351. MR 2382644 (2009b:62016)

[DET06] D.L. Donoho, M. Elad, and V.N. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory **52** (2006), no. 1, 6–18. MR 2237332 (2007d:94007)

[Don06] D.L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), no. 4, 1289–1306. MR 2241189 (2007e:94013)

[GLR08] V. Guruswami, J. R. Lee, and A. Razborov, *Almost euclidean subspaces of ℓ_1 via expander codes*, Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2008, pp. 353–362.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman, *High-dimensional problems: p n*, The elements of statistical learning (2009), 1–50.

[Ind07] P. Indyk, *Uncertainty principles, extractors, and explicit embeddings of l2 into l1*, Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, ACM, 2007, pp. 615–620.

[IS10] P. Indyk and S. Szarek, *Almost-euclidean subspaces of l 1 n\ ell_1^ n via tensor products: A simple approach to randomness reduction*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (2010), 632–641.

[JN10] A. Juditsky and A. Nemirovski, *Accuracy guarantees for ℓ_1-recovery.*

[Kas77] B.S. Kashin, *The widths of certain finite-dimensional sets and classes of smooth functions*, Izv. Akad. Nauk SSSR Ser. Mat. **41** (1977), no. 2, 334–351, 478. MR 0481792 (58 #1891)

[KT07] B.S. Kashin and V.N. Temlyakov, *A remark on the problem of compressed sensing*, Mat. Zametki **82** (2007), no. 6, 829–837. MR 2399963 (2009k:41027)

[Sid68] Z. Sidák, *On multivariate normal probabilities of rectangles: Their dependence on correlations*, Ann. Math. Statist. **39** (1968), 1425–1434. MR 0230403 (37 #5965)

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288. MR 1379242 (96j:62134)

[vdGB09] S.A. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the Lasso*, Electron. J. Stat. **3** (2009), 1360–1392. MR 2576316

[ZY06] P. Zhao and B. Yu, *On model selection consistency of Lasso*, J. Mach. Learn. Res. **7** (2006), 2541–2563. MR 2274449

Yohann de Castro is with the Département de Mathématiques, Université Paris-Sud, Faculté des Sciences d'Orsay, 91405 Orsay, France. This article was written mostly during his Ph.D. at the Institut de Mathématiques de Toulouse.

E-mail address: `yohann.decastro@math.u-psud.fr`