Submitted to the Annals of Statistics

SIMULTANEOUS ANALYSIS OF LASSO AND DANTZIG SELECTOR*

By Peter J. Bickel , Ya'acov Ritov and Alexandre Tsybakov

We exhibit an approximate equivalence between the Lasso estimator and Dantzig selector. For both methods we derive parallel oracle inequalities for the prediction risk in the general nonparametric regression model, as well as bounds on the ℓ_p estimation loss for $1 \leq p \leq 2$ in the linear model when the number of variables can be much larger than the sample size.

1. Introduction. During the last few years a great deal of attention has been focused on the ℓ_1 penalized least squares (Lasso) estimator of parameters in high-dimensional linear regression when the number of variables can be much larger than the sample size [8–10, 15, 16, 18–20, 24, 25]. Quite recently, Candes and Tao [7] have proposed a new estimate for such linear models, the Dantzig selector, for which they establish optimal ℓ_2 rate properties under a sparsity scenario, i.e., when the number of non-zero components of the true vector of parameters is small.

Lasso estimators have been also studied in the nonparametric regression setup [2–5, 11, 12, 17]. In particular, Bunea et al. [2–5] obtain sparsity oracle inequalities for the prediction loss in this context and point out the implications for minimax estimation in classical non-parametric regression settings, as well as for the problem of aggregation of estimators. An analog of Lasso for density estimation with similar properties (SPADES) is proposed in [6]. Modified versions of Lasso estimators (non-quadratic terms and/or penalties slightly different from ℓ_1) for nonparametric regression with random design are suggested and studied under prediction loss in [13, 23]. Sparsity oracle inequalities for the Dantzig selector with random design are obtained in [14]. In linear fixed design regression, Meinshausen and Yu [16] establish a bound on the ℓ_2 loss for the coefficients of Lasso which is quite different from the bound on the same loss for the Dantzig selector proven in [7].

*Partially supported by NSF grant DMS-0605236, ISF Grant, and France-Berkeley Fund.

AMS 2000 subject classifications: Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20

Keywords and phrases: Linear models, Model selection, Nonparametric statistics

The main message of this paper is that under a sparsity scenario, the Lasso and the Dantzig selector exhibit similar behavior, both for linear regression and for nonparametric regression models, for ℓ_2 prediction loss and for ℓ_p loss in the coefficients for $1 \leq p \leq 2$. All the results of the paper are non-asymptotic.

Let us specialize to the case of linear regression with many covariates, $y = X\boldsymbol{\beta} + W$ where X is the $n \times M$ deterministic design matrix, with M possibly much larger than n, and W is a vector of i.i.d. standard normal random variables. This is the situation considered most recently by Candes and Tao [7] and Meinshausen and Yu [16]. Here sparsity specifies that the high-dimensional vector $\boldsymbol{\beta}$ has coefficients that are mostly 0. Our key observation is that the deviations from the true regression function of the Dantzig selector and of the Lasso estimate, with high probability lie in a region such that the contribution to their ℓ_1 loss from coordinates of $\boldsymbol{\beta}$ which vanish is of the same order as the contribution from those which do not.

We develop general tools to study these two estimators in parallel. For the fixed design Gaussian regression model we recover, as particular cases, sparsity oracle inequalities for the Lasso, as in Bunea et al. [4], and ℓ_2 bounds for the coefficients of Dantzig selector, as in Candes and Tao [7]. This is obtained as a consequence of more general results, which include:

- Sparsity oracle inequalities for the Dantzig selector in the nonparametric regression model under ℓ_2 prediction loss.
- Sparsity oracle inequalities for the Lasso in the nonparametric regression model under more general assumptions on the design matrix than in [4].
- An approximate equivalence between Lasso and Dantzig selector in nonparametric regression.
- We develop geometrical assumptions which are considerably weaker than those of Candes and Tao [7] for the Dantzig selector and Bunea et al. [4] for the Lasso. In the context of linear regression where the number of variables is possibly much larger than the sample size these assumptions imply the result of [7] for the ℓ_2 loss and generalize it to ℓ_p loss, $1 \leq p \leq 2$, and to prediction loss. Our bounds for the Lasso differ from those for Dantzig selector only in numerical constants.

We begin, in the next section, by defining the Lasso and Dantzig procedures and the notation. We then give some basic properties of the two procedures, introducing notation and two important technical lemmas. In Section 3 we develop our key geometric assumptions, and compare them to those of [7] and [16] as well as to ones appearing in [4] and [5]. We note a weakness of our assumptions, and hence also these of the authors we cited, and show also a way of remedying them. Sections 4, 5 give the equivalence results and sparsity oracle inequalities for the Lasso and Dantzig estimators in the general nonparametric regression model. Section 6 focuses on linear regression and includes a final discussion.

2. Basic properties of Lasso and Dantzig solutions. Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be a sample of independent random pairs with

$$Y_i = f(Z_i) + W_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $f : \mathcal{Z} \to \mathbb{R}$ is an unknown regression function to be estimated, $\mathcal{Z}$ is a Borel subset of $\mathbb{R}^d$, the Z_i's are fixed elements in $\mathcal{Z}$ and the regression errors W_i are Gaussian. Let $\mathcal{F}_M = \{f_1, \ldots, f_M\}$ be a finite dictionary of functions $f_j : \mathcal{Z} \to \mathbb{R}$, $j = 1, \ldots, M$. We assume throughout that $M \geq 2$. Depending on the statistical targets, the dictionary $\mathcal{F}_M$ can be of different nature. For instance, it can be a collection of basis functions used to approximate f in the nonparametric regression model. Another example is related to the aggregation problem where the f_j are estimators arising from M different methods. They can also correspond to M different values of the tuning parameter of the same method. Without much loss of generality, these estimators f_j are treated as fixed functions: the results are viewed as being conditioned on the sample the f_j are based on.

For any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, define $\mathsf{f}_\lambda(z) = \sum_{j=1}^M \lambda_j f_j(z)$. The estimates we consider are all of the form $\mathsf{f}_{\tilde\lambda}(\cdot)$ where $\tilde\lambda$ is data determined.

Let

$$M(\lambda) = \sum_{j=1}^M I_{\{\lambda_j \neq 0\}} = |J(\lambda)|$$

denote the number of non-zero coordinates of λ, where $I_{\{\cdot\}}$ denotes the indicator function, $J(\lambda) = \{j \in \{1, \ldots, M\} : \lambda_j \neq 0\}$, and $|J|$ denotes the cardinality of J. The value $M(\lambda)$ characterizes the *sparsity* of the vector λ: the smaller $M(\lambda)$, the “sparser” λ.

Introduce the residual sum of squares

$$\widehat{S}(\lambda) = \frac{1}{n}\sum_{i=1}^n \{Y_i - \mathsf{f}_\lambda(Z_i)\}^2,$$

for all $\lambda \in \mathbb{R}^M$. Define the Lasso solution $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)$ by

$$\widehat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \widehat{S}(\lambda) + 2r \sum_{j=1}^{M} \|f_j\|_n |\lambda_j| \right\}, \tag{2.2}$$

where $r > 0$ is some tuning constant, and introduce the corresponding Lasso estimator

$$\widetilde{f}(x) = \mathsf{f}_{\widehat{\lambda}}(x) = \sum_{j=1}^{M} \widehat{\lambda}_j f_j(z). \tag{2.3}$$

Here and below $\|\cdot\|_n$ stands for the empirical norm:

$$\|g\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} g^2(Z_i)}$$

for any $g : \mathcal{Z} \to \mathbb{R}$.

The criterion in (2.2) is convex in λ, so that standard convex optimization procedures can be used to compute $\widehat{\lambda}$. We refer to [9, 18, 19, 22] for detailed discussion of these optimization problems and fast algorithms.

For a vector $\mathbf{\Delta} \in \mathbb{R}^M$ and a subset $J \subset \{1, \ldots, M\}$ we denote by $\mathbf{\Delta}_J$ the vector in $\mathbb{R}^M$ which has the same coordinates as $\mathbf{\Delta}$ on J and zero coordinates on the complement J^c of J.

We also introduce the matrix $X = (f_j(Z_i))_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, M$ and the vectors $y = (Y_1, \ldots, Y_n)^T$, $\mathbf{f} = (f(Z_1), \ldots, f(Z_n))^T$, $W = (W_1, \ldots, W_n)^T$. We will write $|x|_p$ for the ℓ_p norm of $x \in \mathbb{R}^M$, $1 \le p \le \infty$.

With this notation,

$$y = \mathbf{f} + W.$$

The Dantzig estimator of the regression function f is defined by

$$\widetilde{f}_D(z) = \mathsf{f}_{\widehat{\lambda}_D}(z) = \sum_{j=1}^{M} \widehat{\lambda}_{j,D} f_j(z). \tag{2.4}$$

where $\widehat{\lambda}_D = (\widehat{\lambda}_{1,D}, \ldots, \widehat{\lambda}_{M,D})$ is the Dantzig selector, i.e., a solution of the minimization problem

$$\widehat{\lambda}_D = \arg\min \left\{ |\lambda|_1 : \ \left| \frac{1}{n} D^{-1/2} X^T (y - X\lambda) \right|_\infty \le r \right\} \tag{2.5}$$

with some $r > 0$ and the diagonal matrix

$$D = \mathrm{diag}\{\|f_1\|_n^2, \ldots, \|f_M\|_n^2\}.$$

Here and below we suppose that $\|f_j\|_n \neq 0$, $j = 1, \dots, M$. Set

$$f_{\max} = \max_{1\le j\le M} \|f_j\|_n\,, \quad f_{\min} = \min_{1\le j\le M} \|f_j\|_n\,.$$

The Dantzig selector is computationally feasible, since it reduces to a linear programming problem [7].

It is easy to see that the Lasso solution obeys the Dantzig constraint. In fact, the necessary and sufficient condition of the minimum in (2.2) is that 0 belongs to the subgradient of the convex function $\lambda \mapsto n^{-1}|y - X\lambda|_2^2 + 2r|D^{1/2}\lambda|_1$. This implies that the Lasso selector $\widehat{\lambda}$ satisfies the Dantzig constraint:

$$\Big|\frac{1}{n} D^{-1/2} X^T (y - X\widehat{\lambda})\Big|_\infty \le r. \tag{2.6}$$

Therefore, by the definition of Dantzig selector, we have $|\widehat{\lambda}_D|_1 \le |\widehat{\lambda}|_1$.

We conclude this section with two lemmata, whose proofs are given in the appendix.

LEMMA 1. *Let W_i be independent $\mathcal{N}(0,\sigma^2)$ random variables with $\sigma^2 > 0$ and let $\widetilde{f}$ be the Lasso estimator defined by (2.3) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}},$$

for some $A > 2\sqrt{2}$. Then for all $M \ge 2$, $n \ge 1$, with probability of at least $1 - M^{1-A^2/8}$ we have simultaneously for all $\lambda \in \mathbb{R}^M$:

$$\begin{aligned} &\|\widetilde{f} - f\|_n^2 + r\sum_{j=1}^M \|f_j\|_n |\widehat{\lambda}_j - \lambda_j| \\ &\quad\le\ \|\mathsf{f}_\lambda - f\|_n^2 + 4r \sum_{j\in J(\lambda)} \|f_j\|_n |\widehat{\lambda}_j - \lambda_j| \\ &\quad\le\ \|\mathsf{f}_\lambda - f\|_n^2 + 4r\sqrt{M(\lambda)}\sqrt{\sum_{j\in J(\lambda)} \|f_j\|_n^2 |\widehat{\lambda}_j - \lambda_j|^2}, \end{aligned} \tag{2.7}$$

and

$$\Big|\frac{1}{n} X^T (\mathbf{f} - X\widehat{\lambda})\Big|_\infty \le 3r f_{\max}/2. \tag{2.8}$$

Furthermore, with the same probability

$$M(\widehat{\lambda}) \le 4\phi_{\max} f_{\min}^{-2} \left(\|\widetilde{f} - f\|_n^2 / r^2 \right) \tag{2.9}$$

where $\phi_{\max}$ *denotes the maximal eigenvalue of the matrix* $X^T X/n$.

LEMMA 2. *Let* $\lambda \in \mathbb{R}^M$ *satisfy the Dantzig constraint*

$$\left| \frac{1}{n} D^{-1/2} X^T (y - X\lambda) \right|_\infty \le r \tag{2.10}$$

and set $\mathbf{\Delta} = \widehat{\lambda}_D - \lambda$, $J_0 = J(\lambda)$. *Then*

$$|\mathbf{\Delta}_{J_0^c}|_1 \le |\mathbf{\Delta}_{J_0}|_1. \tag{2.11}$$

Further, let the assumptions of Lemma 1 be satisfied with $A > \sqrt{2}$. *Then for all* $M \ge 2$, $n \ge 1$ *with probability of at least* $1 - M^{1-A^2/2}$ *we have*

$$\left| \frac{1}{n} X^T (\mathbf{f} - X\widehat{\lambda}_D) \right|_\infty \le 2r f_{\max}. \tag{2.12}$$

3. Restricted eigenvalue assumptions. For any $n \ge 1$, $M \ge 2$, consider the Gram matrix

$$\Psi_n = \frac{1}{n} X^T X = \left(\frac{1}{n} \sum_{i=1}^n f_j(Z_i) f_{j'}(Z_i) \right)_{1 \le j, j' \le M}.$$

We now introduce the key assumptions on the Gram matrix that are needed to guarantee nice statistical properties of Lasso and Dantzig selector. Under the sparsity scenario we are typically interested in the case where $M > n$, and even $M \gg n$. Then the matrix Ψ_n is degenerate, which can be written as

$$\min_{\mathbf{\Delta} \in \mathbb{R}^M : \mathbf{\Delta} \ne 0} \frac{(\mathbf{\Delta}^T \Psi_n \mathbf{\Delta})^{1/2}}{|\mathbf{\Delta}|_2} \equiv \min_{\mathbf{\Delta} \in \mathbb{R}^M : \mathbf{\Delta} \ne 0} \frac{|X\mathbf{\Delta}|_2}{\sqrt{n}|\mathbf{\Delta}|_2} = 0.$$

Clearly, ordinary least squares does not work in this case, since it requires positive definiteness of Ψ_n, i.e.

$$\min_{\mathbf{\Delta} \in \mathbb{R}^M : \mathbf{\Delta} \ne 0} \frac{|X\mathbf{\Delta}|_2}{\sqrt{n}|\mathbf{\Delta}|_2} > 0. \tag{3.1}$$

It turns out that the Lasso and Dantzig selector require much weaker assumptions: the minimum in (3.1) can be replaced by the minimum over a restricted set of vectors, and the norm $|\mathbf{\Delta}|_2$ in the denominator of the condition can be replaced by the ℓ_2 norm of only a part of $\mathbf{\Delta}$. The resulting conditions will be referred to as *restricted eigenvalue* (RE) assumptions.

Our first RE assumption is stated as follows, where s is an integer such that $1 \le s \le M$, and c_0 is a positive number:

Assumption RE(s, c_0):

$$\kappa(s, c_0) \triangleq \min_{J_0 \subseteq \{1,\dots,M\}: |J_0| \le s} \quad \min_{\mathbf{\Delta} \neq 0: |\mathbf{\Delta}_{J_0^c}|_1 \le c_0 |\mathbf{\Delta}_{J_0}|_1} \quad \frac{|X\mathbf{\Delta}|_2}{\sqrt{n}|\mathbf{\Delta}_{J_0}|_2} > 0.$$

The integer s here plays the role of an upper bound on the sparsity $M(\lambda)$ of a vector of coefficients λ. We will usually interpret J_0 as the set of nonzero coefficients of λ. To explain the role of the constant c_0, we may note that the vector of Dantzig residuals $\mathbf{\Delta}$ satisfies $|\mathbf{\Delta}_{J_0^c}|_1 \le c_0 |\mathbf{\Delta}_{J_0}|_1$ with $c_0 = 1$, cf. (2.11). Similar inequality holds for the vector of Lasso residuals $\mathbf{\Delta} = \widehat{\lambda} - \lambda$, but this time with $c_0 = 3$, and with probability $1 - M^{1-A^2/8}$, in the particular case of Lemma 1 where λ is such that $\|\mathsf{f}_\lambda - f\|_n = 0$ and $\|f_j\|_n \equiv 1$ (cf. (2.7)).

To introduce the second assumption we need some more notation. For integers s, m such that $1 \le s \le M/2$ and $m \ge s$, $s + m \le M$, for a vector $\mathbf{\Delta} \in \mathbb{R}^M$ and a set of indices $J_0 \subseteq \{1, \dots, M\}$ with $|J_0| \le s$, denote by J_1 the subset of $\{1, \dots, M\}$ corresponding to m largest in absolute value coordinates of $\mathbf{\Delta}$ outside of J_0 and define $J_{01} \triangleq J_0 \cup J_1$.

Assumption RE(s, m, c_0):

$$\kappa(s, m, c_0) \triangleq \min_{J_0 \subseteq \{1,\dots,M\}: |J_0| \le s} \quad \min_{\mathbf{\Delta} \neq 0: |\mathbf{\Delta}_{J_0^c}|_1 \le c_0 |\mathbf{\Delta}_{J_0}|_1} \quad \frac{|X\mathbf{\Delta}|_2}{\sqrt{n}|\mathbf{\Delta}_{J_{01}}|_2} > 0.$$

Note that Assumption RE(s, c_0) is less restrictive than RE(s, m, c_0). For our bounds on the prediction loss and on the ℓ_1 loss of the Lasso and Dantzig estimators we will only need Assumption RE(s, c_0). The stronger Assumption RE(s, m, c_0) will be required exclusively for the bounds on the ℓ_p loss with $1 < p \le 2$.

Note also that Assumptions RE(s', c_0) and RE(s', m, c_0) imply Assumptions RE(s, c_0) and RE(s, m, c_0) respectively if $s' > s$.

Assumptions RE(s, c_0) and RE(s, m, c_0) are implied by several simple sufficient conditions. We now consider some of them.

For a real number $1 \le u \le M$ we introduce the following "restricted" eigenvalues:

$$\phi_{\min}(u) = \min_{x \in \mathbb{R}^M : 1 \le M(x) \le u} \frac{x^T \Psi_n x}{|x|_2^2}, \quad \phi_{\max}(u) = \max_{x \in \mathbb{R}^M : 1 \le M(x) \le u} \frac{x^T \Psi_n x}{|x|_2^2}.$$

Denote by X_J the $n \times |J|$ submatrix of X obtained by removing from X the columns that do not correspond to the indices in J, and for $1 \le m, m' \le M$ introduce the "restricted" correlations

$$\theta_{m,m'} = \max \left\{ \frac{1}{n} \mathbf{c}_J^T X_J^T X_{J'} \mathbf{c}_{J'} : \ J \cap J' = \emptyset, \ |J| \le m, \ |J'| \le m', \ |\mathbf{c}_J|_2 \le 1, \ |\mathbf{c}_{J'}|_2 \le 1 \right\}$$

where $\mathbf{c}_J \in \mathbb{R}^{|J|}$, $\mathbf{c}_{J'} \in \mathbb{R}^{|J'|}$.

A sufficient condition for RE(s, c_0) and RE(s, m, c_0) with $m = s$ to hold is given, for example, by the following assumption on the Gram matrix.

Assumption 1. *Assume*

$$\phi_{\min}(2s) > c_0 \theta_{s,2s}$$

for some integer $1 \le s \le M/2$ *and a constant* $c_0 > 0$.

This condition with $c_0 = 1$ appeared in [7], in connection with the Dantzig selector. Assumption 1 is more general: we can have here an arbitrary constant $c_0 > 0$ which will allow us to cover not only the Dantzig selector but also the Lasso estimators, and to prove oracle inequalities for the prediction loss when the model is nonparametric.

Our second sufficient condition for RE(s, c_0) and RE(s, m, c_0) does not need bounds on correlations. Only bounds on the minimal and maximal eigenvalues of "small" submatrices of the Gram matrix Ψ_n are involved.

Assumption 2. *Assume*

$$m \phi_{\min}(s + m) > c_0 s \phi_{\max}(m)$$

for some integers s, m *such that* $1 \le s \le M/2$, $m \ge s$, *and* $s + m \le M$, *and a constat* $c_0 > 0$.

Assumption 2 can be viewed as a weakening of the condition on $\phi_{\min}$ in [16]. Indeed, taking $s + m = s \log n$ (we admit w.l.o.g. that $s \log n$ is an integer and $n > 3$) and assuming that $\phi_{\max}(\cdot)$ is uniformly bounded by a constant we get that Assumption 2 is equivalent to

$$\phi_{\min}(s \log n) > c/\log n \tag{3.2}$$

where $c > 0$ is a constant. The corresponding slightly stronger assumption in [16] is stated in asymptotic form (for $s = s_n \to \infty$):

$$\liminf_n \phi_{\min}(s_n \log n) > 0.$$

The following two constants are useful when Assumptions 1 and 2 are considered:

$$\kappa_1(s, c_0) = \sqrt{\phi_{\min}(2s)} \left(1 - \frac{c_0\, \theta_{s,2s}}{\phi_{\min}(2s)}\right)$$

and

$$\kappa_2(s, m, c_0) = \sqrt{\phi_{\min}(s+m)} \left(1 - c_0 \sqrt{\frac{s\, \phi_{\max}(m)}{m\, \phi_{\min}(s+m)}}\right).$$

The next lemma shows that if Assumptions 1 or 2 are satisfied, then the quadratic form $x^T \Psi_n x$ is positive definite on some restricted sets of vectors x. The construction of the lemma is inspired by Candes and Tao [7] and covers, in particular, the corresponding result in [7].

LEMMA 3. *Fix an integer $1 \le s \le M/2$ and a constant $c_0 > 0$.*

(i) Let Assumption 1 be satisfied. Then Assumptions RE(s, c_0) and RE(s, s, c_0) hold with $\kappa(s, c_0) = \kappa(s, s, c_0) = \kappa_1(s, c_0)$. Moreover, for any subset J_0 of $\{1, \dots, M\}$ with cardinality $|J_0| \le s$, and any $\mathbf{\Delta} \in \mathbb{R}^M$ such that

$$|\mathbf{\Delta}_{J_0^c}|_1 \le c_0 |\mathbf{\Delta}_{J_0}|_1 \tag{3.3}$$

we have

$$\frac{1}{\sqrt{n}} |P_{01} X \mathbf{\Delta}|_2 \ge \kappa_1(s, c_0) |\mathbf{\Delta}_{J_{01}}|_2 \tag{3.4}$$

where P_{01} is the projector in $\mathbb{R}^M$ on the linear span of the columns of $X_{J_{01}}$.

(ii) Let Assumption 2 be satisfied. Then Assumptions RE(s,c_0) and RE(s,m,c_0) hold with $\kappa(s, c_0) = \kappa(s, m, c_0) = \kappa_2(s, m, c_0)$. Moreover, for any subset J_0 of $\{1, \dots, M\}$ with cardinality $|J_0| \le s$, and any $\mathbf{\Delta} \in \mathbb{R}^M$ such that (3.3) holds we have

$$\frac{1}{\sqrt{n}} |P_{01} X \mathbf{\Delta}|_2 \ge \kappa_2(s, m, c_0) |\mathbf{\Delta}_{J_{01}}|_2. \tag{3.5}$$

There exist other sufficient conditions for Assumptions RE(s,c_0) and RE(s,m,c_0) to hold. We mention here two of them implying Assumption RE(s,c_0). The first one is the following [1].

Assumption 3. *For an integer s such that $1 \le s \le M$ we have*

$$\phi_{\min}(s) > 2c_0\theta_{s,1}\sqrt{s}$$

where $c_0 > 0$ is a constant.

To argue that Assumption 3 implies RE(s,c_0) it suffices to remark that

$$\begin{aligned} \frac{1}{n}|X\mathbf{\Delta}|_2^2 &\ge \frac{1}{n}\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0}\mathbf{\Delta}_{J_0} - \frac{2}{n}|\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0^c}\mathbf{\Delta}_{J_0^c}| \\ &\ge \phi_{\min}(s)|\mathbf{\Delta}_{J_0}|_2^2 - \frac{2}{n}|\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0^c}\mathbf{\Delta}_{J_0^c}| \end{aligned} \tag{3.6}$$

and, if (3.3) holds,

$$\begin{aligned} |\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0^c}\mathbf{\Delta}_{J_0^c}|/n &\le |\mathbf{\Delta}_{J_0^c}|_1 \max_{j\in J_0^c}|\mathbf{\Delta}_{J_0}^T X_{J_0}^T \mathbf{x}_{(j)}|/n \\ &\le \theta_{s,1}|\mathbf{\Delta}_{J_0^c}|_1|\mathbf{\Delta}_{J_0}|_2 \\ &\le c_0\theta_{s,1}\sqrt{s}|\mathbf{\Delta}_{J_0}|_2^2. \end{aligned}$$

Another type of assumption related to "mutual coherence" [8] is discussed in the connection to Lasso in [4, 5]. We state it here in a slightly different form.

Assumption 4. *For an integer s such that $1 \le s \le M$ we have*

$$\phi_{\min}(s) > 2c_0\theta_{1,1}s \tag{3.7}$$

where $c_0 > 0$ is a constant.

It is easy to see that Assumption 4 implies RE(s,c_0). Indeed, if (3.3) holds,

$$\begin{aligned} \frac{1}{n}|X\mathbf{\Delta}|_2^2 &\ge \frac{1}{n}\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0}\mathbf{\Delta}_{J_0} - 2\theta_{1,1}|\mathbf{\Delta}_{J_0^c}|_1|\mathbf{\Delta}_{J_0}|_1 \\ &\ge \phi_{\min}(s)|\mathbf{\Delta}_{J_0}|_2^2 - 2c_0\theta_{1,1}|\mathbf{\Delta}_{J_0}|_1^2 \\ &\ge (\phi_{\min}(s) - 2c_0\theta_{1,1}s)|\mathbf{\Delta}_{J_0}|_2^2. \end{aligned} \tag{3.8}$$

If all the diagonal elements of matrix X^TX/n are equal to 1 (and thus $\theta_{1,1}$ coincides with the mutual coherence [8]), a simple sufficient condition for Assumption RE(s,c_0) to hold is given by

$$\theta_{1,1} < \frac{1}{(1+2c_0)s}\,. \tag{3.9}$$

In fact, separating the diagonal and off-diagonal terms of the quadratic form we get

$$\mathbf{\Delta}_{J_0}^T X_{J_0}^T X_{J_0} \mathbf{\Delta}_{J_0}/n \geq |\mathbf{\Delta}_{J_0}|_2^2 - \theta_{1,1}|\mathbf{\Delta}_{J_0}|_1^2 \geq |\mathbf{\Delta}_{J_0}|_2^2(1-\theta_{1,1}s).$$

Combining this inequality with (3.8) we see that Assumption RE(s,c_0) is satisfied whenever (3.9) holds.

Unfortunately, Assumption RE(s,c_0) has some weakness. Let, for example, f_j, $j = 1, \ldots, 2^m - 1$, be the Haar wavelet basis on $[0,1]$ ($M = 2^m$) and consider $Z_i = i/n$, $i = 1, \ldots, n$. If $M \gg n$, it is clear that $\phi_{\min}(1) = 0$ since there are functions f_j on the highest resolution level whose supports (of length M^{-1}) contain no points Z_i. So, none of the Assumptions 1 – 4 holds. Intuitively, the problem arises only because we include very high resolution components. Therefore, we may try to restrict the set J_0 in RE(s,c_0) to low resolution components, which is quite reasonable because the "true" or "interesting" vectors of parameters λ are often characterized by such J_0. This idea is formalized in Section 5, cf. Corollary 1, see also a remark after Theorem 6.2 in Section 6.

4. Approximate equivalence.

In this section we prove a type of approximate equivalence between Lasso and Dantzig selector. It is expressed as closeness of the prediction losses $\|\widetilde{f}_D - f\|_n^2$ and $\|\widetilde{f} - f\|_n^2$ when the number of non-zero components of Lasso or Dantzig selector is small as compared to the sample size.

THEOREM 4.1. *Let W_i be independent $\mathcal{N}(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Let Assumption RE(s, 1) be satisfied with $1 \leq s \leq M$. Consider the Dantzig estimator $\widetilde{f}_D$ defined by (2.4) – (2.5) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

and the Lasso estimator $\widetilde{f}$ defined by (2.2) – (2.3) with the same r. Then, for all $n \geq 1$ and $A > \sqrt{2}$ with probability at least $1 - M^{1-A^2/2}$ we have that if $M(\widehat{\lambda}) \leq s$ then

$$\|\widetilde{f}_D - f\|_n^2 \leq \|\widetilde{f} - f\|_n^2 + \frac{16 f_{\max}^2 A^2 \sigma^2}{\kappa^2}\left(\frac{M(\widehat{\lambda})\log M}{n}\right) \tag{4.1}$$

and for $A > 2\sqrt{2}$ *with probability at least* $1 - M^{1-A^2/8}$ *we have that if* $M(\widehat{\lambda}) \le s$,

$$\|\widetilde{f} - f\|_n^2 \le \|\widetilde{f}_D - f\|_n^2 + \frac{9 f_{\max}^2 A^2 \sigma^2}{\kappa^2} \left(\frac{M(\widehat{\lambda}) \log M}{n} \right) \tag{4.2}$$

where $\kappa = \kappa(s, 1)$.

Proof. Set $\mathbf{\Delta} = \widehat{\lambda} - \widehat{\lambda}_D$. We have

$$\frac{1}{n}|\mathbf{f} - X\widehat{\lambda}|_2^2 = \frac{1}{n}|\mathbf{f} - X\widehat{\lambda}_D|_2^2 - \frac{2}{n}\mathbf{\Delta}^T X^T(\mathbf{f} - X\widehat{\lambda}_D) + \frac{1}{n}|X\mathbf{\Delta}|_2^2.$$

This and (2.12) yield

$$\begin{aligned} \|\widetilde{f}_D - f\|_n^2 &\le \|\widetilde{f} - f\|_n^2 + 2|\mathbf{\Delta}|_1 \left| \frac{1}{n} X^T(\mathbf{f} - X\widehat{\lambda}_D) \right|_\infty - \frac{1}{n}|X\mathbf{\Delta}|_2^2 \\ &\le \|\widetilde{f} - f\|_n^2 + 4 f_{\max} r |\mathbf{\Delta}|_1 - \frac{1}{n}|X\mathbf{\Delta}|_2^2 \end{aligned} \tag{4.3}$$

where the last inequality holds with probability at least $1 - M^{1-A^2/2}$. Since the Lasso solution $\widehat{\lambda}$ satisfies the Dantzig constraint, we can apply Lemma 2 with $\lambda = \widehat{\lambda}$, which yields

$$|\mathbf{\Delta}_{J_0^c}|_1 \le |\mathbf{\Delta}_{J_0}|_1 \tag{4.4}$$

with $J_0 = J(\widehat{\lambda})$. By Assumption RE($s$, 1) we get

$$\frac{1}{\sqrt{n}}|X\mathbf{\Delta}|_2 \ge \kappa |\mathbf{\Delta}_{J_0}|_2 \tag{4.5}$$

where $\kappa = \kappa(s, 1)$. Using (4.4) and (4.5) we obtain

$$|\mathbf{\Delta}|_1 \le 2|\mathbf{\Delta}_{J_0}|_1 \le 2M^{1/2}(\widehat{\lambda})\; |\mathbf{\Delta}_{J_0}|_2 \le \frac{2M^{1/2}(\widehat{\lambda})}{\kappa\sqrt{n}}|X\mathbf{\Delta}|_2. \tag{4.6}$$

Finally, from (4.3) and (4.6) we get that, with probability at least $1 - M^{1-A^2/2}$,

$$\begin{aligned} \|\widetilde{f}_D - f\|_n^2 &\le \|\widetilde{f} - f\|_n^2 + \frac{8 f_{\max} r M^{1/2}(\widehat{\lambda})}{\kappa\sqrt{n}}|X\mathbf{\Delta}|_2 - \frac{1}{n}|X\mathbf{\Delta}|_2^2 \\ &\le \|\widetilde{f} - f\|_n^2 + \frac{16 f_{\max}^2 r^2 M(\widehat{\lambda})}{\kappa^2}, \end{aligned}$$

by (2.8) and (2.12). This proves (4.1).

To show (4.2) we act as in (4.3), up to the inversion of roles of $\widehat{\lambda}$ and $\widehat{\lambda}_D$, and we use (2.8). This yields that, with probability at least $1 - M^{1-A^2/8}$,

$$\begin{aligned} \|\widetilde{f} - f\|_n^2 &\le \|\widetilde{f}_D - f\|_n^2 + 2|\mathbf{\Delta}|_1 \left|\frac{1}{n} X^T(\mathbf{f} - X\widehat{\lambda})\right|_\infty - \frac{1}{n}|X\mathbf{\Delta}|_2^2 \\ &\le \|\widetilde{f}_D - f\|_n^2 + 3f_{\max} r|\mathbf{\Delta}|_1 - \frac{1}{n}|X\mathbf{\Delta}|_2^2. \end{aligned} \tag{4.7}$$

The proof of (4.2) now parallels that of (4.1) up to a difference in numerical constants.

We also have the following result that we state for simplicity under the assumption that $\|f_j\|_n = 1,\ j = 1, \dots, M$.

Theorem 4.2. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, and let $\|f_j\|_n = 1,\ j = 1, \dots, M$. Let Assumption RE(s, 5) be satisfied for some $1 \le s \le M$. Consider the Dantzig estimator $\widetilde{f}_D$ defined by (2.4) – (2.5) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. Let $\widetilde{f}$ be the Lasso estimator defined by (2.2) – (2.3) with the same r. Then, for all $n \ge 1$ with probability at least $1 - M^{1-A^2/8}$ we have that if $M(\widehat{\lambda}_D) \le s$ then

$$\|\widetilde{f} - f\|_n^2 \le 10\|\widetilde{f}_D - f\|_n^2 + \frac{81\, A^2\sigma^2}{\kappa^2}\left(\frac{M(\widehat{\lambda}_D)\log M}{n}\right)$$

where $\kappa = \kappa(s, 5)$.

Proof. Set again $\mathbf{\Delta} = \widehat{\lambda} - \widehat{\lambda}_D$. We apply (2.7) with $\lambda = \widehat{\lambda}_D$ which yields that, with probability at least $1 - M^{1-A^2/8}$,

$$|\mathbf{\Delta}|_1 \le 4|\mathbf{\Delta}_{J_0}|_1 + \|\widetilde{f}_D - f\|_n^2/r \tag{4.8}$$

where now $J_0 = J(\widehat{\lambda}_D)$. Consider the two cases: (i) $\|\widetilde{f}_D - f\|_n^2 > 2r|\mathbf{\Delta}_{J_0}|_1$ and (ii) $\|\widetilde{f}_D - f\|_n^2 \le 2r|\mathbf{\Delta}_{J_0}|_1$. In case (i) inequality (4.7) with $f_{\max} = 1$ immediately implies

$$\|\widetilde{f} - f\|_n^2 \le 10\|\widetilde{f}_D - f\|_n^2 \tag{4.9}$$

and the theorem follows. In case (ii) we get from (4.8) that

$$|\mathbf{\Delta}|_1 \le 6|\mathbf{\Delta}_{J_0}|_1 \tag{4.10}$$

and thus $|\mathbf{\Delta}_{J_0^c}|_1 \le 5|\mathbf{\Delta}_{J_0}|_1$. We can therefore apply Assumption RE(s, 5) which yields, similarly to (4.6),

$$|\mathbf{\Delta}|_1 \le 6M^{1/2}(\widehat{\lambda}_D)\ |\mathbf{\Delta}_{J_0}|_2 \le \frac{6M^{1/2}(\widehat{\lambda}_D)}{\kappa\sqrt{n}}|X\mathbf{\Delta}|_2 \tag{4.11}$$

where $\kappa = \kappa(s, 5)$. Plugging (4.11) into (4.7) we finally get that, in case (ii),

$$\begin{aligned} \|\widetilde{f} - f\|_n^2 &\le \|\widetilde{f}_D - f\|_n^2 + \frac{18\,rM^{1/2}(\widehat{\lambda}_D)}{\kappa\sqrt{n}}|X\mathbf{\Delta}|_2 - \frac{1}{n}|X\mathbf{\Delta}|_2^2 \\ &\le \|\widetilde{f}_D - f\|_n^2 + \frac{81\,r^2M(\widehat{\lambda}_D)}{\kappa^2}. \end{aligned} \tag{4.12}$$

REMARK. The approximate equivalence is essentially that of the rates as Theorem 4.1 exhibits. A statement free of $M(\lambda)$ holds for linear regression, see discussion after Theorem 6.2 and Theorem 6.3 below.

5. Oracle inequalities for prediction loss.

Here we prove sparsity oracle inequalities for the prediction loss of Lasso and Dantzig estimators. A general discussion of sparsity oracle inequalities can be found in [21]. Such inequalities have been recently obtained for the Lasso type estimators in a number of settings [2–6, 13, 23]. In particular, the regression model with fixed design that we study here is considered in [2–4]. The assumptions on the Gram matrix Ψ_n in [2–4] are more restrictive than ours: in those papers either Ψ_n is positive definite or a mutual coherence condition similar to (3.9) is imposed.

THEOREM 5.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and an integer $1 \le s \le M$. Let Assumption RE(s, c_0) be satisfied with $c_0 = 3 + 4/\varepsilon$. Consider the Lasso estimator $\widetilde{f}$ defined by (2.2) – (2.3) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

for some $A > 2\sqrt{2}$. *Then, for all* $n \geq 1$ *with probability at least* $1 - M^{1-A^2/8}$ *we have*

$$\|\widetilde{f} - f\|_n^2 \leq (1+\varepsilon) \inf_{\substack{\lambda \in \mathbb{R}^M: \\ M(\lambda) \leq s}} \left\{ \|\mathsf{f}_\lambda - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\kappa^2} \left(\frac{M(\lambda) \log M}{n} \right) \right\} \tag{5.1}$$

where $\kappa = \kappa(s, 3 + 4/\varepsilon)$ *and* $C(\varepsilon) > 0$ *is a constant depending only on* ε.

Proof. Fix an arbitrary $\lambda \in \mathbb{R}^M$ with $M(\lambda) \leq s$. Set $\mathbf{\Delta} = D^{1/2}(\widehat{\lambda} - \lambda)$, $J_0 = J(\lambda)$. On the event $\mathcal{A}$, we get from the first line in (2.7) that

$$\begin{aligned} \|\widetilde{f} - f\|_n^2 + r|\mathbf{\Delta}|_1 &\leq \|\mathsf{f}_\lambda - f\|_n^2 + 4r \sum_{j \in J_0} \|f_j\|_n |\widehat{\lambda}_j - \lambda_j| \\ &= \|\mathsf{f}_\lambda - f\|_n^2 + 4r|\mathbf{\Delta}_{J_0}|_1, \end{aligned} \tag{5.2}$$

and from the second line in (2.7) that

$$\|\widetilde{f} - f\|_n^2 \leq \|\mathsf{f}_\lambda - f\|_n^2 + 4r\sqrt{M(\lambda)}\,|\mathbf{\Delta}_{J_0}|_2. \tag{5.3}$$

Consider separately the cases where

$$4r|\mathbf{\Delta}_{J_0}|_1 \leq \varepsilon \|\mathsf{f}_\lambda - f\|_n^2 \tag{5.4}$$

and

$$\varepsilon \|\mathsf{f}_\lambda - f\|_n^2 < 4r|\mathbf{\Delta}_{J_0}|_1. \tag{5.5}$$

In case (5.4), the result of the theorem trivially follows from (5.2). So, we will only consider the case (5.5). All the subsequent inequalities are valid on the event $\mathcal{A} \cap \mathcal{A}_1$ where $\mathcal{A}_1$ is defined by (5.5). On this event we get from (5.2) that

$$|\mathbf{\Delta}|_1 \leq 4(1 + 1/\varepsilon)|\mathbf{\Delta}_{J_0}|_1 \tag{5.6}$$

which implies $|\mathbf{\Delta}_{J_0^c}|_1 \leq (3 + 4/\varepsilon)|\mathbf{\Delta}_{J_0}|_1$. We now use Assumption RE(s, $3 + 4/\varepsilon$). This yields

$$\begin{aligned} \kappa^2 |\mathbf{\Delta}_{J_0}|_2^2 &\leq \frac{1}{n}|X\mathbf{\Delta}|_2^2 = \frac{1}{n}(\widehat{\lambda} - \lambda)^T D^{1/2} X^T X D^{1/2} (\widehat{\lambda} - \lambda) \\ &\leq \frac{f_{\max}^2}{n}(\widehat{\lambda} - \lambda)^T X^T X (\widehat{\lambda} - \lambda) = f_{\max}^2 \|\widetilde{f} - \mathsf{f}_\lambda\|_n^2 \end{aligned} \tag{5.7}$$

where $\kappa = \kappa(s, 3 + 4/\varepsilon)$. Combining this with (5.3) we find

$$
\begin{aligned}
\|\widetilde{f} - f\|_n^2 &\le \|\mathsf{f}_\lambda - f\|_n^2 + 4r f_{\max}\kappa^{-1}\sqrt{M(\lambda)}\,\|\widetilde{f} - \mathsf{f}_\lambda\|_n \\
&\le \|\mathsf{f}_\lambda - f\|_n^2 + 4r f_{\max}\kappa^{-1}\sqrt{M(\lambda)}\left(\|\widetilde{f} - f\|_n + \|\mathsf{f}_\lambda - f\|_n\right).
\end{aligned} \tag{5.8}
$$

This inequality is of the same form as (A.4) in [4]. A standard decoupling argument as in [4] using inequality $2xy \le x^2/b + by^2$ with $b > 1$, $x = r\kappa^{-1}\sqrt{M(\lambda)}$, and y being either $\|\widetilde{f} - f\|_n$ or $\|\mathsf{f}_\lambda - f\|_n$ yields that

$$
\|\widetilde{f} - f\|_n^2 \le \frac{b+1}{b-1}\|\mathsf{f}_\lambda - f\|_n^2 + \frac{8b^2 f_{\max}^2}{(b-1)\kappa^2} r^2 M(\lambda), \quad \forall\, b > 1. \tag{5.9}
$$

Taking $b = 1 + 2/\varepsilon$ in the last display finishes the proof of the theorem.

We now state as a corollary a softer version of Theorem 5.1 that can be used to eliminate the pathologies mentioned at the end of Section 3. For this purpose we define

$$
\mathcal{J}_{s,\gamma,c_0} = \left\{ J_0 \subset \{1,\ldots,M\} : |J_0| \le s \ \text{ and } \min_{|\mathbf{\Delta}_{J_0^c}| \le c_0 |\mathbf{\Delta}_{J_0}|} \frac{|X\mathbf{\Delta}|_2}{\sqrt{n}|\mathbf{\Delta}_{J_0}|_2} \ge \gamma \right\}
$$

where $\gamma > 0$ is a constant, and set

$$
\Lambda_{s,\gamma,c_0} = \{\lambda : J(\lambda) \in \mathcal{J}_{s,\gamma,c_0}\}.
$$

In similar way, we define $\mathcal{J}_{s,\gamma,m,c_0}$ and Λ_{s,γ,m,c_0} corresponding to Assumption RE(s, m, c_0).

Corollary 1. *Let W_i, s and the Lasso estimator $\widetilde{f}$ be the same as in Theorem 5.1. Then, for all $n \ge 1$ and $\varepsilon > 0$, $\gamma > 0$, with probability at least $1 - M^{1-A^2/8}$ we have*

$$
\begin{aligned}
&\|\widetilde{f} - f\|_n^2 \\
&\quad \le (1+\varepsilon) \inf_{\lambda \in \bar{\Lambda}_{s,\gamma,\varepsilon}} \left\{ \|\mathsf{f}_\lambda - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\gamma^2} \left(\frac{M(\lambda)\log M}{n} \right) \right\}
\end{aligned} \tag{5.10}
$$

where $\bar{\Lambda}_{s,\gamma,\varepsilon} = \{\lambda \in \Lambda_{s,\gamma,3+4/\varepsilon} : M(\lambda) \le s\}$.

To obtain this corollary it suffices to observe that the proof of Theorem 5.1 goes through if we drop Assumption RE(s, c_0) but we assume instead that $\lambda \in \Lambda_{s,\gamma,3+4/\varepsilon}$ and we replace κ by γ.

We would like now to get a sparsity oracle inequality similar to that of Theorem 5.1 for the Dantzig estimator $\widetilde{f}_D$. We will need a mild additional assumption on f. This is due to the fact that not every $\lambda \in \mathbb{R}^M$ obeys to the Dantzig constraint, and thus we cannot assure the key relation (2.11) for all $\lambda \in \mathbb{R}^M$. One possibility would be to prove inequality as (5.1) where the infimum on the right hand side is taken over λ satisfying not only $M(\lambda) \le s$ but also the Dantzig constraint. However, this seems not very intuitive since we cannot guarantee that the corresponding f_λ gives a good approximation of the unknown function f. Therefore we choose another approach (cf. [5]): we consider f satisfying the *weak sparsity* property relative to the dictionary $f_1, \dots, f_M$. That is, we assume that there exist an integer s and constant $C_0 < \infty$ such that the set

$$\Lambda_s = \left\{ \lambda \in \mathbb{R}^M : \ M(\lambda) \le s, \ \|\mathsf{f}_\lambda - f\|_n^2 \le \frac{C_0 f_{\max}^2 r^2}{\kappa^2} M(\lambda) \right\} \tag{5.11}$$

is non-empty. Here κ is the same as in Theorem 5.1. The second inequality in (5.11) says that the "bias" term $\|\mathsf{f}_\lambda - f\|_n^2$ cannot be much larger than the "variance term" $\sim f_{\max}^2 r^2 \kappa^{-2} M(\lambda)$, cf. (5.1). Weak sparsity is milder than the sparsity property in the usual sense: the latter means that f admits the exact representation $f = \mathsf{f}_{\lambda^*}$ for some $\lambda^* \in \mathbb{R}^M$, with hopefully small $M(\lambda^*) = s$.

COROLLARY 2. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$. Let f obey the weak sparsity assumption for some $C_0 < \infty$ and some s such that $1 \le \max(C_1(\varepsilon), 1)s \le M$ where*

$$C_1(\varepsilon) = 4\left[(1+\varepsilon)C_0 + C(\varepsilon)\right] \frac{\phi_{\max} f_{\max}^2}{\kappa^2 f_{\min}^2}$$

and $C(\varepsilon)$ is the constant in Theorem 5.1. Let Assumption RE($\max(C_1(\varepsilon), 1)s$, c_0) be satisfied with $c_0 = 3 + 4/\varepsilon$. Consider the Dantzig estimator $\widetilde{f}_D$ defined by (2.4) – (2.5) with

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. Then, for all $n \ge 1$, with probability at least $1 - M^{1-A^2/8}$ we have

$$\begin{aligned} &\|\widetilde{f}_D - f\|_n^2 \\ &\le (1+\varepsilon) \inf_{\lambda \in \mathbb{R}^M : M(\lambda) = s} \|\mathsf{f}_\lambda - f\|_n^2 + C_2(\varepsilon) \frac{f_{\max}^2 A^2 \sigma^2}{\kappa_0^2} \left(\frac{s \log M}{n}\right). \end{aligned} \tag{5.12}$$

Here $C_2(\varepsilon) = 16C_1(\varepsilon) + C(\varepsilon)$ *and* $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$.

Proof. Due to the weak sparsity assumption there exists $\bar\lambda \in \mathbb{R}^M$ with $M(\bar\lambda) \le s$ such that $\|\mathsf{f}_{\bar\lambda} - f\|_n^2 \le C_0 f_{\max}^2 r^2 \kappa^{-2} M(\bar\lambda)$ where $\kappa = \kappa(s, 3 + 4/\varepsilon)$ is the same as in Theorem 5.1. Using this together with Theorem 5.1 and (2.9) we obtain that, with probability at least $1 - M^{1-A^2/8}$,

$$M(\widehat\lambda) \le C_1(\varepsilon) M(\bar\lambda) \le C_1(\varepsilon)s.$$

This and Theorem 4.1 imply

$$\|\widetilde f_D - f\|_n^2 \le \|\widetilde f - f\|_n^2 + \frac{16C_1(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\kappa_0^2} \left(\frac{s \log M}{n}\right)$$

where $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$. Applying Theorem 5.1 once again we get the result.

Note that the sparsity oracle inequality (5.12) is slightly weaker than the analogous inequality (5.1) for the Lasso: we have here $\inf_{\lambda\in\mathbb{R}^M:\, M(\lambda)=s}$ instead of $\inf_{\lambda\in\mathbb{R}^M:\, M(\lambda)\le s}$ in (5.1).

6. Special case: linear regression.

In this section we assume that the vector of observations $y = (Y_1, \dots, Y_n)^T$ is of the form

$$y = X\boldsymbol\beta^* + W \tag{6.1}$$

where X is an $n\times M$ deterministic matrix, $\boldsymbol\beta^* \in \mathbb{R}^M$ and $W = (W_1, \dots, W_n)^T$.

We do **not** assume that $\boldsymbol\beta^*$ is uniquely defined. On the contrary, we expect to have M at least of order of n and typically much larger. In this case, if $\boldsymbol\beta^* = \boldsymbol\beta_0$ satisfies (6.1) there exists an $(M - n)$-dimensional affine space $\{\boldsymbol\beta^* : X\boldsymbol\beta^* = X\boldsymbol\beta_0\}$ of vectors satisfying (6.1). The results of this section are valid for any $\boldsymbol\beta^*$ such that (6.1) holds, in particular, for $\boldsymbol\beta^{**}$ that gives the sparsest representation of $E(y)$, i.e., such that

$$\boldsymbol\beta^{**} = \arg \min_{\boldsymbol\beta^*:\, X\boldsymbol\beta^* = E(y)} M(\boldsymbol\beta^*).$$

Our goal is to estimate both $X\boldsymbol\beta^*$ for purposes of prediction and $\boldsymbol\beta^*$ itself for purposes of model selection. We will see that meaningful results are obtained when the sparsity index $M(\boldsymbol\beta^*)$ is small.

It will be assumed throughout this section that the diagonal elements of the matrix $X^T X/n$ are all equal to 1 (this is equivalent to the condition $\|f_j\|_n = 1,\ j = 1, \dots, M$, in the notation of previous sections). Then the Lasso estimator of $\boldsymbol{\beta}^*$ in (6.1) is defined by

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^M} \left\{ \frac{1}{n} |y - X\boldsymbol{\beta}|_2^2 + 2r|\boldsymbol{\beta}|_1 \right\}. \tag{6.2}$$

The correspondence between the notation here and that of the previous sections is the following: for $\boldsymbol{\beta} = \lambda$ we have

$$\|\mathsf{f}_\lambda\|_n^2 = |X\,\boldsymbol{\beta}|_2^2/n, \quad \|\mathsf{f}_\lambda - f\|_n^2 = |X\,(\boldsymbol{\beta} - \boldsymbol{\beta}^*)|_2^2/n, \quad \|\widetilde{f} - f\|_n^2 = |X\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2/n.$$

The Dantzig selector for linear model (6.1) is defined by

$$\widehat{\boldsymbol{\beta}}_D = \arg\min_{\boldsymbol{\beta} \in \Lambda} |\boldsymbol{\beta}|_1 \tag{6.3}$$

where

$$\Lambda = \left\{ \boldsymbol{\beta} \in \mathbb{R}^M : \ \left| \frac{1}{n} X^T (y - X\boldsymbol{\beta}) \right|_\infty \le r \right\}$$

is the set of all $\boldsymbol{\beta}$ satisfying the Dantzig constraint.

We first get bounds on the rate of convergence of Dantzig selector.

THEOREM 6.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, let all the diagonal elements of the matrix $X^T X/n$ be equal to 1, and $M(\boldsymbol{\beta}^*) = s$ where $1 \le s \le M$. Let Assumption RE(s,1) be satisfied. Consider the Dantzig selector $\widehat{\boldsymbol{\beta}}_D$ defined by (6.3) with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}$$

and $A > \sqrt{2}$. Then, for all $n \ge 1$, with probability at least $1 - M^{1 - A^2/2}$ we have

$$|\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*|_1 \le \frac{8A}{\kappa^2} \sigma\, s \sqrt{\frac{\log M}{n}}, \tag{6.4}$$

$$|X(\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*)|_2^2 \le \frac{16A^2}{\kappa^2}\, \sigma^2\, s \log M \tag{6.5}$$

where $\kappa = \kappa(s, 1)$. In addition, if Assumption RE(s,m,1) is satisfied, then with the same probability as above, simultaneously for all $1 < p \le 2$ we have

$$|\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*|_p^p \le 2^{p-1} 8 \left\{ 1 + \sqrt{\frac{s}{m}} \right\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2} \sqrt{\frac{\log M}{n}} \right)^p \tag{6.6}$$

where $\kappa = \kappa(s, m, 1)$.

Note that, since $s \le m$, the factor in curly brackets in (6.6) is bounded by a constant independent of s and m. Under Assumption 1 with $c_0 = 1$ (which is less general than RE(s,s,1), cf. Lemma 3(i)) a bound of the form (6.6) for the case $p = 2$ is established by Candes and Tao [7].

Bounds on the rate of convergence of the Lasso selector are quite similar to those obtained in Theorem 6.1. They are given by the following result.

THEOREM 6.2. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Let all the diagonal elements of the matrix $X^T X/n$ be equal to 1, and $M(\boldsymbol{\beta}^*) = s$ where $1 \le s \le M$. Let Assumption RE(s,3) be satisfied. Consider the Lasso selector $\widehat{\boldsymbol{\beta}}$ defined by (6.2) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

and $A > 2\sqrt{2}$. Then, for all $n \ge 1$, with probability at least $1 - M^{1-A^2/8}$ we have

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \le \frac{16A}{\kappa^2}\sigma\, s\sqrt{\frac{\log M}{n}}, \tag{6.7}$$

$$|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2 \le \frac{16A^2}{\kappa^2}\,\sigma^2\, s\log M, \tag{6.8}$$

$$M(\widehat{\boldsymbol{\beta}}) \le \frac{64\phi_{\max}}{\kappa^2}\, s \tag{6.9}$$

where $\kappa = \kappa(s, 3)$. In addition, if Assumption RE(s,m,3) is satisfied, then with the same probability as above, simultaneously for all $1 < p \le 2$ we have

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_p^p \le 16\left\{1 + 3\sqrt{\frac{s}{m}}\right\}^{2(p-1)} s\left(\frac{A\sigma}{\kappa^2}\sqrt{\frac{\log M}{n}}\right)^p \tag{6.10}$$

where $\kappa = \kappa(s, m, 3)$.

Assumptions RE$(s, 1)$ respectively RE$(s, 3)$ can be dropped in Theorem 6.1 and 6.2 if we assume $\beta^* \in \Lambda_{s,\gamma,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate. Then (6.4), (6.5) or respectively (6.7), (6.8) hold with $\kappa = \gamma$. This is analogous to Corollary 1. Similarly (6.6) and (6.10) hold with $\kappa = \gamma$ if $\beta^* \in \Lambda_{s,\gamma,m,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate.

Observe that combining Theorems 6.1 and 6.2 we can immediately get bounds for the differences between Lasso and Dantzig selectors $|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_D|_p^p$

and $|X(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_D)|_2^2$. Such bounds have the same form as those of Theorems 6.1 and 6.2, up to numerical constants. Another way of estimating these differences follows directly from the proof of Theorem 6.1. It suffices to observe that the only property of $\boldsymbol{\beta}^*$ used in that proof is the fact that $\boldsymbol{\beta}^*$ satisfies the Dantzig constraint, which is also true for the Lasso solution $\widehat{\boldsymbol{\beta}}$. So, we can replace $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$ and s by $M(\widehat{\boldsymbol{\beta}})$ everywhere in Theorem 6.1. Generalizing a bit more, we easily derive the following fact.

THEOREM 6.3. *The result of Theorem 6.1 remains valid if we replace there* $|\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*|_p^p$ *by* $\sup\{|\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}|_p^p : \boldsymbol{\beta} \in \Lambda, M(\boldsymbol{\beta}) = s\}$ *for* $1 \le p \le 2$ *and* $|X(\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*)|_2^2$ *by* $\sup\{|X(\widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta})|_2^2 : \boldsymbol{\beta} \in \Lambda, M(\boldsymbol{\beta}) = s\}$ *respectively. Here* Λ *is the set of all vectors satisfying the Dantzig constraint.*

REMARKS.

1. We would like to emphasize that Theorems 6.1 and 6.2 are true for any $\boldsymbol{\beta}^*$ satisfying (6.1), in particular, when the parameter $\boldsymbol{\beta}^*$ is non-identifiable. Even more, Theorem 6.3 applies to certain values of $\boldsymbol{\beta}$ that do not come from the model (6.1) at all. Note that Assumptions RE(s,1) and RE(s,m,1) do not imply identifiability. In fact, they do not guarantee that $\phi_{\min}(2s) > 0$ which is an evident necessary condition for identifiability, cf. [7]. The lack of identifiability is not a contradiction, even when we deal with the ℓ_p loss on the coefficients. Indeed, Theorems 6.1 and 6.2 only give non-asymptotic upper bounds on the loss, with some probability and under some conditions. The probability depends on M and the conditions depend on n and M: recall that Assumptions RE(s,1) and RE(s,m,1) are imposed on the $n \times M$ matrix X. To deduce asymptotic convergence (as $n \to \infty$ and/or as $M \to \infty$) from Theorems 6.1 and 6.2 we would need some very strong additional properties, such as simultaneous validity of Assumption RE(s,1) or RE(s,m,1) (with one and the same constant κ) for infinitely many n and M.

In particular, we see that the identifiability argument emphasized by Candes and Tao [7] to justify a qualified positivity of $\phi_{\min}(2s)$ in their conditions is not really a matter of importance. We get the same and more general results without identifiability. What is more, we can use Theorems 6.1 – 6.3 in a paradoxical way, aiming to deduce some geometric facts from probabilistic statements, for example: "in very high dimensions M and for reasonably large sample sizes n the set of all very sparse vectors $\boldsymbol{\beta}^*$ satisfying the model (6.1) is necessarily very well concentrated".

2. For the smallest value of A (which is $A = 2\sqrt{2}$) the constants in the bound of Theorem 6.2 for the Lasso are larger than the corresponding nu-

merical constants for the Dantzig selector given in Theorem 6.1, again for the smallest admissible value $A = \sqrt{2}$. There is not much margin for improvement, which probably suggests that for the parametric linear model (6.1), under the assumption that all the diagonal elements of the matrix $X^T X/n$ are equal to 1, the Dantzig selector might be better than Lasso. However, this remark should be considered with caution, since Theorems 6.1 and 6.2 only give upper bounds. Note also that Dantzig selector has certain defects as compared to Lasso when the model is nonparametric, as discussed in Section 5. In particular, to obtain sparsity oracle inequalities for Dantzig selector we need some restrictions on f, for example the weak sparsity property. On the other hand, sparsity oracle inequality (5.1) for the Lasso is valid with no restriction on f.

3. Proofs of Theorems 6.1 and 6.2 differ mainly in the value of the tuning constant: $c_0 = 1$ in Theorem 6.1 and $c_0 = 3$ in Theorem 6.2. Note that since the Lasso solution satisfies the Dantzig constraint we could have obtained a result similar to Theorem 6.2, though with less accurate numerical constants, by simply conducting the proof of Theorem 6.1 with $c_0 = 3$. However, we act differently: we deduce (A.17) directly from (2.7), and not from (A.11). This is done only for the sake of improving the constants: in fact, using (A.11) with $c_0 = 3$ would yield (A.17) with the doubled constant on the right hand side.

4. For Dantzig selector in the linear regression model and under Assumptions 1 or 2 some further improvement of constants in the ℓ_p bounds for the coefficients can be achieved by applying the general version of Lemma 3 with the projector P_{01} inside. We do not pursue this issue here.

5. All our results are stated with probabilities at least $1 - M^{1-A^2/2}$ or $1 - M^{1-A^2/8}$. These are reasonable (but not the most accurate) lower bounds on the probabilities $\mathbb{P}(\mathcal{B})$ and $\mathbb{P}(\mathcal{A})$ respectively: we have chosen them just for readability. Inspection of (A.1) shows that they can be refined to $1 - 2M\Phi(A\sqrt{\log M})$ and $1 - 2M\Phi(A\sqrt{\log M}/2)$ respectively where $\Phi(\cdot)$ is the standard normal c.d.f.

APPENDIX A: PROOFS

Proof of Lemma 1. The result (2.7) is essentially Lemma 1 from [5]. For completeness, we give its proof. Set $r_{n,j} = r\|f_j\|_n$. By definition,

$$\widehat{S}(\widehat{\lambda}) + 2\sum_{j=1}^{M} r_{n,j}|\widehat{\lambda}_j| \le \widehat{S}(\lambda) + 2\sum_{j=1}^{M} r_{n,j}|\lambda_j|$$

for all $\lambda \in \mathbb{R}^M$, which is equivalent to

$$\|\widetilde{f} - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}|\widehat{\lambda}_j| \le \|\mathsf{f}_\lambda - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}|\lambda_j| + \frac{2}{n}\sum_{i=1}^n W_i(\widetilde{f} - \mathsf{f}_\lambda)(Z_i).$$

Define the random variables $V_j = n^{-1}\sum_{i=1}^n f_j(Z_i)W_i, \quad 1 \le j \le M$, and the event

$$\mathcal{A} = \bigcap_{j=1}^M \{2|V_j| \le r_{n,j}\}.$$

Using an elementary bound on the tails of Gaussian disribution we find that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\text{(A.1)} \quad \begin{aligned} \mathbb{P}\{\mathcal{A}^c\} &\le \sum_{j=1}^M \mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n} r_{n,j}/2\} \le M\,\mathbb{P}\{|\eta| \ge r\sqrt{n}/(2\sigma)\} \\ &\le M\exp\left(-\frac{nr^2}{8\sigma^2}\right) = M\exp\left(-\frac{A^2\log M}{8}\right) = M^{1-A^2/8} \end{aligned}$$

where $\eta \sim \mathcal{N}(0,1)$. On the event $\mathcal{A}$ we have

$$\|\widetilde{f} - f\|_n^2 \quad\le\quad \|\mathsf{f}_\lambda - f\|_n^2 + \sum_{j=1}^M r_{n,j}|\widehat{\lambda}_j - \lambda_j| + \sum_{j=1}^M 2r_{n,j}|\lambda_j| - \sum_{j=1}^M 2r_{n,j}|\widehat{\lambda}_j|.$$

Adding the term $\sum_{j=1}^M r_{n,j}|\widehat{\lambda}_j - \lambda_j|$ to both sides of this inequality yields, on $\mathcal{A}$,

$$\|\widetilde{f} - f\|_n^2 + \sum_{j=1}^M r_{n,j}|\widehat{\lambda}_j - \lambda_j| \quad\le\quad \|\mathsf{f}_\lambda - f\|_n^2 + 2\sum_{j=1}^M r_{n,j}\left(|\widehat{\lambda}_j - \lambda_j| + |\lambda_j| - |\widehat{\lambda}_j|\right).$$

Now, $|\widehat{\lambda}_j - \lambda_j| + |\lambda_j| - |\widehat{\lambda}_j| = 0$ for $j \notin J(\lambda)$, so that on $\mathcal{A}$ we get (2.7).

To prove (2.8) it suffices to note that on $\mathcal{A}$ we have

$$\text{(A.2)} \qquad \left|\frac{1}{n}D^{-1/2}X^TW\right|_\infty \le r/2.$$

Now, $y = \mathbf{f} + W$, and (2.8) follows from (2.6), (A.2).

We finally prove (2.9). The necessary and sufficient condition for $\widehat{\lambda}$ to be the Lasso solution can be written in the form

$$\text{(A.3)} \quad \begin{aligned} \frac{1}{n}\mathbf{x}_{(j)}^T(y - X\widehat{\lambda}) &= r\|f_j\|_n \operatorname{sign}(\widehat{\lambda}_j) \quad \text{if } \widehat{\lambda}_j \ne 0, \\ \left|\frac{1}{n}\mathbf{x}_{(j)}^T(y - X\widehat{\lambda})\right| &\le r\|f_j\|_n \qquad \text{if } \widehat{\lambda}_j = 0 \end{aligned}$$

where $\mathbf{x}_{(j)}$ denotes the jth column of X, $j = 1, \ldots, M$. Next, (A.2) yields that on $\mathcal{A}$ we have

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^T W\right| \leq r\|f_j\|_n/2, \quad j = 1, \ldots, M. \tag{A.4}$$

Combining (A.3) and (A.4) we get

$$\left|\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{f} - X\widehat{\lambda})\right| \geq r\|f_j\|_n/2 \quad \text{if } \widehat{\lambda}_j \neq 0. \tag{A.5}$$

Therefore,

$$\begin{aligned}\frac{1}{n^2}(\mathbf{f} - X\widehat{\lambda})^T XX^T(\mathbf{f} - X\widehat{\lambda}) &= \frac{1}{n^2}\sum_{j=1}^{M}\left(\mathbf{x}_{(j)}^T(\mathbf{f} - X\widehat{\lambda})\right)^2 \\ &\geq \frac{1}{n^2}\sum_{j:\,\widehat{\lambda}_j\neq 0}\left(\mathbf{x}_{(j)}^T(\mathbf{f} - X\widehat{\lambda})\right)^2 \\ &= M(\widehat{\lambda})r^2\|f_j\|_n^2/4 \geq f_{\min}^2 M(\widehat{\lambda})r^2/4.\end{aligned}$$

Since the matrices X^TX/n and XX^T/n have the same maximal eigenvalues,

$$\frac{1}{n^2}(\mathbf{f} - X\widehat{\lambda})^T XX^T(\mathbf{f} - X\widehat{\lambda}) \leq \frac{\phi_{\max}}{n}|\mathbf{f} - X\widehat{\lambda}|_2^2 = \phi_{\max}\|f - \widetilde{f}\|_n^2$$

and we deduce (2.9) from the last two displays.

Proof of Lemma 2. Inequality(2.11) follows immediately from the definition of Dantzig selector, cf. [7]. To prove (2.12) consider the event

$$\mathcal{B} = \left\{\left|\frac{1}{n}D^{-1/2}X^TW\right|_\infty \leq r\right\} = \bigcap_{j=1}^{M}\{|V_j| \leq r_{n,j}\}.$$

Analogously to (A.1), $\mathbb{P}\{\mathcal{B}^c\} \leq M^{1-A^2/2}$. On the other hand, $y = \mathbf{f} + W$ and using the definition of Dantzig selector it is easy to see that (2.12) is satisfied on $\mathcal{B}$.

Proof of Lemma 3. Consider a partition J_0^c into subsets of size m, with the last subset of size $\leq m$: $J_0^c = \cup_{k=1}^K J_k$ where $K \geq 1$, $|J_k| = m$ for $k = 1, \ldots, K-1$ and $|J_K| \leq m$, such that J_k is the set of indices corresponding to m largest in absolute value coordinates of $\mathbf{\Delta}$ outside $\cup_{j=1}^{k-1} J_j$ (for $k < K$) and J_K is the remaining subset. We have

$$\begin{aligned}|P_{01} X \mathbf{\Delta}|_2 &\geq |P_{01} X \mathbf{\Delta}_{J_{01}}|_2 - \Big|\sum_{k=2}^{K} P_{01} X \mathbf{\Delta}_{J_k}\Big|_2 \\ &= |X \mathbf{\Delta}_{J_{01}}|_2 - \Big|\sum_{k=2}^{K} P_{01} X \mathbf{\Delta}_{J_k}\Big|_2 \\ &\geq |X \mathbf{\Delta}_{J_{01}}|_2 - \sum_{k=2}^{K} |P_{01} X \mathbf{\Delta}_{J_k}|_2 .\end{aligned} \tag{A.6}$$

We will prove first part (ii) of the lemma. Since for $k \geq 1$ the vector $\mathbf{\Delta}_{J_k}$ has only m non-zero components we obtain

$$\frac{1}{\sqrt{n}} |P_{01} X \mathbf{\Delta}_{J_k}|_2 \leq \frac{1}{\sqrt{n}} |X \mathbf{\Delta}_{J_k}|_2 \leq \sqrt{\phi_{\max}(m)}\, |\mathbf{\Delta}_{J_k}|_2 . \tag{A.7}$$

Next, as in [7], we observe that $|\mathbf{\Delta}_{J_{k+1}}|_2 \leq |\mathbf{\Delta}_{J_k}|_1/\sqrt{m}$, $k = 1, \ldots, K-1$, and therefore

$$\sum_{k=2}^{K} |\mathbf{\Delta}_{J_k}|_2 \leq \frac{|\mathbf{\Delta}_{J_0^c}|_1}{\sqrt{m}} \leq \frac{c_0 |\mathbf{\Delta}_{J_0}|_1}{\sqrt{m}} \leq c_0 \sqrt{\frac{s}{m}}\, |\mathbf{\Delta}_{J_0}|_2 \leq c_0 \sqrt{\frac{s}{m}}\, |\mathbf{\Delta}_{J_{01}}|_2 \tag{A.8}$$

where we used (3.3). From (A.6) – (A.8) we find

$$\begin{aligned}\frac{1}{\sqrt{n}} |X \mathbf{\Delta}|_2 &\geq \frac{1}{\sqrt{n}} |X \mathbf{\Delta}_{J_{01}}|_2 - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}}\, |\mathbf{\Delta}_{J_{01}}|_2 \\ &\geq \left(\sqrt{\phi_{\min}(s+m)} - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} \right) |\mathbf{\Delta}_{J_{01}}|_2\end{aligned} \tag{A.9}$$

which proves part (ii) of the lemma.

The proof of part (i) is analogous. The only difference is that we replace in the above argument m by s and instead of (A.7) we use the following bound (cf. [7]):

$$\frac{1}{\sqrt{n}} |P_{01} X \mathbf{\Delta}_{J_k}|_2 \leq \frac{\theta_{s,2s}}{\sqrt{\phi_{\min}(2s)}}\, |\mathbf{\Delta}_{J_k}|_2 . \tag{A.10}$$

Proof of Theorem 6.1. Set $\mathbf{\Delta} = \widehat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}^*$ and $J_0 = J(\boldsymbol{\beta}^*)$. Using Lemma 2 with $\lambda = \boldsymbol{\beta}^*$ we get that on the event $\mathcal{B}$ (i.e., with probability at least

$1 - M^{1-A^2/2}$): (i) $\frac{1}{n}|X^T X\mathbf{\Delta}|_\infty \le 2r$, and (ii) inequality (3.3) holds with $c_0 = 1$. Therefore, on $\mathcal{B}$ we have

$$\text{(A.11)} \qquad \begin{aligned} \frac{1}{n}|X\mathbf{\Delta}|_2^2 &= \frac{1}{n}\mathbf{\Delta}^T X^T X \mathbf{\Delta} \\ &\le \frac{1}{n}\left|X^T X\mathbf{\Delta}\right|_\infty |\mathbf{\Delta}|_1 \\ &\le 2r\Big(|\mathbf{\Delta}_{J_0}|_1 + |\mathbf{\Delta}_{J_0^c}|_1\Big) \\ &\le 2(1+c_0)r|\mathbf{\Delta}_{J_0}|_1 \\ &\le 2(1+c_0)r\sqrt{s}\,|\mathbf{\Delta}_{J_0}|_2 = 4r\sqrt{s}\,|\mathbf{\Delta}_{J_0}|_2 \end{aligned}$$

since $c_0 = 1$. From Assumption RE(s,1) we get that

$$\frac{1}{n}|X\mathbf{\Delta}|_2^2 \ge \kappa^2 |\mathbf{\Delta}_{J_0}|_2^2$$

where $\kappa = \kappa(s, 1)$. This and (A.11) yield that, on $\mathcal{B}$,

$$\text{(A.12)} \qquad \frac{1}{n}|X\mathbf{\Delta}|_2^2 \le 16r^2 s/\kappa^2, \quad |\mathbf{\Delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2.$$

The first inequality in (A.12) implies (6.5). Next, (6.4) is straightforward in view of the second inequality in (A.12) of the following relations (with $c_0 = 1$):

$$\text{(A.13)} \qquad |\mathbf{\Delta}|_1 = |\mathbf{\Delta}_{J_0}|_1 + |\mathbf{\Delta}_{J_0^c}|_1 \le (1+c_0)|\mathbf{\Delta}_{J_0}|_1 \le (1+c_0)\sqrt{s}|\mathbf{\Delta}_{J_0}|_2$$

that hold on $\mathcal{B}$. It remains to prove (6.6). It is easy to see that the kth largest in absolute value element of $\mathbf{\Delta}_{J_0^c}$ satisfies $|\mathbf{\Delta}_{J_0^c}|_{(k)} \le |\mathbf{\Delta}_{J_0^c}|_1/k$. Thus

$$\text{(A.14)} \qquad |\mathbf{\Delta}_{J_{01}^c}|_2^2 \le |\mathbf{\Delta}_{J_0^c}|_1^2 \sum_{k \ge m+1} \frac{1}{k^2} \le \frac{1}{m}|\mathbf{\Delta}_{J_0^c}|_1^2$$

and since (3.3) holds on $\mathcal{B}$ (with $c_0 = 1$) we find

$$|\mathbf{\Delta}_{J_{01}^c}|_2 \le \frac{c_0|\mathbf{\Delta}_{J_0}|_1}{\sqrt{m}} \le c_0|\mathbf{\Delta}_{J_0}|_2\sqrt{\frac{s}{m}} \le c_0|\mathbf{\Delta}_{J_{01}}|_2\sqrt{\frac{s}{m}}\ .$$

Therefore, on $\mathcal{B}$,

$$\text{(A.15)} \qquad |\mathbf{\Delta}|_2 \le \left(1 + c_0\sqrt{\frac{s}{m}}\right)|\mathbf{\Delta}_{J_{01}}|_2.$$

On the other hand, it follows from (A.11) that

$$\frac{1}{n}|X\mathbf{\Delta}|_2^2 \le 4r\sqrt{s}\,|\mathbf{\Delta}_{J_{01}}|_2.$$

Combining this inequality with Assumption RE(s,m,1) we obtain that, on $\mathcal{B}$,

$$|\boldsymbol{\Delta}_{J_{01}}|_2 \le 4r\sqrt{s}/\kappa^2.$$

Recalling that $c_0 = 1$ and applying the last inequality together with (A.15) we get

$$|\boldsymbol{\Delta}|_2^2 \le 16\left(1 + c_0\sqrt{\frac{s}{m}}\right)^2 (r\sqrt{s}/\kappa^2)^2. \tag{A.16}$$

It remains to note that (6.6) is a direct consequence of (6.4) and (A.16). This follows from the fact that inequalities $\sum_{j=1}^M a_j \le b_1$ and $\sum_{j=1}^M a_j^2 \le b_2$ with $a_j \ge 0$ imply

$$\sum_{j=1}^M a_j^p = \sum_{j=1}^M a_j^{2-p} a_j^{2p-2} \le \left(\sum_{j=1}^M a_j\right)^{2-p} \left(\sum_{j=1}^M a_j^2\right)^{p-1} \le b_1^{2-p} b_2^{p-1}, \quad \forall\, 1 < p \le 2.$$

Proof of Theorem 6.2. Set $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $J_0 = J(\boldsymbol{\beta}^*)$. Using (2.7) where we put $\lambda = \boldsymbol{\beta}^*$, $r_{n,j} \equiv r$ and $\|\mathsf{f}_\lambda - f\|_n = 0$ we get that, on the event $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\Delta}|_2^2 \le 4r\sqrt{s}|\boldsymbol{\Delta}_{J_0}|_2 \tag{A.17}$$

and (3.3) holds with $c_0 = 3$ on the same event. Thus, by Assumption RE(s,3) and the last inequality we obtain that, on $\mathcal{A}$,

$$\frac{1}{n}|X\boldsymbol{\Delta}|_2^2 \le 16r^2 s/\kappa^2, \quad |\boldsymbol{\Delta}_{J_0}|_2 \le 4r\sqrt{s}/\kappa^2 \tag{A.18}$$

where $\kappa = \kappa(s, 3)$. The first inequality here coincides with (6.8). Next, (6.9) follows immediately from (2.9) and (6.8). To show (6.7) it suffices to note that on the event $\mathcal{A}$ the relations (A.13) hold with $c_0 = 3$, to apply the second inequality in (A.18) and to use (A.1).

Finally, the proof of (6.10) follows exactly the same lines as that of (6.6): the only difference is that one should set $c_0 = 3$ in (A.15), (A.16), as well as in the display preceding (A.15).

REFERENCES

[1] Bickel, P.J. (2007). Discussion of "The Dantzig selector: statistical estimation when p is much larger than n", by Candes and Tao. *Annals of Statistics*, to appear.

[2] Bunea F., Tsybakov, A.B. and Wegkamp M.H. (2004). Aggregation for regression learning. Preprint LPMA, Universities Paris 6 – Paris 7, n° 948, available at arXiv:math.ST/0410214 and at https://hal.ccsd.cnrs.fr/ccsd-00003205

[3] Bunea F., Tsybakov, A.B. and Wegkamp M.H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence* v.4005 (Lugosi, G. and Simon, H.U.,eds.), Springer-Verlag, Berlin-Heidelberg, 379–391.

[4] Bunea F., Tsybakov, A.B. and Wegkamp M.H. (2007). Aggregation for Gaussian regression. *Annals of Statistics*, to appear.

[5] Bunea F., Tsybakov, A.B. and Wegkamp M.H. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169-194.

[6] Bunea F., Tsybakov, A.B. and Wegkamp M.H. (2007). Sparse density estimation with ℓ_1 penalties. *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007), Lecture Notes in Artificial Intelligence* v.4539 (N.H. Bshouty and C.Gentile, eds.), Springer-Verlag, Berlin-Heidelberg, 530-543.

[7] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, to appear.

[8] Donoho, D.L., Elad, M. and Temlyakov, V. (2006). Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. *IEEE Trans. on Information Theory* **52** 6–18.

[9] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–451.

[10] Fu, W. and Knight, K. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics* **28** 1356-1378.

[11] Greenshtein, E. and Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.

[12] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Annals of Statistics* **28** 681–712.

[13] Koltchinskii, V. (2006). Sparsity in penalized empirical risk minimization. *Manuscript.*

[14] Koltchinskii, V. (2007). Dantzig selector and sparsity oracle inequalities. *Manuscript.*

[15] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.

[16] Meinshausen, N. and Yu, B. (2006). Lasso type recovery of sparse representations for high dimensional data. *Manuscript*.

[17] Nemirovski, A. (2000). *Topics in Non-parametric Statistics*. Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998, Lecture Notes in Mathematics, v. 1738, Springer: New York.

[18] Osborne, M.R., Presnell, B. and Turlach, B.A (2000a). On the Lasso and its dual. *Journal of Computational and Graphical Statistics* **9** 319 – 337.

[19] Osborne, M.R., Presnell, B. and Turlach, B.A (2000b). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20** 389 – 404.

[20] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **58** 267–288.

[21] Tsybakov, A.B. (2006). Discussion of "Regularization in Statistics", by P.Bickel and B.Li. *TEST* **15** 303-310.

[22] Turlach, B.A. (2005). On algorithms for solving least squares problems under an L1 penalty or an L1 constraint. *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*, American Statistical Association, Alexandria, VA, pp. 2572-2577.

[23] van de Geer, S.A. (2006). High dimensional generalized linear models and the Lasso. Research report No.133. Seminar für Statistik, ETH, Zürich.

[24] Zhang, C.-H. and Huang, J. (2006). Model-selection consistency of the Lasso in high-dimensional regression. *Manuscript.*

[25] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Department of Statistics
University of California at Berkeley,
CA USA
E-mail: bickel@stat.berkeley.edu

Jerusalem, Israel
E-mail: yaacov.ritov@gmail.com

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris VI, France.
E-mail: tsybakov@ccr.jussieu.fr