

SIMULTANEOUS ANALYSIS OF LASSO AND DANTZIG SELECTOR¹

BY PETER J. BICKEL, YA'ACOV RITOV AND ALEXANDRE B. TSYBAKOV

*University of California at Berkeley, The Hebrew University and
Université Paris VI and CREST*

We show that, under a sparsity scenario, the Lasso estimator and the Dantzig selector exhibit similar behavior. For both methods, we derive, in 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, 9, 11, 17, 18, 20–22, 26] and [27]. Quite recently, Candès 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; that is, when the number of nonzero components of the true vector of parameters is small.

Lasso estimators have also been studied in the nonparametric regression setup [2–4, 12, 13, 19] and [5]. In particular, Bunea, Tsybakov and Wegkamp [2–5] obtain sparsity oracle inequalities for the prediction loss in this context and point out the implications for minimax estimation in classical nonparametric 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 (nonquadratic terms and/or penalties slightly different from ℓ_1) for nonparametric regression with random design are suggested and studied under prediction loss in [14] and [25]. Sparsity oracle inequalities for the Dantzig selector with random design are obtained in [15]. In linear fixed design regression, Meinshausen and Yu [18] establish a bound on the ℓ_2 loss for the coefficients of Lasso that is quite different from the bound on the same loss for the Dantzig selector proven in [7].

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

Received August 2007; revised April 2008.

¹Supported in part by NSF Grant DMS-06-05236, ISF grant, France-Berkeley Fund, the Grant ANR-06-BLAN-0194 and the European Network of Excellence PASCAL.

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

Key words and phrases. Linear models, model selection, nonparametric statistics.

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 nonasymptotic.

Let us specialize to the case of linear regression with many covariates, $\mathbf{y} = X\beta + \mathbf{w}$, where X is the $n \times M$ deterministic design matrix, with M possibly much larger than n , and \mathbf{w} is a vector of i.i.d. standard normal random variables. This is the situation considered most recently by Candès and Tao [7] and Meinshausen and Yu [18]. Here, sparsity specifies that the high-dimensional vector β has coefficients that are mostly 0.

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, Tsybakov and Wegkamp [4], and ℓ_2 bounds for the coefficients of Dantzig selector, as in Candès and Tao [7]. This is obtained as a consequence of our more general results, which are the following:

- In the nonparametric regression model, we prove sparsity oracle inequalities for the Dantzig selector; that is, bounds on the prediction loss in terms of the best possible (oracle) approximation under the sparsity constraint.
- Similar sparsity oracle inequalities are proved for the Lasso in the nonparametric regression model, and this is done under more general assumptions on the design matrix than in [4].
- We prove that, for nonparametric regression, the Lasso and the Dantzig selector are approximately equivalent in terms of the prediction loss.
- We develop geometrical assumptions that are considerably weaker than those of Candès and Tao [7] for the Dantzig selector and Bunea, Tsybakov and Wegkamp [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. In Section 3, we present our key geometric assumptions. Some sufficient conditions for these assumptions are given in Section 4, where they are also compared to those of [7] and [18], as well as to ones appearing in [4] and [5]. We note a weakness of our assumptions, and, hence, of those in the papers we cited, and we discuss a way of slightly remedying them. Sections 5 and 6 give some equivalence results and sparsity oracle inequalities for the Lasso and Dantzig estimators in the general nonparametric regression model. Section 7 focuses on the linear regression model and includes a final discussion. Two important technical lemmas are given in Appendix B as well as most of the proofs.

2. Definitions and notation. Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be a sample of independent random pairs with

$$Y_i = f(Z_i) + W_i, \quad i = 1, \dots, n,$$

where $f : \mathcal{Z} \rightarrow \mathbb{R}$ 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 $\mathcal{F}_M = \{f_1, \dots, f_M\}$ be a finite dictionary of functions $f_j : \mathcal{Z} \rightarrow \mathbb{R}$, $j = 1, \dots, M$. We assume throughout that $M \geq 2$.

Depending on the statistical targets, the dictionary \mathcal{F}_M can contain qualitatively different parts. For instance, it can be a collection of basis functions used to approximate f in the nonparametric regression model (e.g., wavelets, splines with fixed knots, step functions). 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 that the f_j are based on.

The selection of the dictionary can be very important to make the estimation of f possible. We assume implicitly that f can be well approximated by a member of the span of \mathcal{F}_M . However, this is not enough. In this paper, we have in mind the situation where $M \gg n$, and f can be estimated reasonably only because it can be approximated by a linear combination of a small number of members of \mathcal{F}_M , or, in other words, it has a sparse approximation in the span of \mathcal{F}_M . But, when sparsity is an issue, equivalent bases can have different properties. A function that has a sparse representation in one basis may not have it in another, even if both of them span the same linear space.

Consider the matrix $X = (f_j(Z_i))_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, M$ and the vectors $\mathbf{y} = (Y_1, \dots, Y_n)^T$, $\mathbf{f} = (f(Z_1), \dots, f(Z_n))^T$, $\mathbf{w} = (W_1, \dots, W_n)^T$. With the notation

$$\mathbf{y} = \mathbf{f} + \mathbf{w},$$

we will write $|x|_p$ for the ℓ_p norm of $x \in \mathbb{R}^M$, $1 \leq p \leq \infty$. The notation $\|\cdot\|_n$ stands for the empirical norm

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

for any $g : \mathcal{Z} \rightarrow \mathbb{R}$. We suppose that $\|f_j\|_n \neq 0$, $j = 1, \dots, M$. Set

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

For any $\beta = (\beta_1, \dots, \beta_M) \in \mathbb{R}^M$, define $f_\beta = \sum_{j=1}^M \beta_j f_j$ or, explicitly, $f_\beta(z) = \sum_{j=1}^M \beta_j f_j(z)$ and $\mathbf{f}_\beta = X\beta$. The estimates we consider are all of the form $f_{\tilde{\beta}}(\cdot)$, where $\tilde{\beta}$ is data determined. Since we consider mainly sparse vectors $\tilde{\beta}$, it will be convenient to define the following. Let

$$\mathcal{M}(\beta) = \sum_{j=1}^M I_{\{\beta_j \neq 0\}} = |J(\beta)|$$

denote the number of nonzero coordinates of β , where $I_{\{\cdot\}}$ denotes the indicator function $J(\beta) = \{j \in \{1, \dots, M\} : \beta_j \neq 0\}$ and $|J|$ denotes the cardinality of J . The value $\mathcal{M}(\beta)$ characterizes the *sparsity* of the vector β . The smaller $\mathcal{M}(\beta)$, the “sparser” β . For a vector $\delta \in \mathbb{R}^M$ and a subset $J \subset \{1, \dots, M\}$, we denote by δ_J the vector in \mathbb{R}^M that has the same coordinates as δ on J and zero coordinates on the complement J^c of J .

Introduce the residual sum of squares

$$\widehat{S}(\beta) = \frac{1}{n} \sum_{i=1}^n \{Y_i - f_\beta(Z_i)\}^2$$

for all $\beta \in \mathbb{R}^M$. Define the Lasso solution $\widehat{\beta}_L = (\widehat{\beta}_{1,L}, \dots, \widehat{\beta}_{M,L})$ by

$$(2.1) \quad \widehat{\beta}_L = \arg \min_{\beta \in \mathbb{R}^M} \left\{ \widehat{S}(\beta) + 2r \sum_{j=1}^M \|f_j\|_n |\beta_j| \right\},$$

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

$$(2.2) \quad \widehat{f}_L(x) = f_{\widehat{\beta}_L}(x) = \sum_{j=1}^M \widehat{\beta}_{j,L} f_j(z).$$

The criterion in (2.1) is convex in β , so that standard convex optimization procedures can be used to compute $\widehat{\beta}_L$. We refer to [9, 10, 20, 21, 24] and [16] for detailed discussion of these optimization problems and fast algorithms.

A necessary and sufficient condition of the minimizer in (2.1) is that 0 belongs to the subdifferential of the convex function $\beta \mapsto n^{-1} |y - X\beta|_2^2 + 2r |D^{1/2} \beta|_1$. This implies that the Lasso selector $\widehat{\beta}_L$ satisfies the constraint

$$(2.3) \quad \left| \frac{1}{n} D^{-1/2} X^T (y - X\widehat{\beta}_L) \right|_\infty \leq r,$$

where D is the diagonal matrix

$$D = \text{diag}\{\|f_1\|_n^2, \dots, \|f_M\|_n^2\}.$$

More generally, we will say that $\beta \in \mathbb{R}^M$ satisfies the Dantzig constraint if β belongs to the set

$$\left\{ \beta \in \mathbb{R}^M : \left| \frac{1}{n} D^{-1/2} X^T (y - X\beta) \right|_\infty \leq r \right\}.$$

The Dantzig estimator of the regression function f is based on a particular solution of (2.3), the Dantzig selector $\widehat{\beta}_D$, which is defined as a vector having the smallest ℓ_1 norm among all β satisfying the Dantzig constraint

$$(2.4) \quad \widehat{\beta}_D = \arg \min \left\{ |\beta|_1 : \left| \frac{1}{n} D^{-1/2} X^T (y - X\beta) \right|_\infty \leq r \right\}.$$

The Dantzig estimator is defined by

$$(2.5) \quad \widehat{f}_D(z) = f_{\widehat{\beta}_D}(z) = \sum_{j=1}^M \widehat{\beta}_{j,D} f_j(z),$$

where $\widehat{\beta}_D = (\widehat{\beta}_{1,D}, \dots, \widehat{\beta}_{M,D})$ is the Dantzig selector. By the definition of Dantzig selector, we have $|\widehat{\beta}_D|_1 \leq |\widehat{\beta}_L|_1$.

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

Finally, for any $n \geq 1, M \geq 2$, we consider the Gram matrix

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

and let ϕ_{\max} denote the maximal eigenvalue of Ψ_n .

3. Restricted eigenvalue assumptions. We now introduce the key assumptions on the Gram matrix that are needed to guarantee nice statistical properties of the Lasso and Dantzig selectors. 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

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

Clearly, ordinary least squares does not work in this case, since it requires positive definiteness of Ψ_n ; that is,

$$(3.1) \quad \min_{\delta \in \mathbb{R}^M: \delta \neq 0} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} > 0.$$

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 $|\delta|_2$ in the denominator of the condition can be replaced by the ℓ_2 norm of only a part of δ .

One of the properties of both the Lasso and the Dantzig selectors is that, for the linear regression model, the residuals $\delta = \widehat{\beta}_L - \beta$ and $\delta = \widehat{\beta}_D - \beta$ satisfy, with probability close to 1,

$$(3.2) \quad |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1,$$

where $J_0 = J(\beta)$ is the set of nonzero coefficients of the true parameter β of the model. For the linear regression model, the vector of Dantzig residuals δ satisfies (3.2) with probability close 1 if $c_0 = 1$ and M is large [cf. (B.9) and the fact that β of the model satisfies the Dantzig constraint with probability close to 1 if M is

large]. A similar inequality holds for the vector of Lasso residuals $\delta = \widehat{\beta}_L - \beta$, but this time with $c_0 = 3$ [cf. Corollary B.2].

Now, for example, consider the case where the elements of the Gram matrix Ψ_n are close to those of a positive definite $(M \times M)$ -matrix Ψ . Denote, by $\varepsilon_n \triangleq \max_{i,j} |(\Psi_n - \Psi)_{i,j}|$, the maximal difference between the elements of the two matrices. Then, for any δ satisfying (3.2), we get

$$\begin{aligned}
 \frac{\delta^T \Psi_n \delta}{|\delta|_2^2} &= \frac{\delta^T \Psi \delta + \delta^T (\Psi_n - \Psi) \delta}{|\delta|_2^2} \\
 &\geq \frac{\delta^T \Psi \delta}{|\delta|_2^2} - \frac{\varepsilon_n |\delta|_1^2}{|\delta|_2^2} \\
 (3.3) \quad &\geq \frac{\delta^T \Psi \delta}{|\delta|_2^2} - \varepsilon_n \left(\frac{(1 + c_0) |\delta_{J_0}|_1}{|\delta_{J_0}|_2} \right)^2 \\
 &\geq \frac{\delta^T \Psi \delta}{|\delta|_2^2} - \varepsilon_n (1 + c_0)^2 |J_0|.
 \end{aligned}$$

Thus, for δ satisfying (3.2), which are the vectors that we have in mind, and for $\varepsilon_n |J_0|$ small enough, the LHS of (3.3) is bounded away from 0. This means that we have a kind of “restricted” positive definiteness, which is valid only for the vectors satisfying (3.2). This suggests the following conditions, which will suffice for the main argument of the paper. We refer to these conditions as *restricted eigenvalue* (RE) assumptions.

ASSUMPTION RE(s, c_0). For some integer s such that $1 \leq s \leq M$ and a positive number c_0 , the following condition holds:

$$\kappa(s, c_0) \triangleq \min_{\substack{J_0 \subseteq \{1, \dots, M\}, \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0, \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|X \delta|_2}{\sqrt{n} |\delta_{J_0}|_2} > 0.$$

The integer s here plays the role of an upper bound on the sparsity $\mathcal{M}(\beta)$ of a vector of coefficients β .

Note that, if Assumption RE(s, c_0) is satisfied with $c_0 \geq 1$, then

$$\min\{|X \delta|_2 : \mathcal{M}(\delta) \leq 2s, \delta \neq 0\} > 0.$$

In other words, the square submatrices of size $\leq 2s$ of the Gram matrix are necessarily positive definite. Indeed, suppose that, for some $\delta \neq 0$, we have simultaneously $\mathcal{M}(\delta) \leq 2s$ and $X \delta = 0$. Partition $J(\delta)$ in two sets $J(\delta) = I_0 \cup I_1$, such that $|I_i| \leq s, i = 0, 1$. Without loss of generality, suppose that $|\delta_{I_1}|_1 \leq |\delta_{I_0}|_1$. Since, clearly, $|\delta_{I_1}|_1 = |\delta_{I_1^c}|_1$ and $c_0 \geq 1$, we have $|\delta_{I_0^c}|_1 \leq c_0 |\delta_{I_0}|_1$. Hence, $\kappa(s, c_0) = 0$, a contradiction.

To introduce the second assumption, we need some notation. For integers s, m such that $1 \leq s \leq M/2$ and $m \geq s, s + m \leq M$, a vector $\delta \in \mathbb{R}^M$ and a set of indices $J_0 \subseteq \{1, \dots, M\}$ with $|J_0| \leq s$; denote by J_1 the subset of $\{1, \dots, M\}$ corresponding to the m largest in absolute value coordinates of δ outside of J_0 , and define $J_{01} \triangleq J_0 \cup J_1$. Clearly, J_1 and J_{01} depend on m , but we do not indicate this in our notation for the sake of brevity.

ASSUMPTION $\text{RE}(s, m, c_0)$.

$$\kappa(s, m, c_0) \triangleq \min_{\substack{J_0 \subseteq \{1, \dots, M\}, \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0, \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|X\delta|_2}{\sqrt{n} |\delta_{J_{01}}|_2} > 0.$$

Note that the only difference between the two assumptions is in the denominators, and $\kappa(s, m, c_0) \leq \kappa(s, c_0)$. As written, for fixed n , the two assumptions are equivalent. However, asymptotically for large n , Assumption $\text{RE}(s, c_0)$ is less restrictive than $\text{RE}(s, m, c_0)$, since the ratio $\kappa(s, m, c_0)/\kappa(s, c_0)$ may tend to 0 if s and m depend on n . For our bounds on the prediction loss and on the ℓ_1 loss of the Lasso and Dantzig estimators, we will only need Assumption $\text{RE}(s, c_0)$. Assumption $\text{RE}(s, m, c_0)$ will be required exclusively for the bounds on the ℓ_p loss with $1 < p \leq 2$.

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

4. Discussion of the RE assumptions. There exist several simple sufficient conditions for Assumptions $\text{RE}(s, c_0)$ and $\text{RE}(s, m, c_0)$ to hold. Here, we discuss some of them.

For a real number $1 \leq u \leq M$, we introduce the following quantities that we will call *restricted eigenvalues*:

$$\begin{aligned} \phi_{\min}(u) &= \min_{x \in \mathbb{R}^M: 1 \leq \mathcal{M}(x) \leq u} \frac{x^T \Psi_n x}{|x|_2^2}, \\ \phi_{\max}(u) &= \max_{x \in \mathbb{R}^M: 1 \leq \mathcal{M}(x) \leq u} \frac{x^T \Psi_n x}{|x|_2^2}. \end{aligned}$$

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 $1 \leq m_1, m_2 \leq M$, introduce the following quantities called *restricted correlations*:

$$\theta_{m_1, m_2} = \max \left\{ \frac{c_1^T X_{I_1}^T X_{I_2} c_2}{n |c_1|_2 |c_2|_2} : I_1 \cap I_2 = \emptyset, |I_i| \leq m_i, c_i \in \mathbb{R}^{I_i} \setminus \{0\}, i = 1, 2 \right\}.$$

In Lemma 4.1, below, we show that a sufficient condition for $\text{RE}(s, c_0)$ and $\text{RE}(s, s, c_0)$ to hold is given, for example, by the following assumption on the Gram matrix.

ASSUMPTION 1. Assume that

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

for some integer $1 \leq s \leq M/2$ 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, in that we can have an arbitrary constant $c_0 > 0$ that 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 $\text{RE}(s, c_0)$ and $\text{RE}(s, m, c_0)$ 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 that

$$m\phi_{\min}(s + m) > c_0^2s\phi_{\max}(m)$$

for some integers s, m , such that $1 \leq s \leq M/2$, $m \geq s$ and $s + m \leq M$, and a constant $c_0 > 0$.

Assumption 2 can be viewed as a weakening of the condition on ϕ_{\min} in [18]. Indeed, taking $s + m = s \log n$ (we assume, without loss of generality, 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

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

where $c > 0$ is a constant. The corresponding, slightly stronger, assumption in [18] is stated in asymptotic form, for $s = s_n \rightarrow \infty$, as

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

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 Candès and Tao [7] and covers, in particular, the corresponding result in [7].

LEMMA 4.1. Fix an integer $1 \leq s \leq M/2$ 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 $|J_0| \leq s$, and any $\delta \in \mathbb{R}^M$ such that

$$(4.1) \quad |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1,$$

we have

$$\frac{1}{\sqrt{n}} |P_{01} X \delta|_2 \geq \kappa_1(s, c_0) |\delta_{J_0}|_2,$$

where P_{01} is the projector in \mathbb{R}^M on the linear span of the columns of X_{J_0} .

(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 $|J_0| \leq s$, and any $\delta \in \mathbb{R}^M$ such that (4.1) holds, we have

$$\frac{1}{\sqrt{n}} |P_{01} X \delta|_2 \geq \kappa_2(s, m, c_0) |\delta_{J_0}|_2.$$

The proof of the lemma is given in Appendix A.

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

ASSUMPTION 3. For an integer s such that $1 \leq s \leq M$, 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

$$\begin{aligned} \frac{1}{n} |X \delta|_2^2 &\geq \frac{1}{n} \delta_{J_0}^T X^T X \delta_{J_0} - \frac{2}{n} |\delta_{J_0}^T X^T X \delta_{J_0^c}| \\ &\geq \phi_{\min}(s) |\delta_{J_0}|_2^2 - \frac{2}{n} |\delta_{J_0}^T X^T X \delta_{J_0^c}| \end{aligned}$$

and, if (4.1) holds,

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

Another type of assumption related to “mutual coherence” [8] is discussed in connection to Lasso in [4, 5]. We state it in two different forms, which are given below.

ASSUMPTION 4. For an integer s such that $1 \leq s \leq M$, we have

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

where $c_0 > 0$ is a constant.

It is easy to see that Assumption 4 implies $\text{RE}(s, c_0)$. Indeed, if (4.1) holds,

$$\begin{aligned} \frac{1}{n}|X\delta|_2^2 &\geq \frac{1}{n}\delta_{J_0}^T X^T X \delta_{J_0} - 2\theta_{1,1}|\delta_{J_0^c}|_1 |\delta_{J_0}|_1 \\ (4.2) \qquad &\geq \phi_{\min}(s)|\delta_{J_0}|_2^2 - 2c_0\theta_{1,1}|\delta_{J_0}|_1^2 \\ &\geq (\phi_{\min}(s) - 2c_0\theta_{1,1}s)|\delta_{J_0}|_2^2. \end{aligned}$$

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

ASSUMPTION 5. All the diagonal elements of the Gram matrix Ψ_n are equal to 1, and for an integer s , such that $1 \leq s \leq M$, we have

$$(4.3) \qquad \theta_{1,1} < \frac{1}{(1 + 2c_0)s},$$

where $c_0 > 0$ is a constant.

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

$$\delta_{J_0}^T X^T X \delta_{J_0}/n \geq |\delta_{J_0}|_2^2 - \theta_{1,1}|\delta_{J_0}|_1^2 \geq |\delta_{J_0}|_2^2(1 - \theta_{1,1}s).$$

Combining this inequality with (4.2), we see that Assumption $\text{RE}(s, c_0)$ is satisfied whenever (4.3) holds.

Unfortunately, Assumption $\text{RE}(s, c_0)$ has some weakness. Let, for example, f_j , $j = 1, \dots, 2^m - 1$, be the Haar wavelet basis on $[0, 1]$ ($M = 2^m$), and consider $Z_i = i/n$, $i = 1, \dots, n$. If $M \gg n$, then 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 Assumptions 1–4 hold. A less severe, although similar, situation is when we consider step functions $f_j(t) = I_{\{t < j/M\}}$ for $t \in [0, 1]$. It is clear that $\phi_{\min}(2) = O(1/M)$, although sparse representation in this basis is very natural. Intuitively, the problem arises only because we include very high resolution components. Therefore, we may try to restrict the set J_0 in $\text{RE}(s, c_0)$ to low resolution components, which is quite reasonable, because the “true” or “interesting” vectors of parameters β are often characterized by such J_0 . This idea is formalized in Section 6 (cf. Corollary 6.2, see also a remark after Theorem 7.2 in Section 7).

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

THEOREM 5.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \geq 1$, $M \geq 2$. Let Assumption RE($s, 1$) be satisfied with $1 \leq s \leq M$. Consider the Dantzig estimator \hat{f}_D defined by (2.5)–(2.4) with*

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

where $A > 2\sqrt{2}$, and consider the Lasso estimator \hat{f}_L defined by (2.1)–(2.2) with the same r .

If $\mathcal{M}(\hat{\beta}_L) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have

$$(5.1) \quad \left| \|\hat{f}_D - f\|_n^2 - \|\hat{f}_L - f\|_n^2 \right| \leq 16A^2 \frac{\mathcal{M}(\hat{\beta}_L)\sigma^2}{n} \frac{f_{\max}^2}{\kappa^2(s, 1)} \log M.$$

Note that the RHS of (5.1) is bounded by a product of three factors (and a numerical constant which, unfortunately, equals at least 128). The first factor $\mathcal{M}(\hat{\beta}_L)\sigma^2/n \leq s\sigma^2/n$ corresponds to the error rate for prediction in regression with s parameters. The two other factors, $\log M$ and $f_{\max}^2/\kappa^2(s, 1)$, can be regarded as a price to pay for the large number of regressors. If the Gram matrix Ψ_n equals the identity matrix (the white noise model), then there is only the $\log M$ factor. In the general case, there is another factor $f_{\max}^2/\kappa^2(s, 1)$ representing the extent to which the Gram matrix is ill-posed for estimation of sparse vectors.

We also have the following result that we state, for simplicity, under the assumption that $\|f_j\|_n = 1$, $j = 1, \dots, M$. It gives a bound in the spirit of Theorem 5.1 but with $\mathcal{M}(\hat{\beta}_D)$ rather than $\mathcal{M}(\hat{\beta}_L)$ on the right-hand side.

THEOREM 5.2. *Let the assumptions of Theorem 5.1 hold, but with RE($s, 5$) in place of RE($s, 1$), and let $\|f_j\|_n = 1$, $j = 1, \dots, M$. If $\mathcal{M}(\hat{\beta}_D) \leq s$, then, with probability at least $1 - M^{1-A^2/8}$, we have*

$$(5.2) \quad \|\hat{f}_L - f\|_n^2 \leq 10\|\hat{f}_D - f\|_n^2 + 81A^2 \frac{\mathcal{M}(\hat{\beta}_D)\sigma^2}{n} \frac{\log M}{\kappa^2(s, 5)}.$$

REMARK. The approximate equivalence is essentially that of the rates as Theorem 5.1 exhibits. A statement free of $\mathcal{M}(\beta)$ holds for linear regression, see discussion after Theorems 7.2 and 7.3 below.

6. Oracle inequalities for prediction loss. Here, we prove sparsity oracle inequalities for the prediction loss of the Lasso and Dantzig estimators. These inequalities allow us to bound the difference between the prediction errors of the estimators and the best sparse approximation of the regression function (by an oracle that knows the truth but is constrained by sparsity). The results of this section, together with those of Section 5, show that the distance between the prediction losses of the Dantzig and Lasso estimators is of the same order as the distances between them and their oracle approximations.

A general discussion of sparsity oracle inequalities can be found in [23]. Such inequalities have been recently obtained for the Lasso type estimators in a number of settings [2–6, 14] and [25]. 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 (4.3) is imposed.

THEOREM 6.1. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1$, $M \geq 2$, $1 \leq s \leq M$. Let Assumption RE($s, 3 + 4/\varepsilon$) be satisfied. Consider the Lasso estimator \hat{f}_L defined by (2.1)–(2.2) with*

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

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

$$(6.1) \quad \begin{aligned} & \|\hat{f}_L - f\|_n^2 \\ & \leq (1 + \varepsilon) \inf_{\substack{\beta \in \mathbb{R}^M: \\ \mathcal{M}(\beta) \leq s}} \left\{ \|\beta - f\|_n^2 + \frac{C(\varepsilon) f_{\max}^2 A^2 \sigma^2 \mathcal{M}(\beta) \log M}{\kappa^2(s, 3 + 4/\varepsilon) n} \right\}, \end{aligned}$$

where $C(\varepsilon) > 0$ is a constant depending only on ε .

We now state, as a corollary, a softer version of Theorem 6.1 that can be used to eliminate the pathologies mentioned at the end of Section 4. For this purpose, we define

$$\mathcal{J}_{s,\gamma,c_0} = \left\{ J_0 \subset \{1, \dots, M\} : |J_0| \leq s \text{ and } \min_{\substack{\delta \neq 0, \\ |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|X\delta|_2}{\sqrt{n}|\delta_{J_0}|_2} \geq \gamma \right\},$$

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

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

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 6.2. *Let W_i, s and the Lasso estimator \widehat{f}_L be the same as in Theorem 6.1. Then, for all $n \geq 1, \varepsilon > 0$, and $\gamma > 0$, with probability at least $1 - M^{1-A^2/8}$ we have*

$$\|\widehat{f}_L - f\|_n^2 \leq (1 + \varepsilon) \inf_{\beta \in \bar{\Lambda}_{s,\gamma,\varepsilon}} \left\{ \|f_\beta - f\|_n^2 + \frac{C(\varepsilon)f_{\max}^2 A^2 \sigma^2}{\gamma^2} \left(\frac{\mathcal{M}(\beta) \log M}{n} \right) \right\},$$

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

To obtain this corollary, it suffices to observe that the proof of Theorem 6.1 goes through if we drop Assumption RE($s, 3 + 4/\varepsilon$), but we assume instead that $\beta \in \Lambda_{s,\gamma,3+4/\varepsilon}$, and we replace $\kappa(s, 3 + 4/\varepsilon)$ by γ .

We would like now to get a sparsity oracle inequality similar to that of Theorem 6.1 for the Dantzig estimator \widehat{f}_D . We will need a mild additional assumption on f . This is due to the fact that not every $\beta \in \mathbb{R}^M$ obeys the Dantzig constraint; thus, we cannot assure the key relation (B.9) for all $\beta \in \mathbb{R}^M$. One possibility would be to prove inequality as (6.1), where the infimum on the right hand side is taken over β satisfying not only $\mathcal{M}(\beta) \leq s$ but also the Dantzig constraint. However, this seems not to be very intuitive, since we cannot guarantee that the corresponding f_β gives a good approximation of the unknown function f . Therefore, we choose another approach (cf. [5]), in which 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

$$(6.2) \quad \Lambda_s = \left\{ \beta \in \mathbb{R}^M : \mathcal{M}(\beta) \leq s, \|f_\beta - f\|_n^2 \leq \frac{C_0 f_{\max}^2 r^2}{\kappa^2(s, 3 + 4/\varepsilon)} \mathcal{M}(\beta) \right\}$$

is nonempty. The second inequality in (6.2) says that the “bias” term $\|f_\beta - f\|_n^2$ cannot be much larger than the “variance term” $\sim f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\beta)$ [cf. (6.1)]. Weak sparsity is milder than the sparsity property in the usual sense. The latter means that f admits the exact representation $f = f_{\beta^*}$, for some $\beta^* \in \mathbb{R}^M$, with hopefully small $\mathcal{M}(\beta^*) = s$.

PROPOSITION 6.3. *Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix some $\varepsilon > 0$ and integers $n \geq 1, M \geq 2$. Let f obey the weak sparsity assumption for some $C_0 < \infty$ and some s such that $1 \leq s \max\{C_1(\varepsilon), 1\} \leq M$, where*

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

and $C(\varepsilon)$ is the constant in Theorem 6.1. Suppose, further, that Assumption RE($s \max\{C_1(\varepsilon), 1\}, 3 + 4/\varepsilon$) is satisfied. Consider the Dantzig estimator \widehat{f}_D defined by (2.5)–(2.4) with

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

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

$$(6.3) \quad \begin{aligned} & \|\widehat{f}_D - f\|_n^2 \\ & \leq (1 + \varepsilon) \inf_{\beta \in \mathbb{R}^M: \mathcal{M}(\beta)=s} \|f_\beta - f\|_n^2 + C_2(\varepsilon) \frac{f_{\max}^2 A^2 \sigma^2}{\kappa_0^2} \left(\frac{s \log M}{n} \right). \end{aligned}$$

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

Note that the sparsity oracle inequality (6.3) is slightly weaker than the analogous inequality (6.1) for the Lasso. Here, we have $\inf_{\beta \in \mathbb{R}^M: \mathcal{M}(\beta)=s}$ instead of $\inf_{\beta \in \mathbb{R}^M: \mathcal{M}(\beta) \leq s}$ in (6.1).

7. Special case. Parametric estimation in linear regression. In this section, we assume that the vector of observations $\mathbf{y} = (Y_1, \dots, Y_n)^T$ is of the form

$$(7.1) \quad \mathbf{y} = X\beta^* + \mathbf{w},$$

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

We consider dimension M that can be of order n and even much larger. Then, β^* is, in general, not uniquely defined. For $M > n$, if (7.1) is satisfied for $\beta^* = \beta_0$, then there exists an affine space $\mathcal{U} = \{\beta^* : X\beta^* = X\beta_0\}$ of vectors satisfying (7.1). The results of this section are valid for any β^* such that (7.1) holds. However, we will suppose that Assumption RE(s, c_0) holds with $c_0 \geq 1$ and that $\mathcal{M}(\beta^*) \leq s$. Then, the set $\mathcal{U} \cap \{\beta^* : \mathcal{M}(\beta^*) \leq s\}$ reduces to a single element (cf. Remark 2 at the end of this section). In this sense, there is a unique sparse solution of (7.1).

Our goal in this section, unlike that of the previous ones, is to estimate both $X\beta^*$ for the purpose of prediction and β^* itself for purpose of model selection. We will see that meaningful results are obtained when the sparsity index $\mathcal{M}(\beta^*)$ is small.

It will be assumed throughout this section that the diagonal elements of the Gram matrix $\Psi_n = X^T X/n$ 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 β^* in (7.1) is defined by

$$(7.2) \quad \widehat{\beta}_L = \arg \min_{\beta \in \mathbb{R}^M} \left\{ \frac{1}{n} |\mathbf{y} - X\beta|_2^2 + 2r|\beta|_1 \right\}.$$

The correspondence between the notation here and that of the previous sections is

$$\begin{aligned} \|f_\beta\|_n^2 &= |X\beta|_2^2/n, & \|f_\beta - f\|_n^2 &= |X(\beta - \beta^*)|_2^2/n, \\ \|\widehat{f}_L - f\|_n^2 &= |X(\widehat{\beta}_L - \beta^*)|_2^2/n. \end{aligned}$$

The Dantzig selector for linear model (7.1) is defined by

$$(7.3) \quad \widehat{\beta}_D = \arg \min_{\beta \in \Lambda} |\beta|_1,$$

where

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

is the set of all β satisfying the Dantzig constraint.

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

THEOREM 7.1. *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 $\mathcal{M}(\beta^*) \leq s$, where $1 \leq s \leq M$, $n \geq 1$, $M \geq 2$. Let Assumption RE($s, 1$) be satisfied. Consider the Dantzig selector $\hat{\beta}_D$ defined by (7.3) with*

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

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

$$(7.4) \quad |\hat{\beta}_D - \beta^*|_1 \leq \frac{8A}{\kappa^2(s, 1)} \sigma s \sqrt{\frac{\log M}{n}},$$

$$(7.5) \quad |X(\hat{\beta}_D - \beta^*)|_2^2 \leq \frac{16A^2}{\kappa^2(s, 1)} \sigma^2 s \log M.$$

If Assumption RE($s, m, 1$) is satisfied, then, with the same probability as above, simultaneously for all $1 < p \leq 2$, we have

$$(7.6) \quad |\hat{\beta}_D - \beta^*|_p^p \leq 2^{p-1} 8 \left\{ 1 + \sqrt{\frac{s}{m}} \right\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2(s, m, 1)} \sqrt{\frac{\log M}{n}} \right)^p.$$

Note that, since $s \leq m$, the factor in curly brackets in (7.6) is bounded by a constant independent of s and m . Under Assumption 1 in Section 4, with $c_0 = 1$ [which is less general than RE($s, s, 1$), cf. Lemma 4.1(i)], a bound of the form (7.6) 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 7.1. They are given by the following result.

THEOREM 7.2. *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 let $\mathcal{M}(\beta^*) \leq s$, where $1 \leq s \leq M$, $n \geq 1$, $M \geq 2$. Let Assumption RE($s, 3$) be satisfied. Consider the Lasso estimator $\hat{\beta}_L$ defined by (7.2) with*

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

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

$$(7.7) \quad |\widehat{\beta}_L - \beta^*|_1 \leq \frac{16A}{\kappa^2(s, 3)} \sigma s \sqrt{\frac{\log M}{n}},$$

$$(7.8) \quad |X(\widehat{\beta}_L - \beta^*)|_2^2 \leq \frac{16A^2}{\kappa^2(s, 3)} \sigma^2 s \log M,$$

$$(7.9) \quad \mathcal{M}(\widehat{\beta}_L) \leq \frac{64\phi_{\max}}{\kappa^2(s, 3)} s.$$

If Assumption RE($s, m, 3$) is satisfied, then, with the same probability as above, simultaneously for all $1 < p \leq 2$, we have

$$(7.10) \quad |\widehat{\beta}_L - \beta^*|_p^p \leq 16 \left\{ 1 + 3\sqrt{\frac{s}{m}} \right\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2(s, m, 3)} \sqrt{\frac{\log M}{n}} \right)^p.$$

Inequalities of the form similar to (7.7) and (7.8) can be deduced from the results of [3] under more restrictive conditions on the Gram matrix (the mutual coherence assumption, cf. Assumption 5 of Section 4).

Assumptions RE($s, 1$) and RE($s, 3$), respectively, can be dropped in Theorems 7.1 and 7.2 if we assume $\beta^* \in \Lambda_{s,\gamma,c_0}$ with $c_0 = 1$ or $c_0 = 3$ as appropriate. Then, (7.4) and (7.5) or, respectively, (7.7) and (7.8) hold with $\kappa = \gamma$. This is analogous to Corollary 6.2. Similarly, (7.6) and (7.10) 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 7.1 and 7.2, we can immediately get bounds for the differences between Lasso and Dantzig selector $|\widehat{\beta}_L - \widehat{\beta}_D|_p^p$ and $|X(\widehat{\beta}_L - \widehat{\beta}_D)|_2^2$. Such bounds have the same form as those of Theorems 7.1 and 7.2, up to numerical constants. Another way of estimating these differences follows directly from the proof of Theorem 7.1. It suffices to observe that the only property of β^* used in that proof is the fact that β^* satisfies the Dantzig constraint on the event of given probability, which is also true for the Lasso solution $\widehat{\beta}_L$. So, we can replace β^* by $\widehat{\beta}_L$ and s by $\mathcal{M}(\widehat{\beta}_L)$ everywhere in Theorem 7.1. Generalizing a bit more, we easily derive the following fact.

THEOREM 7.3. *The result of Theorem 7.1 remains valid if we replace $|\widehat{\beta}_D - \beta^*|_p^p$ by $\sup\{|\widehat{\beta}_D - \beta|_p^p : \beta \in \Lambda, \mathcal{M}(\beta) \leq s\}$ for $1 \leq p \leq 2$ and $|X(\widehat{\beta}_D - \beta^*)|_2^2$ by $\sup\{|X(\widehat{\beta}_D - \beta)|_2^2 : \beta \in \Lambda, \mathcal{M}(\beta) \leq s\}$, respectively. Here, Λ is the set of all vectors satisfying the Dantzig constraint.*

REMARKS.

1. Theorems 7.1 and 7.2 only give nonasymptotic 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, c_0) and RE(s, m, c_0) are imposed on the $n \times M$ matrix X . To deduce asymptotic conver-

gence (as $n \rightarrow \infty$ and/or as $M \rightarrow \infty$) from Theorems 7.1 and 7.2, we would need some very strong additional properties, such as simultaneous validity of Assumption RE(s, c_0) or RE(s, m, c_0) (with one and the same constant κ) for infinitely many n and M .

2. Note that neither Assumption RE(s, c_0) or RE(s, m, c_0) implies identifiability of β^* in the linear model (7.1). However, the vector β^* appearing in the statements of Theorems 7.1 and 7.2 is uniquely defined, because we additionally suppose that $\mathcal{M}(\beta^*) \leq s$ and $c_0 \geq 1$. Indeed, if there exists a β' such that $X\beta' = X\beta^*$, and $\mathcal{M}(\beta') \leq s$, then, in view of assumption RE(s, c_0) with $c_0 \geq 1$, we necessarily have $\beta^* = \beta'$ [cf. discussion following the definition of RE(s, c_0)]. On the other hand, Theorem 7.3 applies to certain values of β that do not come from the model (7.1) at all.

3. For the smallest value of A (which is $A = 2\sqrt{2}$) the constants in the bound of Theorem 7.2 for the Lasso are larger than the corresponding numerical constants for the Dantzig selector given in Theorem 7.1, again, for the smallest admissible value $A = \sqrt{2}$. On the contrary, the Dantzig selector has certain defects as compared to Lasso when the model is nonparametric, as discussed in Section 6. In particular, to obtain sparsity oracle inequalities for the Dantzig selector, we need some restrictions on f , for example, the weak sparsity property. On the other hand, the sparsity oracle inequality (6.1) for the Lasso is valid with no restriction on f .

4. The proofs of Theorems 7.1 and 7.2 differ mainly in the value of the tuning constant, which is $c_0 = 1$ in Theorem 7.1 and $c_0 = 3$ in Theorem 7.2. Note that, since the Lasso solution satisfies the Dantzig constraint, we could have obtained a result similar to Theorem 7.2, but with less accurate numerical constants, by simply conducting the proof of Theorem 7.1 with $c_0 = 3$. However, we act differently, and we deduce (B.30) directly from (B.1) and not from (B.25). This is done only for the sake of improving the constants. In fact, using (B.25) with $c_0 = 3$ would yield (B.30) with the doubled constant on the right-hand side.

5. For the 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 4.1 with the projector P_{01} inside. We do not pursue this issue here.

6. All of 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 for readability. Inspection of (B.4) 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

PROOF OF LEMMA 4.1. Consider a partition J_0^c into subsets of size m , with the last subset of size $\leq m$: $J_0^c = \bigcup_{k=1}^K J_k$, where $K \geq 1$, $|J_k| = m$ for

$k = 1, \dots, K - 1$ and $|J_K| \leq m$, such that J_k is the set of indices corresponding to m largest in absolute value coordinates of δ outside $\cup_{j=1}^{k-1} J_j$ (for $k < K$) and J_K is the remaining subset. We have

$$\begin{aligned}
 |P_{01} X \delta|_2 &\geq |P_{01} X \delta_{J_{01}}|_2 - \left| \sum_{k=2}^K P_{01} X \delta_{J_k} \right|_2 \\
 \text{(A.1)} \quad &= |X \delta_{J_{01}}|_2 - \left| \sum_{k=2}^K P_{01} X \delta_{J_k} \right|_2 \\
 &\geq |X \delta_{J_{01}}|_2 - \sum_{k=2}^K |P_{01} X \delta_{J_k}|_2.
 \end{aligned}$$

We will prove first part (ii) of the lemma. Since for $k \geq 1$ the vector δ_{J_k} has only m nonzero components, we obtain

$$\text{(A.2)} \quad \frac{1}{\sqrt{n}} |P_{01} X \delta_{J_k}|_2 \leq \frac{1}{\sqrt{n}} |X \delta_{J_k}|_2 \leq \sqrt{\phi_{\max}(m)} |\delta_{J_k}|_2.$$

Next, as in [7], we observe that $|\delta_{J_{k+1}}|_2 \leq |\delta_{J_k}|_1 / \sqrt{m}$, $k = 1, \dots, K - 1$. Therefore,

$$\text{(A.3)} \quad \sum_{k=2}^K |\delta_{J_k}|_2 \leq \frac{|\delta_{J_0^c}|_1}{\sqrt{m}} \leq \frac{c_0 |\delta_{J_0}|_1}{\sqrt{m}} \leq c_0 \sqrt{\frac{s}{m}} |\delta_{J_0}|_2 \leq c_0 \sqrt{\frac{s}{m}} |\delta_{J_{01}}|_2,$$

where we used (4.1). From (A.1)–(A.3), we find

$$\begin{aligned}
 \frac{1}{\sqrt{n}} |X \delta|_2 &\geq \frac{1}{\sqrt{n}} |X \delta_{J_{01}}|_2 - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} |\delta_{J_{01}}|_2 \\
 &\geq \left(\sqrt{\phi_{\min}(s+m)} - c_0 \sqrt{\phi_{\max}(m)} \sqrt{\frac{s}{m}} \right) |\delta_{J_{01}}|_2,
 \end{aligned}$$

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.2), we use the bound (cf. [7])

$$\frac{1}{\sqrt{n}} |P_{01} X \delta_{J_k}|_2 \leq \frac{\theta_{s,2s}}{\sqrt{\phi_{\min}(2s)}} |\delta_{J_k}|_2. \quad \square$$

APPENDIX B: TWO LEMMAS AND THE PROOFS OF THE RESULTS

LEMMA B.1. Fix $M \geq 2$ and $n \geq 1$. Let W_i be independent $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$, and let \hat{f}_L be the Lasso estimator defined by (2.2) with

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

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

$$\begin{aligned}
 & \|\widehat{f}_L - f\|_n^2 + r \sum_{j=1}^M \|f_j\|_n |\widehat{\beta}_{j,L} - \beta_j| \\
 \text{(B.1)} \quad & \leq \|f_\beta - f\|_n^2 + 4r \sum_{j \in J(\beta)} \|f_j\|_n |\widehat{\beta}_{j,L} - \beta_j| \\
 & \leq \|f_\beta - f\|_n^2 + 4r \sqrt{\mathcal{M}(\beta)} \sqrt{\sum_{j \in J(\beta)} \|f_j\|_n^2 |\widehat{\beta}_{j,L} - \beta_j|^2},
 \end{aligned}$$

and

$$\text{(B.2)} \quad \left| \frac{1}{n} X^T (\mathbf{f} - X \widehat{\beta}_L) \right|_\infty \leq 3r f_{\max} / 2.$$

Furthermore, with the same probability,

$$\text{(B.3)} \quad \mathcal{M}(\widehat{\beta}_L) \leq 4\phi_{\max} f_{\min}^{-2} (\|\widehat{f}_L - f\|_n^2 / r^2),$$

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

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

$$\widehat{S}(\widehat{\beta}_L) + 2 \sum_{j=1}^M r_{n,j} |\widehat{\beta}_{j,L}| \leq \widehat{S}(\beta) + 2 \sum_{j=1}^M r_{n,j} |\beta_j|$$

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

$$\begin{aligned}
 & \|\widehat{f}_L - f\|_n^2 + 2 \sum_{j=1}^M r_{n,j} |\widehat{\beta}_{j,L}| \\
 & \leq \|f_\beta - f\|_n^2 + 2 \sum_{j=1}^M r_{n,j} |\beta_j| + \frac{2}{n} \sum_{i=1}^n W_i (\widehat{f}_L - f_\beta)(Z_i).
 \end{aligned}$$

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

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

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

$$\begin{aligned}
 \mathbb{P}\{\mathcal{A}^c\} & \leq \sum_{j=1}^M \mathbb{P}\{\sqrt{n}|V_j| > \sqrt{n}r_{n,j}/2\} \leq M \mathbb{P}\{|\eta| \geq r\sqrt{n}/(2\sigma)\} \\
 \text{(B.4)} \quad & \leq M \exp\left(-\frac{nr^2}{8\sigma^2}\right) = M \exp\left(-\frac{A^2 \log M}{8}\right) = M^{1-A^2/8},
 \end{aligned}$$

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

$$\|\widehat{f}_L - f\|_n^2 \leq \|f_\beta - f\|_n^2 + \sum_{j=1}^M r_{n,j} |\widehat{\beta}_{j,L} - \beta_j| + \sum_{j=1}^M 2r_{n,j} |\beta_j| - \sum_{j=1}^M 2r_{n,j} |\widehat{\beta}_{j,L}|.$$

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

$$\begin{aligned} & \|\widehat{f}_L - f\|_n^2 + \sum_{j=1}^M r_{n,j} |\widehat{\beta}_{j,L} - \beta_j| \\ & \leq \|f_\beta - f\|_n^2 + 2 \sum_{j=1}^M r_{n,j} (|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}|). \end{aligned}$$

Now, $|\widehat{\beta}_{j,L} - \beta_j| + |\beta_j| - |\widehat{\beta}_{j,L}| = 0$ for $j \notin J(\beta)$, so that, on \mathcal{A} , we get (B.1).

To prove (B.2) it suffices to note that, on \mathcal{A} , we have

$$(B.5) \quad \left| \frac{1}{n} D^{-1/2} X^T W \right|_\infty \leq r/2.$$

Now, $\mathbf{y} = \mathbf{f} + \mathbf{w}$, and (B.2) follows from (2.3) and (B.5).

We finally prove (B.3). The necessary and sufficient condition for $\widehat{\beta}_L$ to be the Lasso solution can be written in the form

$$(B.6) \quad \begin{aligned} & \frac{1}{n} \mathbf{x}_{(j)}^T (y - X\widehat{\beta}_L) = r \|f_j\|_n \text{sign}(\widehat{\beta}_{j,L}) \quad \text{if } \widehat{\beta}_{j,L} \neq 0, \\ & \left| \frac{1}{n} \mathbf{x}_{(j)}^T (y - X\widehat{\beta}_L) \right| \leq r \|f_j\|_n \quad \text{if } \widehat{\beta}_{j,L} = 0, \end{aligned}$$

where $\mathbf{x}_{(j)}$ denotes the j th column of X , $j = 1, \dots, M$. Next, (B.5) yields that, on \mathcal{A} , we have

$$(B.7) \quad \left| \frac{1}{n} \mathbf{x}_{(j)}^T W \right| \leq r \|f_j\|_n / 2, \quad j = 1, \dots, M.$$

Combining (B.6) and (B.7), we get

$$(B.8) \quad \left| \frac{1}{n} \mathbf{x}_{(j)}^T (\mathbf{f} - X\widehat{\beta}_L) \right| \geq r \|f_j\|_n / 2 \quad \text{if } \widehat{\beta}_{j,L} \neq 0.$$

Therefore,

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

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

$$\frac{1}{n^2}(\mathbf{f} - X\hat{\beta}_L)^T X X^T (\mathbf{f} - X\hat{\beta}_L) \leq \frac{\phi_{\max}}{n} |\mathbf{f} - X\hat{\beta}_L|_2^2 = \phi_{\max} \|f - \hat{f}_L\|_n^2,$$

and we deduce (B.3) from the last two displays. \square

COROLLARY B.2. *Let the assumptions of Lemma B.1 be satisfied and $\|f_j\|_n = 1, j = 1, \dots, M$. Consider the linear regression model $\mathbf{y} = X\beta + \mathbf{w}$. Then, with probability at least $1 - M^{1-A^2/8}$, we have*

$$|\delta_{J_0^c}|_1 \leq 3|\delta_{J_0}|_1,$$

where $J_0 = J(\beta)$ is the set of nonzero coefficients of β and $\delta = \hat{\beta}_L - \beta$.

PROOF. Use the first inequality in (B.1) and the fact that $f = f_\beta$ for the linear regression model. \square

LEMMA B.3. *Let $\beta \in \mathbb{R}^M$ satisfy the Dantzig constraint*

$$\left| \frac{1}{n} D^{-1/2} X^T (\mathbf{y} - X\beta) \right|_\infty \leq r$$

and set $\delta = \hat{\beta}_D - \beta, J_0 = J(\beta)$. Then,

$$(B.9) \quad |\delta_{J_0^c}|_1 \leq |\delta_{J_0}|_1.$$

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

$$(B.10) \quad \left| \frac{1}{n} X^T (\mathbf{f} - X\hat{\beta}_D) \right|_\infty \leq 2r f_{\max}.$$

PROOF OF LEMMA B.3. Inequality (B.9) follows immediately from the definition of Dantzig selector (cf. [7]). To prove (B.10), consider the event

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

Analogously to (B.4), $\mathbb{P}\{\mathcal{B}^c\} \leq M^{1-A^2/2}$. On the other hand, $\mathbf{y} = \mathbf{f} + \mathbf{w}$, and, using the definition of Dantzig selector, it is easy to see that (B.10) is satisfied on \mathcal{B} . \square

PROOF OF THEOREM 5.1. Set $\delta = \hat{\beta}_L - \hat{\beta}_D$. We have

$$\frac{1}{n} |\mathbf{f} - X\hat{\beta}_L|_2^2 = \frac{1}{n} |\mathbf{f} - X\hat{\beta}_D|_2^2 - \frac{2}{n} \delta^T X^T (\mathbf{f} - X\hat{\beta}_D) + \frac{1}{n} |X\delta|_2^2.$$

This and (B.10) yield

$$\begin{aligned}
 \|\widehat{f}_D - f\|_n^2 &\leq \|\widehat{f}_L - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left| \frac{1}{n} X^T(\mathbf{f} - X\widehat{\beta}_D) \right|_\infty - \frac{1}{n} |X\boldsymbol{\delta}|_2^2 \\
 &\leq \|\widehat{f}_L - f\|_n^2 + 4f_{\max} r |\boldsymbol{\delta}|_1 - \frac{1}{n} |X\boldsymbol{\delta}|_2^2,
 \end{aligned}
 \tag{B.11}$$

where the last inequality holds with probability at least $1 - M^{1-A^2/2}$. Since the Lasso solution $\widehat{\beta}_L$ satisfies the Dantzig constraint, we can apply Lemma B.3 with $\beta = \widehat{\beta}_L$, which yields

$$|\boldsymbol{\delta}_{J_0^c}|_1 \leq |\boldsymbol{\delta}_{J_0}|_1 \tag{B.12}$$

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

$$\frac{1}{\sqrt{n}} |X\boldsymbol{\delta}|_2 \geq \kappa |\boldsymbol{\delta}_{J_0}|_2, \tag{B.13}$$

where $\kappa = \kappa(s, 1)$. Using (B.12) and (B.13), we obtain

$$|\boldsymbol{\delta}|_1 \leq 2|\boldsymbol{\delta}_{J_0}|_1 \leq 2\mathcal{M}^{1/2}(\widehat{\beta}_L) |\boldsymbol{\delta}_{J_0}|_2 \leq \frac{2\mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}} |X\boldsymbol{\delta}|_2. \tag{B.14}$$

Finally, from (B.11) and (B.14), we get that, with probability at least $1 - M^{1-A^2/2}$,

$$\begin{aligned}
 \|\widehat{f}_D - f\|_n^2 &\leq \|\widehat{f}_L - f\|_n^2 + \frac{8f_{\max} r \mathcal{M}^{1/2}(\widehat{\beta}_L)}{\kappa\sqrt{n}} |X\boldsymbol{\delta}|_2 - \frac{1}{n} |X\boldsymbol{\delta}|_2^2 \\
 &\leq \|\widehat{f}_L - f\|_n^2 + \frac{16f_{\max}^2 r^2 \mathcal{M}(\widehat{\beta}_L)}{\kappa^2},
 \end{aligned}
 \tag{B.15}$$

where the RHS follows (B.2), (B.10) and another application of (B.14). This proves one side of the inequality.

To show the other side of the bound on the difference, we act as in (B.11), up to the inversion of roles of $\widehat{\beta}_L$ and $\widehat{\beta}_D$, and we use (B.2). This yields that, with probability at least $1 - M^{1-A^2/8}$,

$$\begin{aligned}
 \|\widehat{f}_L - f\|_n^2 &\leq \|\widehat{f}_D - f\|_n^2 + 2|\boldsymbol{\delta}|_1 \left| \frac{1}{n} X^T(\mathbf{f} - X\widehat{\beta}_L) \right|_\infty - \frac{1}{n} |X\boldsymbol{\delta}|_2^2 \\
 &\leq \|\widehat{f}_D - f\|_n^2 + 3f_{\max} r |\boldsymbol{\delta}|_1 - \frac{1}{n} |X\boldsymbol{\delta}|_2^2.
 \end{aligned}
 \tag{B.16}$$

This is analogous to (B.11). Now, paralleling the proof leading to (B.15), we obtain

$$\|\widehat{f}_L - f\|_n^2 \leq \|\widehat{f}_D - f\|_n^2 + \frac{9f_{\max}^2 r^2 \mathcal{M}(\widehat{\beta}_L)}{\kappa^2}. \tag{B.17}$$

The theorem now follows from (B.15) and (B.17). \square

PROOF OF THEOREM 5.2. Set, again, $\delta = \widehat{\beta}_L - \widehat{\beta}_D$. We apply (B.1) with $\beta = \widehat{\beta}_D$, which yields that, with probability at least $1 - M^{1-A^2/8}$,

$$(B.18) \quad |\delta|_1 \leq 4|\delta_{J_0}|_1 + \|\widehat{f}_D - f\|_n^2/r,$$

where, now, $J_0 = J(\widehat{\beta}_D)$. Consider the following two cases: (i) $\|\widehat{f}_D - f\|_n^2 > 2r|\delta_{J_0}|_1$ and (ii) $\|\widehat{f}_D - f\|_n^2 \leq 2r|\delta_{J_0}|_1$. In case (i), inequality (B.16) with $f_{\max} = 1$ immediately implies

$$\|\widehat{f}_L - f\|_n^2 \leq 10\|\widehat{f}_D - f\|_n^2,$$

and the theorem follows. In case (ii), we get, from (B.18), that

$$|\delta|_1 \leq 6|\delta_{J_0}|_1$$

and thus $|\delta_{J_0^c}|_1 \leq 5|\delta_{J_0}|_1$. We can therefore apply Assumption RE($s, 5$), which yields, similarly to (B.14),

$$(B.19) \quad |\delta|_1 \leq 6\mathcal{M}^{1/2}(\widehat{\beta}_D)|\delta_{J_0}|_2 \leq \frac{6\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\delta|_2,$$

where $\kappa = \kappa(s, 5)$. Plugging (B.19) into (B.16) we finally get that, in case (ii),

$$(B.20) \quad \begin{aligned} \|\widehat{f}_L - f\|_n^2 &\leq \|\widehat{f}_D - f\|_n^2 + \frac{18r\mathcal{M}^{1/2}(\widehat{\beta}_D)}{\kappa\sqrt{n}}|X\delta|_2 - \frac{1}{n}|X\delta|_2^2 \\ &\leq \|\widehat{f}_D - f\|_n^2 + \frac{81r^2\mathcal{M}(\widehat{\beta}_D)}{\kappa^2}. \end{aligned} \quad \square$$

PROOF OF THEOREM 6.1. Fix an arbitrary $\beta \in \mathbb{R}^M$ with $\mathcal{M}(\beta) \leq s$. Set $\delta = D^{1/2}(\widehat{\beta}_L - \beta)$, $J_0 = J(\beta)$. On the event \mathcal{A} , we get, from the first line in (B.1), that

$$(B.21) \quad \begin{aligned} \|\widehat{f}_L - f\|_n^2 + r|\delta|_1 &\leq \|f_\beta - f\|_n^2 + 4r \sum_{j \in J_0} \|f_j\|_n |\widehat{\beta}_{j,L} - \beta_j| \\ &= \|f_\beta - f\|_n^2 + 4r|\delta_{J_0}|_1, \end{aligned}$$

and from the second line in (B.1) that

$$(B.22) \quad \|\widehat{f}_L - f\|_n^2 \leq \|f_\beta - f\|_n^2 + 4r\sqrt{\mathcal{M}(\beta)}|\delta_{J_0}|_2.$$

Consider, separately, the cases where

$$(B.23) \quad 4r|\delta_{J_0}|_1 \leq \varepsilon\|f_\beta - f\|_n^2$$

and

$$(B.24) \quad \varepsilon\|f_\beta - f\|_n^2 < 4r|\delta_{J_0}|_1.$$

In case (B.23), the result of the theorem trivially follows from (B.21). So, we will only consider the case (B.24). All of the subsequent inequalities are valid on the

event $\mathcal{A} \cap \mathcal{A}_1$, where \mathcal{A}_1 is defined by (B.24). On this event, we get, from (B.21), that

$$|\delta|_1 \leq 4(1 + 1/\varepsilon)|\delta_{J_0}|_1,$$

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

$$\begin{aligned} \kappa^2 |\delta_{J_0}|_2^2 &\leq \frac{1}{n} |X\delta|_2^2 = \frac{1}{n} (\widehat{\beta}_K - \beta)^T D^{1/2} X^T X D^{1/2} (\widehat{\beta}_L - \beta) \\ &\leq \frac{f_{\max}^2}{n} (\widehat{\beta}_L - \beta)^T X^T X (\widehat{\beta}_L - \beta) = f_{\max}^2 \|\widehat{f}_L - f_\beta\|_n^2, \end{aligned}$$

where $\kappa = \kappa(s, 3 + 4/\varepsilon)$. Combining this with (B.22), we find

$$\begin{aligned} \|\widehat{f}_L - f\|_n^2 &\leq \|f_\beta - f\|_n^2 + 4r f_{\max} \kappa^{-1} \sqrt{\mathcal{M}(\beta)} \|\widehat{f}_L - f_\beta\|_n \\ &\leq \|f_\beta - f\|_n^2 + 4r f_{\max} \kappa^{-1} \sqrt{\mathcal{M}(\beta)} (\|\widehat{f}_L - f\|_n + \|f_\beta - f\|_n). \end{aligned}$$

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

$$\|\widehat{f}_L - f\|_n^2 \leq \frac{b+1}{b-1} \|f_\beta - f\|_n^2 + \frac{8b^2 f_{\max}^2}{(b-1)\kappa^2} r^2 \mathcal{M}(\beta) \quad \forall b > 1.$$

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

PROOF OF PROPOSITION 6.3. Due to the weak sparsity assumption, there exists $\bar{\beta} \in \mathbb{R}^M$ with $\mathcal{M}(\bar{\beta}) \leq s$ such that $\|f_{\bar{\beta}} - f\|_n^2 \leq C_0 f_{\max}^2 r^2 \kappa^{-2} \mathcal{M}(\bar{\beta})$, where $\kappa = \kappa(s, 3 + 4/\varepsilon)$ is the same as in Theorem 6.1. Using this together with Theorem 6.1 and (B.3), we obtain that, with probability at least $1 - M^{1-A^2/8}$,

$$\mathcal{M}(\widehat{\beta}_L) \leq C_1(\varepsilon) \mathcal{M}(\bar{\beta}) \leq C_1(\varepsilon) s.$$

This and Theorem 5.1 imply

$$\|\widehat{f}_D - f\|_n^2 \leq \|\widehat{f}_L - f\|_n^2 + \frac{16C_1(\varepsilon) f_{\max}^2 A^2 \sigma^2}{\kappa_0^2} \left(\frac{s \log M}{n} \right),$$

where $\kappa_0 = \kappa(\max(C_1(\varepsilon), 1)s, 3 + 4/\varepsilon)$. Once Again, applying Theorem 6.1, we get the result. \square

PROOF OF THEOREM 7.1. Set $\delta = \widehat{\beta}_D - \beta^*$ and $J_0 = J(\beta^*)$. Using Lemma B.3 with $\beta = \beta^*$, we get that, on the event \mathcal{B} (i.e., with probability at least

$1 - M^{1-A^2/2}$, the following are true: (i) $\frac{1}{n}|X^T X \delta|_\infty \leq 2r$, and (ii) inequality (4.1) holds with $c_0 = 1$. Therefore, on \mathcal{B} we have

$$\begin{aligned}
 \frac{1}{n}|X\delta|_2^2 &= \frac{1}{n}\delta^T X^T X \delta \\
 &\leq \frac{1}{n}|X^T X \delta|_\infty |\delta|_1 \\
 \text{(B.25)} \quad &\leq 2r(|\delta_{J_0}|_1 + |\delta_{J_0^c}|_1) \\
 &\leq 2(1 + c_0)r|\delta_{J_0}|_1 \\
 &\leq 2(1 + c_0)r\sqrt{s}|\delta_{J_0}|_2 = 4r\sqrt{s}|\delta_{J_0}|_2
 \end{aligned}$$

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

$$\frac{1}{n}|X\delta|_2^2 \geq \kappa^2 |\delta_{J_0}|_2^2,$$

where $\kappa = \kappa(s, 1)$. This and (B.25) yield that, on \mathcal{B} ,

$$\text{(B.26)} \quad \frac{1}{n}|X\delta|_2^2 \leq 16r^2s/\kappa^2, \quad |\delta_{J_0}|_2 \leq 4r\sqrt{s}/\kappa^2.$$

The first inequality in (B.26) implies (7.5). Next, (7.4) is straightforward in view of the second inequality in (B.26) and of the relations (with $c_0 = 1$)

$$\text{(B.27)} \quad |\delta|_1 = |\delta_{J_0}|_1 + |\delta_{J_0^c}|_1 \leq (1 + c_0)|\delta_{J_0}|_1 \leq (1 + c_0)\sqrt{s}|\delta_{J_0}|_2$$

that hold on \mathcal{B} . It remains to prove (7.6). It is easy to see that the k th largest in absolute value element of $\delta_{J_0^c}$ satisfies $|\delta_{J_0^c}|_{(k)} \leq |\delta_{J_0^c}|_1/k$. Thus,

$$|\delta_{J_0^c}|_2^2 \leq |\delta_{J_0^c}|_1^2 \sum_{k \geq m+1} \frac{1}{k^2} \leq \frac{1}{m} |\delta_{J_0^c}|_1^2,$$

and, since (4.1) holds on \mathcal{B} (with $c_0 = 1$), we find

$$|\delta_{J_0^c}|_2 \leq \frac{c_0|\delta_{J_0}|_1}{\sqrt{m}} \leq c_0|\delta_{J_0}|_2\sqrt{\frac{s}{m}} \leq c_0|\delta_{J_0}|_2\sqrt{\frac{s}{m}}.$$

Therefore, on \mathcal{B} ,

$$\text{(B.28)} \quad |\delta|_2 \leq \left(1 + c_0\sqrt{\frac{s}{m}}\right) |\delta_{J_0}|_2.$$

On the other hand, it follows from (B.25) that

$$\frac{1}{n}|X\delta|_2^2 \leq 4r\sqrt{s}|\delta_{J_0}|_2.$$

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

$$|\delta_{J_0}|_2 \leq 4r\sqrt{s}/\kappa^2.$$

Recalling that $c_0 = 1$ and applying the last inequality together with (B.28), we get

$$(B.29) \quad |\delta|_2^2 \leq 16 \left(1 + c_0 \sqrt{\frac{s}{m}} \right)^2 (r\sqrt{s}/\kappa^2)^2.$$

It remains to note that (7.6) is a direct consequence of (7.4) and (B.29). This follows from the fact that inequalities $\sum_{j=1}^M a_j \leq b_1$ and $\sum_{j=1}^M a_j^2 \leq b_2$ with $a_j \geq 0$ imply

$$\begin{aligned} \sum_{j=1}^M a_j^p &= \sum_{j=1}^M a_j^{2-p} a_j^{2p-2} \leq \left(\sum_{j=1}^M a_j \right)^{2-p} \left(\sum_{j=1}^M a_j^2 \right)^{p-1} \\ &\leq b_1^{2-p} b_2^{p-1} \quad \forall 1 < p \leq 2. \end{aligned} \quad \square$$

PROOF OF THEOREM 7.2. Set $\delta = \widehat{\beta}_L - \beta^*$ and $J_0 = J(\beta^*)$. Using (B.1), where we put $\beta = \beta^*$, $r_{n,j} \equiv r$ and $\|f_\beta - f\|_n = 0$, we get that, on the event \mathcal{A} ,

$$(B.30) \quad \frac{1}{n} |X\delta|_2^2 \leq 4r\sqrt{s} |\delta_{J_0}|_2$$

and (4.1) 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} ,

$$(B.31) \quad \frac{1}{n} |X\delta|_2^2 \leq 16r^2 s / \kappa^2, \quad |\delta_{J_0}|_2 \leq 4r\sqrt{s} / \kappa^2,$$

where $\kappa = \kappa(s, 3)$. The first inequality here coincides with (7.8). Next, (7.9) follows immediately from (B.3) and (7.8). To show (7.7), it suffices to note that on the event \mathcal{A} the relations (B.27) hold with $c_0 = 3$, to apply the second inequality in (B.31) and to use (B.4).

Finally, the proof of (7.10) follows exactly the same lines as that of (7.6). The only difference is that one should set $c_0 = 3$ in (B.28) and (B.29), as well as in the display preceding (B.28). \square

REFERENCES

- [1] BICKEL, P. J. (2007). Discussion of “The Dantzig selector: Statistical estimation when p is much larger than n ,” by E. Candes and T. Tao. *Ann. Statist.* **35** 2352–2357. MR2382645
- [2] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2004). Aggregation for regression learning. Preprint LPMA, Univ. Paris 6–Paris 7, n° 948. Available at [arXiv:math.ST/0410214](https://arxiv.org/abs/math/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. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)* (G. Lugosi and H. U. Simon, eds.). *Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer, Berlin. MR2280619

- [4] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. [MR2351101](#)
- [5] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.* **1** 169–194. [MR2312149](#)
- [6] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparse density estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007)* (N. H. Bshouty and C. Gentile, eds.). *Lecture Notes in Artificial Intelligence* **4539** 530–543. Springer, Berlin. [MR2397610](#)
- [7] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. [MR2382644](#)
- [8] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse over-complete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. [MR2237332](#)
- [9] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451. [MR2060166](#)
- [10] FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. [MR2415737](#)
- [11] FU, W. and KNIGHT, K. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. [MR1805787](#)
- [12] GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988. [MR2108039](#)
- [13] JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28** 681–712. [MR1792783](#)
- [14] KOLTCHINSKII, V. (2006). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* To appear.
- [15] KOLTCHINSKII, V. (2007). Dantzig selector and sparsity oracle inequalities. Unpublished manuscript.
- [16] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The Group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71.
- [17] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#)
- [18] MEINSHAUSEN, N. and YU, B. (2006). Lasso type recovery of sparse representations for high dimensional data. *Ann. Statist.* To appear.
- [19] NEMIROVSKI, A. (2000). Topics in nonparametric statistics. In *Ecole d’Eté de Probabilités de Saint-Flour XXVIII—1998. Lecture Notes in Math.* **1738**. Springer, New York. [MR1775640](#)
- [20] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000a). On the Lasso and its dual. *J. Comput. Graph. Statist.* **9** 319–337. [MR1822089](#)
- [21] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000b). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–404.
- [22] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- [23] TSYBAKOV, A. B. (2006). Discussion of “Regularization in Statistics,” by P. Bickel and B. Li. *TEST* **15** 303–310. [MR2273731](#)
- [24] TURLACH, B. A. (2005). On algorithms for solving least squares problems under an L1 penalty or an L1 constraint. In *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]* 2572–2577. Amer. Statist. Assoc., Alexandria, VA.
- [25] VAN DE GEER, S. A. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. [MR2396809](#)
- [26] ZHANG, C.-H. and HUANG, J. (2008). Model-selection consistency of the Lasso in high-dimensional regression. *Ann. Statist.* **36** 1567–1594. [MR2435448](#)

- [27] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
CALIFORNIA
USA
E-MAIL: bickel@stat.berkeley.edu

Y. RITOV
DEPARTMENT OF STATISTICS
FACULTY OF SOCIAL SCIENCES
THE HEBREW UNIVERSITY
JERUSALEM 91904
ISRAEL
E-MAIL: yaacov.ritov@gmail.com

A. B. TSYBAKOV
LABORATOIRE DE STATISTIQUE
CREST
3, AVENUE PIERRE LAROUSSE,
92240 MALAKOFF
AND
LPMA (UMR CNRS 1599)
UNIVERSITÉ PARIS VI
4, PLACE JUSSIEU,
75252 PARIS, CEDEX 05
FRANCE
E-MAIL: alexandre.tsybakov@upmc.fr