Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 571361, 6 pages
http://dx.doi.org/10.1155/2013/571361

Research Article

The Oracle Inequalities on Simultaneous Lasso and Dantzig Selector in High-Dimensional Nonparametric Regression

Shiqing Wang and Limin Su

College of Mathematics and Information Sciences, North China University of Water Resources and Electric Power, Zhengzhou 450011, China

Correspondence should be addressed to Shiqing Wang; wangshiqing@ncwu.edu.cn

Received 11 March 2013; Accepted 15 May 2013

Academic Editor: Gianluca Ranzi

During the last few years, a great deal of attention has been focused on Lasso and Dantzig selector in high-dimensional linear regression when the number of variables can be much larger than the sample size. Under a sparsity scenario, the authors (see, e.g., Bickel et al., 2009, Bunea et al., 2007, Candes and Tao, 2007, Candès and Tao, 2007, Donoho et al., 2006, Koltchinskii, 2009, Koltchinskii, 2009, Meinshausen and Yu, 2009, Rosenbaum and Tsybakov, 2010, Tsybakov, 2006, van de Geer, 2008, and Zhang and Huang, 2008) discussed the relations between Lasso and Dantzig selector and derived sparsity oracle inequalities for the prediction risk and bounds on the L_p estimation loss. In this paper, we point out that some of the authors overemphasize the role of some sparsity conditions, and the assumptions based on this sparsity condition may cause bad results. We give better assumptions and the methods that avoid using the sparsity condition. As a comparison with the results by Bickel et al., 2009, more precise oracle inequalities for the prediction risk and bounds on the L_p estimation loss are derived 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 (e.g., see [1–12]). Quite recently, Candes and Tao [13] have proposed the Dantzig estimate for such linear models, and other authors [1, 6, 14–22] have discussed the Dantzig estimate and established the 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 (see [23–26]). In particular, Bunea et al. [23, 24] 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. 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 Koltchinskii [27] and van de Geer [28]. Sparsity oracle inequalities for the Dantzig selector with random design are obtained by Koltchinskii [29]. In linear fixed design regression, Meinshausen and Yu [7] establish a bound on the L_2 loss for the coefficients of Lasso that are quite different from the bound on the same loss for the Dantzig selector proven in Candes and Tao [13]. Bickel et al. [15] show 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 $1 \le p \le 2$. In the nonparametric regression model, they prove sparsity oracle inequalities for the Lasso and the Dantzig selector. Moreover, the Lasso and the Dantzig selector are approximately equivalent in terms of the prediction loss. They develop geometrical assumptions that are considerably weaker than those of Candes and Tao [13] for the Dantzig selector and Bunea et al. [23] for the Lasso.

We give the assumptions equivalent with assumptions by Bickel et al. [15] and derive oracle inequalities that are more precise than Bickel et al.'s [15] for the prediction risk

in the general nonparametric regression model and bounds that are more precise than Bickel et al.'s [15] on the L_p estimation loss in the linear model when the number of variables can be much larger than the sample size. We begin, in the next section, by defining the Lasso and Dantzig procedures and the notation. In Section 3, we present our key three assumptions and discuss the relations between the assumptions and assumptions by Bickel et al. [15]. In Section 4, we give some equivalent results and sparsity oracle inequalities for the Lasso and Dantzig estimators in the general nonparametric regression model and improve corresponding results by Bickel et al. [15]. The concluding remarks are given in Section 5.

2. Definitions and Notations

Unless stated otherwise, all of our notations, definitions, and terminologies follow Bickel et al. [15]. Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be a sample of independent random pairs with

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

where $f : \mathbb{Z} \to \mathbb{R}$ is an unknown regression function to be estimated, $\mathbb{Z}$ is a Borel subset of $\mathbb{R}^d$, the Z_i's are fixed elements in $\mathbb{Z}$, and the regression errors W_i are Gaussian. Let $F_M = \{f_1, \ldots, f_M\}$ be a finite dictionary of functions $f_j : \mathbb{Z} \to \mathbb{R}$, $j = 1, \ldots, M$. We assume throughout that $M \geq 2$.

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

$$\mathbf{y} = \mathbf{f} + \mathbf{w}, \tag{2}$$

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

$$\|g\|_n = \sqrt{\frac{1}{n}|g|_2^2} \tag{3}$$

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

$$f_{\max} = \max_{1 \leq j \leq M} \|f_j\|_n, \qquad f_{\min} = \min_{1 \leq j \leq M} \|f_j\|_n. \tag{4}$$

For any $\beta = (\beta_1, \ldots, \beta_M)^T \in \mathbb{R}^M$ and $Z \in \mathbb{Z}$, define $f_\beta(Z) = \sum_{j=1}^M \beta_j f_j(Z)$ and $\mathbf{f}_\beta = (f_\beta(Z_1), \ldots, f_\beta(Z_n))^T = X\beta$. The estimates we consider are all of the form $f_{\widetilde{\beta}}(\cdot)$, where $\widetilde{\beta}$ is data determined. Since we consider mainly sparse vectors $\widetilde{\beta}$, it will be convenient to define the following. Let

$$M(\beta) = \sum_{j=1}^M I_{\{\beta_j \neq 0\}} = |J(\beta)| \tag{5}$$

denote the number of nonzero coordinates of β, where $I_{\{\cdot\}}$ denotes the indicator function, $J(\beta) = \{j \in \{1, \ldots, M\} : \beta_j \neq 0\}$, and $|J|$ denotes the cardinality of J. For a vector $\delta \in \mathbb{R}^M$ and a subset $J \subset \{1, \ldots, M\}$, 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.

Define the Lasso solution $\widehat{\beta}_L = (\widehat{\beta}_{1,L}, \ldots, \widehat{\beta}_{M,L})^T$ by

$$\widehat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^M} \left\{ \frac{1}{n}|y - X\beta|_2^2 + 2r\sum_{j=1}^M \|f_j\|_n |\beta_j| \right\}, \tag{6}$$

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

$$\widehat{f}_L(Z) = f_{\widehat{\beta}_L}(Z) = \sum_{j=1}^M \widehat{\beta}_{j,L} f_j(Z). \tag{7}$$

The Dantzig selector is defined by

$$\widehat{\beta}_D = \arg\min \left\{ |\beta|_1 : \frac{1}{n}\left|D^{-1/2}X^T(y - X\beta)\right|_\infty \leq r \right\}, \tag{8}$$

where D is the diagonal matrix

$$D = \operatorname{diag}\left\{\|f_1\|_n^2, \|f_2\|_n^2, \ldots, \|f_M\|_n^2\right\}. \tag{9}$$

The Dantzig estimator is defined by

$$\widehat{f}_D(Z) = f_{\widehat{\beta}_D}(Z) = \sum_{j=1}^M \widehat{\beta}_{j,D} f_j(Z), \tag{10}$$

where $\widehat{\beta}_D = (\widehat{\beta}_{1,D}, \ldots, \widehat{\beta}_{M,D})^T$ is the Dantzig selector.

We refer to Bickel et al. [15] for detailed discussion of the Dantzig constraint and the constraint that the Lasso selector satisfies.

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

$$\Psi_n = \frac{1}{n}X^T X = \left(\frac{1}{n}\sum_{i=1}^n f_j(Z_i) f_k(Z_i)\right), \quad 1 \leq j, k \leq M, \tag{11}$$

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

3. Discussion of the Assumptions

Under the sparsity scenario, we are typically interested in the case where $M > n$ and even $M \gg n$. Here, sparsity specifies that the high-dimensional vector β has coefficients that are mostly 0. Clearly, the matrix Ψ_n is degenerate, and ordinary least squares do not work in this case, since the require positive definiteness of Ψ_n. That is,

$$\min_{\delta \in \mathbb{R}^M, \delta \neq 0} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} > 0. \tag{12}$$

It turns out that the Lasso and Dantzig selector require much weaker assumptions. The idea by Bickel et al. [15] is that the minimum in (12) be replaced by the minimum over a restricted set of vectors, and the norm $|\delta|_2$ in the denominator of the condition be replaced by the L_2 norm of only

a part of δ. This is feasible. Because for the linear regression model, the residuals $\delta = \widehat{\beta}_L - \beta$ and $\delta = \widehat{\beta}_D - \beta$ satisfy

$$\left|\delta_{J_0^c}\right|_1 \le c_0 \left|\delta_{J_0}\right|_1 \tag{13}$$

with $c_0 = 1$ by Candes and Tao [13] and $c_0 = 3$ by Bickel et al. [15], respectively, where $c_0 > 0$ and $J_0 = J(\beta)$ is the set of nonzero coefficients of the true parameter β of the model; therefore, for any δ satisfying (13), we have

$$\frac{\delta^T \Psi_n \delta}{|\delta|_2^2} \ge \frac{\delta^T \Psi \delta}{|\delta|_2^2} - \varepsilon_n (1 + c_0)^2 \left|J_0\right|, \tag{14}$$

where Ψ is a positive definite matrix and $\varepsilon_n = \max |(\Psi_n - \Psi)_{ij}|$. Thus, we have a kind of "restricted" positive definiteness if $\varepsilon_n |J_0|$ is small enough. This results in the following restricted eigenvalue (RE) assumption.

Assumption RE(s, c_0) (Bickel et al. [15]). For some integer s such that $1 \le s \le M$ and a positive number c_0, the following condition holds:

$$\kappa(s, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,M\}, |J_0| \le s} \min_{\delta \ne 0, \left|\delta_{J_0^c}\right|_1 \le c_0 \left|\delta_{J_0}\right|_1} \frac{|X\delta|_2}{\sqrt{n}\left|\delta_{J_0}\right|_2} > 0. \tag{15}$$

The purpose of giving this assumption may be in order to facilitate the use of $|\delta_{J_0^c}|_1 \le c_0 |\delta_{J_0}|_1$ since they frequently use it in the proofs of their theorems and so do Candes and Tao [13].

Note that the role of $|\delta_{J_0^c}|_1 \le c_0 |\delta_{J_0}|_1$ is only to restrict set of vectors; that is, $\{\delta \in \mathbb{R}^M : \delta \ne 0\}$ restricts to $\{\delta \in \mathbb{R}^M : \delta \ne 0, |\delta_{J_0^c}|_1 \le c_0 |\delta_{J_0}|_1\}$. Therefore, it is not necessary that the norm $|\delta|_2$ in the denominator of (12) be replaced by the L_2 norm of only a part of δ. We give the following assumptions.

Assumption RE$\tau_1(s, c_0)$. For some integer s such that $1 \le s \le M$ and a positive number c_0, the following condition holds:

$$\tau_1(s, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,M\}, |J_0| \le s} \min_{\delta \ne 0, \left|\delta_{J_0^c}\right|_1 \le c_0 \left|\delta_{J_0}\right|_1} \frac{|X\delta|_2}{\sqrt{n}\left|\delta_{J_0}\right|_1} > 0. \tag{16}$$

Assumption RE$\tau_2(s, c_0)$. For some integer s such that $1 \le s \le M$ and a positive number c_0, the following condition holds:

$$\tau_2(s, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,M\}, |J_0| \le s} \min_{\delta \ne 0, \left|\delta_{J_0^c}\right|_1 \le c_0 \left|\delta_{J_0}\right|_1} \frac{|X\delta|_2}{\sqrt{n}|\delta|_1} > 0. \tag{17}$$

Assumption RE$\tau_3(s, c_0)$. For some integer s such that $1 \le s \le M$ and a positive number c_0, the following condition holds:

$$\tau_3(s, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,M\}, |J_0| \le s} \min_{\delta \ne 0, \left|\delta_{J_0^c}\right|_1 \le c_0 \left|\delta_{J_0}\right|_1} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} > 0. \tag{18}$$

Note that $\tau_2(s, c_0) \le \tau_1(s, c_0) \le \kappa(s, c_0)$ and $\tau_2(s, c_0) \le \tau_3(s, c_0) \le \kappa(s, c_0)$ since $|\delta_{J_0}|_2 \le |\delta_{J_0}|_1 \le |\delta|_1$ and $|\delta_{J_0}|_2 \le |\delta|_2 \le |\delta|_1$. Moreover, it is easy to see that for fixed n, the four assumptions are equivalent, and those assumptions 1–5 by Bickel et al. [15] are all sufficient conditions for assumptions RE$\tau_1(s, c_0)$, RE$\tau_2(s, c_0)$, and RE$\tau_3(s, c_0)$.

In Section 4, we will see that RE$\tau_1(s, c_0)$ and RE$\tau_2(s, c_0)$ are all better than $\kappa(s, c_0)$ since they use $|\delta_{J_0^c}|_1 \le c_0 |\delta_{J_0}|_1$ and $|\delta_{J_0}|_1 \le |J_0|^{1/2} |\delta_{J_0}|_2$ as little as possible. Therefore, the inequalities given are more precise.

4. Comparisons with the Results by Bickel et al.

In the following, we give a bound of the prediction losses $\|\widehat{f}_L - f\|_n^2$ with respect to $\|\widehat{f}_D - f\|_n^2$ when the number of nonzero components of the Lasso or the Dantzig selector is small as compared to the sample size.

Theorem 1. *Let W_i be independent $N(0, \sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \ge 1$, $M \ge 2$. Let assumption RE(s, c_0) or RE$\tau_1(s, c_0)$ be satisfied with $1 \le s \le M$, where $c_0 > 0$, and let $\|f_j\|_n = 1$, $j = 1, \ldots, M$. Consider the Lasso estimator $\widehat{f}_L$ defined by* (6)-(7) *with*

$$r = A\sigma \sqrt{\frac{\log M}{n}}, \tag{19}$$

where $A > 2\sqrt{2}$, and consider the Dantzig estimator $\widehat{f}_D$ defined by (10) *with the same r. If $M(\widehat{\beta}_D) \le s$, then, with probability at least $1 - M^{1 - A^2/8}$, one has*

$$\begin{aligned} \left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|\widehat{f}_D - f\right\|_n^2 + 9A^2 \frac{\sigma^2}{n} \frac{\log M}{\tau_1^2(s, c_0)} \\ &\le \left\|\widehat{f}_D - f\right\|_n^2 + 9A^2 \frac{M(\widehat{\beta}_D)\sigma^2}{n} \frac{\log M}{\kappa^2(s, c_0)}. \end{aligned} \tag{20}$$

Proof. Set $\delta = \widehat{\beta}_L - \widehat{\beta}_D$. We apply (B.1) by Bickel et al. [15] with $\beta = \widehat{\beta}_D$, which yields that, with probability at least $1 - M^{1 - A^2/8}$,

$$\left\|\widehat{f}_L - f\right\|_n^2 \le \left\|\widehat{f}_D - f\right\|_n^2 + 4r\left|\delta_{J_0}\right|_1 - r|\delta|_1, \tag{21}$$

where $J_0 = J(\widehat{\beta}_D)$. From (B.16) by Bickel et al. [15], we have

$$\left\|\widehat{f}_L - f\right\|_n^2 \le \left\|\widehat{f}_D - f\right\|_n^2 + 3r|\delta|_1 - \frac{1}{n}|X\delta|_2^2. \tag{22}$$

Then,

$$\begin{aligned} \left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|\widehat{f}_D - f\right\|_n^2 + 3r\left|\delta_{J_0}\right|_1 - \frac{1}{4n}|X\delta|_2^2. \\ &\le \left\|\widehat{f}_D - f\right\|_n^2 + \left(3r\left|\delta_{J_0}\right|_1 \frac{\sqrt{n}}{|X\delta|_2}\right)^2 \\ &\le \left\|\widehat{f}_D - f\right\|_n^2 + \frac{9r^2}{\tau_1^2(s, c_0)} \\ &\le \left\|\widehat{f}_D - f\right\|_n^2 + \frac{9r^2 M(\widehat{\beta}_D)}{\kappa^2(s, c_0)}. \end{aligned} \tag{23}$$

□

Corollary 2. *Let the conditions of Theorem 1 hold, but with* $RE(s,5)$ *in place of* $RE(s,c_0)$. *If* $M(\widehat{\beta}_D) \le s$, *then, with probability at least* $1 - M^{1-A^2/8}$, *one has*

$$\left\|\widehat{f}_L - f\right\|_n^2 \le \left\|\widehat{f}_D - f\right\|_n^2 + 9A^2 \frac{M\left(\widehat{\beta}_D\right)\sigma^2}{n} \frac{\log M}{\kappa^2(s,5)}. \quad (24)$$

This corollary greatly improves Theorem 5.2 by Bickel et al. [15]. The right-hand side of the inequality of Theorem 5.2 is

$$10\left\|\widehat{f}_D - f\right\|_n^2 + 81A^2 \frac{M\left(\widehat{\beta}_D\right)\sigma^2}{n} \frac{\log M}{\kappa^2(s,5)}. \quad (25)$$

A general discussion of sparsity oracle inequalities can be found in Tsybakov [30]. Here, we prove a sparsity oracle inequality for the prediction loss of the Lasso estimators. Such inequalities have been recently obtained for the Lasso-type estimators in a number of settings, see [15, 23, 24, 27, 28].

Theorem 3. *Let* W_i *be independent* $N(0,\sigma^2)$ *random variables with* $\sigma^2 > 0$. *Fix integers* $n \ge 1$, $M \ge 2$, $1 \le s \le M$. *Let assumption* $RE(s,c_0)$ *or* $RE\tau_1(s,c_0)$ *be satisfied, where* $c_0 > 0$. *Consider the Lasso estimator* $\widehat{f}_L$ *defined by* (6)-(7) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}} \quad (26)$$

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

$$\begin{aligned}\left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|f_\beta - f\right\|_n^2 + \frac{9f_{\max}^2 A^2\sigma^2}{\tau_1^2(s,c_0)} \frac{\log M}{n} \\ &\le \left\|f_\beta - f\right\|_n^2 + \frac{9f_{\max}^2 A^2\sigma^2}{\kappa^2(s,c_0)} \frac{M(\beta)\log M}{n}.\end{aligned} \quad (27)$$

Proof. Fix an arbitrary $\beta \in \mathbb{R}^M$ with $M(\beta) \le s$. Set $\delta = D^{1/2}(\widehat{\beta}_L - \beta)$, $J_0 = J(\beta)$, where $D^{1/2} = \mathrm{diag}\{||f_1||_n, \ldots, ||f_M||_n\}$. On the event A in p1723 by Bickel et al. [15], we get, from the first line in (B.1) by Bickel et al. [15], that

$$\left\|\widehat{f}_L - f\right\|_n^2 \le \left\|f_\beta - f\right\|_n^2 + 4r\left|\delta_{J_0}\right|_1 - r|\delta|_1. \quad (28)$$

Since

$$\begin{aligned}&\left|f - X\widehat{\beta}_L\right|_2^2 \\ &\quad = |f - X\beta|_2^2 - 2\delta^T D^{-1/2} X^T\left(f - X\widehat{\beta}_L\right) - \left|XD^{-1/2}\delta\right|_2^2,\end{aligned} \quad (29)$$

then

$$\begin{aligned}\left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|f_\beta - f\right\|_n^2 \\ &\quad + \frac{2}{n}|\delta|_1\left|D^{-1/2}X^T(f - X\widehat{\beta}_L)\right|_\infty - \frac{1}{n}\left|XD^{-1/2}\delta\right|_2^2.\end{aligned} \quad (30)$$

From (8) and (B.5) by Bickel et al. [15], we have

$$\begin{aligned}&\frac{1}{n}\left|D^{-1/2}X^T(f - X\widehat{\beta}_L)\right|_\infty \le \frac{3r}{2}, \\ &\frac{1}{n}\left|XD^{-1/2}\delta\right|_2^2 = \frac{1}{n}\delta^T D^{-1/2}X^T XD^{-1/2}\delta \ge \frac{1}{nf_{\max}^2}|X\delta|_2^2.\end{aligned} \quad (31)$$

Thus,

$$\begin{aligned}\left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|f_\beta - f\right\|_n^2 + 3r|\delta|_1 - \frac{1}{n}\left|XD^{-1/2}\delta\right|_2^2 \\ &\le \left\|f_\beta - f\right\|_n^2 + 3r|\delta|_1 - \frac{1}{nf_{\max}^2}|X\delta|_2^2.\end{aligned} \quad (32)$$

From (28) and (32), we have

$$\begin{aligned}\left\|\widehat{f}_L - f\right\|_n^2 &\le \left\|f_\beta - f\right\|_n^2 + 3r\left|\delta_{J_0}\right|_1 - \frac{1}{4nf_{\max}^2}|X\delta|_2^2 \\ &\le \left\|f_\beta - f\right\|_n^2 + \left(3r\left|\delta_{J_0}\right|_1 \frac{\sqrt{n}f_{\max}}{|X\delta|_2}\right)^2 \\ &\le \left\|f_\beta - f\right\|_n^2 + \frac{9r^2 f_{\max}^2}{\tau_1^2(s,c_0)} \\ &\le \left\|f_\beta - f\right\|_n^2 + \frac{9r^2 f_{\max}^2}{\kappa^2(s,c_0)} M(\beta).\end{aligned} \quad (33)$$

□

Corollary 4. *Let the conditions of Theorem 3 hold, but with* $RE(s,3+4/\varepsilon)$ *in place of* $RE(s,c_0)$. *Then, with probability at least* $1 - M^{1-A^2/8}$, *one has*

$$\left\|\widehat{f}_L - f\right\|_n^2 \le \left\|f_\beta - f\right\|_n^2 + \frac{9f_{\max}^2 A^2\sigma^2}{\kappa^2(s,3+4/\varepsilon)} \frac{M(\beta)\log M}{n}. \quad (34)$$

This corollary greatly improves Theorem 6.1 by Bickel et al. [15]. The right-hand side of the inequality of Theorem 6.1 is

$$\begin{aligned}(1+\varepsilon) \inf_{\substack{\beta\in\mathbb{R}^M, \\ M(\beta)\le s}} \Bigg\{&\left\|f_\beta - f\right\|_n^2 \\ &+ \frac{C(\varepsilon) f_{\max}^2 A^2\sigma^2}{\kappa^2(s,3+4/\varepsilon)} \frac{M(\beta)\log M}{n}\Bigg\},\end{aligned} \quad (35)$$

where $(1+\varepsilon)C(\varepsilon) = 4(\varepsilon+2)^2/\varepsilon > 9$.

In the following, we assume that the vector of observations $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ is of the form

$$\mathbf{y} = X\beta^* + \mathbf{w}, \quad (36)$$

where X is an $n \times M$ deterministic matrix, $\beta^* \in \mathbb{R}^M$, and $\mathbf{w} = (W_1, \ldots, W_n)^T$. 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 (36) is satisfied for $\beta^* = \beta_0$, then there exists an affine space $U = \{\beta^* : X\beta^* = X\beta_0\}$ of

vectors satisfying (36). The Lasso estimator of β^* in (36) is defined by

$$\widehat{\beta}_L = \arg\min_{\beta\in\mathbb{R}^M} \left\{ \frac{1}{n}\left|\mathbf{y} - X\beta\right|_2^2 + 2r\left|\beta\right|_1 \right\}. \tag{37}$$

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

$$\begin{aligned} \left\|f_\beta\right\|_n^2 &= \frac{1}{n}\left|X\beta\right|_2^2, \\ \left\|f_\beta - f\right\|_n^2 &= \frac{1}{n}\left|X(\beta-\beta^*)\right|_2^2, \\ \left\|\widehat{f}_L - f\right\|_n^2 &= \frac{1}{n}\left|X(\widehat{\beta}_L-\beta^*)\right|_2^2. \end{aligned} \tag{38}$$

Theorem 5. *Let W_i be independent $N(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Let all the diagonal elements of the matrix X^TX/n be equal to 1, and let $M(\beta^*) \le s$, where $1 \le s \le M$, $n \ge 1$, $M \ge 2$. Let assumption $RE(s,c_0)$ or $RE\tau_1(s,c_0)$ be satisfied, where $c_0 > 0$. Consider the Lasso estimator $\widehat{\beta}_L$ defined by* (37) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}} \tag{39}$$

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

$$\left|\widehat{\beta}_L - \beta^*\right|_1 \le \frac{4A}{\tau_1^2(s,c_0)}\sigma\sqrt{\frac{\log M}{n}} \le \frac{4A}{\kappa^2(s,c_0)}\sigma s\sqrt{\frac{\log M}{n}}, \tag{40}$$

$$\begin{aligned} &\left|X(\widehat{\beta}_L - \beta^*)\right|_2^2 \\ &\quad\le \frac{144A^2}{25\tau_1^2(s,c_0)}\sigma^2\log M \le \frac{144A^2}{25\kappa^2(s,c_0)}\sigma^2 s\ \log M, \end{aligned} \tag{41}$$

$$M\left(\widehat{\beta}_L\right) \le \frac{576\phi_{\max}}{25\tau_1^2(s,c_0)} \le \frac{576\phi_{\max}}{25\kappa^2(s,c_0)}s. \tag{42}$$

Proof. Set $\delta = \widehat{\beta}_L - \beta^*$ and $J_0 = J(\beta^*)$. Using (B.1) and (B.2) by Bickel et al. [15], where we put $\beta = \beta^*$, $r_{n,j} \equiv r$, and $\|f_\beta - f\|_n = 0$, we get that, on the event A (i.e., with probability at least $1 - M^{1-A^2/8}$),

$$\frac{1}{n}\left|X\delta\right|_2^2 \le 4r\left|\delta_{J_0}\right|_1 - r\left|\delta\right|_1. \tag{43}$$

From (B.16) by Bickel et al. [15], we have

$$\frac{2}{n}\left|X\delta\right|_2^2 \le 3r\left|\delta\right|_1. \tag{44}$$

Then,

$$\frac{1}{n}\left|X\delta\right|_2^2 \le \frac{12}{5}r\left|\delta_{J_0}\right|_1. \tag{45}$$

By assumption $\mathrm{RE}(s,c_0)$ or $\mathrm{RE}\tau_1(s,c_0)$, we obtain that, on A,

$$\tau_1^2(s,c_0)\left|\delta_{J_0}\right|_1^2 \le \frac{1}{n}\left|X\delta\right|_2^2 \le \frac{12}{5}r\left|\delta_{J_0}\right|_1. \tag{46}$$

Thus,

$$\begin{aligned} \left|\delta_{J_0}\right|_2 \le \left|\delta_{J_0}\right|_1 &\le \frac{12}{5\tau_1^2(s,c_0)}r \le \frac{12}{5\kappa^2(s,c_0)}r\sqrt{s}, \\ \frac{1}{n}\left|X\delta\right|_2^2 &\le \frac{144}{25\tau_1^2(s,c_0)}r^2 \le \frac{144}{25\kappa^2(s,c_0)}r^2 s. \end{aligned} \tag{47}$$

From (43), we have

$$\begin{aligned} r\left|\delta\right|_1 &\le 4r\left|\delta_{J_0}\right|_1 - \frac{1}{n}\left|X\delta\right|_2^2 \\ &\le \left(\frac{2r\left|\delta_{J_0}\right|_1\sqrt{n}}{\left|X\delta\right|_2}\right)^2 \le \frac{4r^2}{\tau_1^2(s,c_0)} \le \frac{4sr^2}{\kappa^2(s,c_0)}. \end{aligned} \tag{48}$$

Thus,

$$\left|\delta\right|_1 \le \frac{4r}{\tau_1^2(s,c_0)} \le \frac{4sr}{\kappa^2(s,c_0)}. \tag{49}$$

The inequalities (49) and (47) coincide with (40) and (41), respectively. Next, (42) follows immediately from (B.3) in Bickel et al. [15] and (41). □

Corollary 6. *Let the conditions of Theorem 5 hold, but with $RE(s,3)$ in place of $RE(s,c_0)$. Then, with probability at least $1 - M^{1-A^2/8}$, one has*

$$\left|\widehat{\beta}_L - \beta^*\right|_1 \le \frac{4A}{\kappa^2(s,3)}\sigma s\sqrt{\frac{\log M}{n}}, \tag{50}$$

$$\left|X(\widehat{\beta}_L - \beta^*)\right|_2^2 \le \frac{144A^2}{25\kappa^2(s,3)}\sigma^2 s\ \log M, \tag{51}$$

$$M\left(\widehat{\beta}_L\right) \le \frac{576\phi_{\max}}{25\kappa^2(s,c_0)}s. \tag{52}$$

This corollary improves Theorem 7.2 by Bickel et al. [15]. The right-hand sides of the inequalities of Theorem 7.2 are

$$\begin{aligned} &\frac{16A}{\kappa^2(s,3)}\sigma s\sqrt{\frac{\log M}{n}}, \qquad \frac{16A^2}{\kappa^2(s,3)}\sigma^2 s\ \log M, \\ &\qquad\qquad \frac{64\phi_{\max}}{\kappa^2(s,3)}s, \end{aligned} \tag{53}$$

respectively. That is, they are 4, 25/9, and 25/9 times as large as (50)–(52), respectively.

5. Conclusions

We point out that $|\delta_{J_0^c}|_1 \le c_0|\delta_{J_0}|_1$ with $c_0 = 1$ by Candes and Tao [13] and $c_0 = 3$ by Bickel et al. [15] are only the sufficient condition of Lasso and Dantzig selector. Their role should not be overemphasized. That is, $|\delta_{J_0^c}|_1 \le c_0|\delta_{J_0}|_1$ should not be deliberately used in any case for solving inequality. We should use $|\delta_{J_0^c}|_1 \le c_0|\delta_{J_0}|_1$ as little as possible when proving inequalities.

In fact, the corresponding results have been enlarged due to the use of $|\delta_{J_0^c}|_1 \leq c_0|\delta_{J_0}|_1$ when solving the problems of Lasso and Dantzig selector. When proving sparsity oracle inequalities for the prediction loss and bounds on the L_p estimation loss, using again $|\delta_{J_0^c}|_1 \leq c_0|\delta_{J_0}|_1$ must be to enlarge the inequalities again and to result in reduced accuracy.

We have seen that $\mathrm{RE}\tau_1(s, c_0)$ and $\mathrm{RE}\tau_2(s, c_0)$ are all much better than $\kappa(s, c_0)$ since in $\mathrm{RE}\tau_1(s, c_0)$ and $\mathrm{RE}\tau_2(s, c_0)$ the use of $|\delta_{J_0^c}|_1 \leq c_0|\delta_{J_0}|_1$ and $|\delta_{J_0}|_1 \leq |J_0|^{1/2}|\delta_{J_0}|_2$ is less. Therefore, the inequalities given are more precise.

References

[1] M. Rosenbaum and A. B. Tsybakov, "Sparse recovery under matrix uncertainty," *The Annals of Statistics*, vol. 38, no. 5, pp. 2620–2651, 2010.

[2] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[4] W. Fu and K. Knight, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[5] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[6] N. Meinshausen, G. Rocha, and B. Yu, "Discussion: a tale of three cousins: Lasso, L_2Boosting and Dantzig," *The Annals of Statistics*, vol. 35, no. 6, pp. 2373–2384, 2007.

[7] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[8] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the LASSO and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, 2000.

[9] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, 2000.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] C. H. Zhang and J. Huang, "The sparsity and bias of the LASSO selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[12] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[13] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[14] P. J. Bickel, "Discussion: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2352–2357, 2007.

[15] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[16] T. T. Cai and J. Lv, "Discussion: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2365–2369, 2007.

[17] E. Candès and T. Tao, "Rejoinder: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2392–2404, 2007.

[18] B. Efron, T. Hastie, and R. Tibshirani, "Discussion: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2358–2364, 2007.

[19] M. P. Friedlander and M. A. Saunders, "Discussion: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2385–2391, 2007.

[20] Y. Ritov, "Discussion: the Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2370–2372, 2007.

[21] F. Ye and C. H. Zhang, "Rate minimaxity of the Lasso and Dantzig selector for the L_q loss in L_r balls," *Journal of Machine Learning Research*, vol. 11, pp. 3519–3540, 2010.

[22] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[23] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Aggregation for Gaussian regression," *The Annals of Statistics*, vol. 35, no. 4, pp. 1674–1697, 2007.

[24] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Sparsity oracle inequalities for the Lasso," *Electronic Journal of Statistics*, vol. 1, pp. 169–194, 2007.

[25] E. Greenshtein and Y. Ritov, "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, vol. 10, no. 6, pp. 971–988, 2004.

[26] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric regression," *The Annals of Statistics*, vol. 28, no. 3, pp. 681–712, 2000.

[27] V. Koltchinskii, "Sparsity in penalized empirical risk minimization," *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, vol. 45, no. 1, pp. 7–57, 2009.

[28] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[29] V. Koltchinskii, "The Dantzig selector and sparsity oracle inequalities," *Bernoulli*, vol. 15, no. 3, pp. 799–828, 2009.

[30] A. B. Tsybakov, "Discussion of "regularization in statistics" by P. Bickel and B. Li," *Test*, vol. 15, no. 2, pp. 303–310, 2006.