Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 946241, 7 pages
http://dx.doi.org/10.1155/2014/946241

Research Article

Weaker Regularity Conditions and Sparse Recovery in High-Dimensional Regression

Shiqing Wang,[1] Yan Shi,[2] and Limin Su[1]

[1] *College of Mathematics and Information Sciences, North China University of Water Resources and Electric Power, Zhengzhou 450045, China*
[2] *Institute of Environmental and Municipal Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450045, China*

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

Received 27 October 2013; Accepted 7 July 2014; Published 17 July 2014

Academic Editor: Yuesheng Xu

Regularity conditions play a pivotal role for sparse recovery in high-dimensional regression. In this paper, we present a weaker regularity condition and further discuss the relationships with other regularity conditions, such as restricted eigenvalue condition. We study the behavior of our new condition for design matrices with independent random columns uniformly drawn on the unit sphere. Moreover, the present paper shows that, under a sparsity scenario, the Lasso estimator and Dantzig selector exhibit similar behavior. Based on both methods, we derive, in parallel, more precise bounds for the estimation loss and the prediction risk in the linear regression model when the number of variables can be much larger than the sample size.

1. Introduction

In the recent years, the problems of statistical inference in high-dimensional setting, in which the dimension of the data p exceeds the sample size n, have attracted a great deal of attention. One concrete instance of a high-dimensional inference problem concerns the standard linear regression model:

$$y = X\beta + W, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is called the design matrix, $\beta \in \mathbb{R}^p$ is an unknown target vector, and $W \in \mathbb{R}^n$ is a stochastic error term, in which the goal is to estimate a vector $\beta \in \mathbb{R}^p$ based on response y and the vector of covariates $X = (X_1, \ldots, X_p)$. In the setting $p \gg n$, the classical linear regression model is unidentifiable, so that it is not meaningful to estimate the parameter vector $\beta \in \mathbb{R}^p$.

However, many high-dimensional regression problems exhibit special structure that can lead to an identifiable model. In particular, sparsity in the regression vector β is an archetypal example of such structure; that is, only a few components of β are different from zero, say s-sparsity; β is then said to be s-sparsity, and there has been a great interest in the study of this problem recently. The use of the ℓ_1-norm penalty to enforce sparsity has been very successful and there have been several methods, such as the Lasso [1] or basis pursuit [2], and the Dantzig selector [3]. Sparsity has also been exploited in a number of other questions, for instance, instrumental variable regression in the presence of endogeneity [4].

There is now a well-developed theory on what conditions are required on the design matrix $X \in \mathbb{R}^{n\times p}$ for such ℓ_1-based relaxations to reliably estimate β; for example, see [5–16]. The restricted eigenvalue (RE) condition due to Bickel et al. [10] is a weaker one of the conditions mentioned above. Wang and Su [7, 13] presented some equivalent conditions with them, respectively, and there is also a large body of work in the high-dimensional setting; for example, see [3, 6, 12, 17–19], which showed a uniform uncertainty principle (UUP, a condition that is stronger than the RE condition; see [10, 20]). In this paper, we consider a restricted eigenvalue condition that is weaker than the RE conditions in [7, 10, 13] under certain setting.

Thus, in the setting of high-dimensional linear regression, the interesting question is accurately estimating the regression vector β and the response $X\beta$ from few and corrupted

observations. In the standard form, under assumptions on the matrix X and with high probability, the estimation bounds are of the form $C\|\beta\|_0(\log(p)/n)^{q/2}$ (e.g., see [7, 8, 13, 21]), and the prediction errors are bounded by $C\log(p)\|\beta\|_0$ (e.g., see [1, 7, 21]), where C is a positive constant.

The main contribution of this paper is the following: we present a restricted eigenvalue assumption that is weaker than the RE conditions in previous paper under certain setting. Using the ℓ_1-norm penalty, our results are more precise than the existing ones. There is an open question that is finding a weaker assumption and obtaining better results no matter under what circumstances.

The remainder of this paper is organized as follows. We begin in Section 2 with some notations and definitions. In Section 3, we introduce some assumptions and discuss the relation between our assumptions and the existing ones. Section 4 contains our main results, and we also show the approximate equivalence between the Lasso and the Dantzig selector. We give three lemmas and the proofs of the theorems in Section 5.

2. Preliminaries

In this section, we introduce some notations and definitions.

Let a vector $\beta \in \mathbb{R}^p$. We denote by

$$M(\beta) = \sum_{j=1}^{p} I_{\{\beta_j \neq 0\}} = |J(\beta)| \tag{2}$$

the number of nonzero coordinates of β, where $I_{\{\cdot\}}$ denotes the indicator function

$$J(\beta) = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\} \tag{3}$$

and $|J|$ the cardinality of J. We use the standard notation

$$\|\beta\|_q = \left(\sum_{i=1}^{p} |\beta_i|^q\right)^{1/q} \tag{4}$$

to stand for the ℓ_q-norm of the vector of β. Moreover, a vector β is said to be k-sparse if $\|\beta\|_0 \leq k$; that is, it has at most k nonzero entries. For a vector $\Delta \in \mathbb{R}^p$ and a subset $J \subset \{1, \ldots, p\}$, we denote by Δ_J the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on J and zero coordinates on the complement J^c of J.

For linear regression model (1), regularized estimation with the ℓ_1-norm penalty, also known as the Lasso [1] or the basis pursuit [2], refers to the following convex optimization problem:

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{\frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1\right\}, \tag{5}$$

where $\lambda > 0$ is a penalization parameter. The Dantzig selector has been introduced by Candes and Tao [3] as

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{subject to } \left\|X^T(y - X\beta)\right\|_\infty \leq \lambda, \tag{6}$$

where $\lambda > 0$ is a tuning parameter. It is known that it can be recast as a linear program. Hence, it is also computationally tractable.

For an integer $1 \leq s \leq p/2$ and s-sparse vector $\beta \in \mathbb{R}^p$, let $\beta_{J_0} \in \mathbb{R}^{|J_0|}$ be a subvector of $\beta \in \mathbb{R}^p$ confined to J_0. One of the common properties of the Lasso and the Dantzig selector is that, for an appropriately chosen λ and a vector $\delta = \widehat{\beta} - \beta$, where $\widehat{\beta}$ is the solution from either the Lasso or the Dantzig selector, it holds with high probability (cf. Lemmas 11 and 12):

$$\left\|\delta_{J_0^c}\right\|_1 \leq c_0\left\|\delta_{J_0}\right\|_1, \tag{7}$$

with $c_0 = 1$ for the Dantzig selector by Candes and Tao [3] and with $c_0 = 3$ for the Lasso by Bickel et al. [9], where $c_0 > 0$ and

$$J_0 = J(\beta) \subset \{1, 2, \ldots, p\} \tag{8}$$

is the set of nonzero coefficients of the true parameter β of the model.

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

$$\Psi_n = \frac{1}{n}X^T X, \tag{9}$$

where X is the designed matrix in model (1) and $X^T \in \mathbb{R}^{p \times n}$ denotes the transpose matrix of X.

3. Discussion of the Assumption

Under the sparsity scenario, we are typically interested in the case where $p > n$, and even $p \gg n$. 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 does not work in this case, since it requires positive definiteness of Ψ_n. That is,

$$\min_{\delta \in \mathbb{R}^p, \delta \neq 0} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta\|_2} > 0. \tag{10}$$

It turns out that the Lasso and Dantzig selector require much weaker assumptions. The idea by Bickel et al. [10] is that the minimum in (10) 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. Note that the role of (7) is to restrict set of vectors $\{\delta \in \mathbb{R}^p : \delta \neq 0\}$ to

$$\left\{\delta \in \mathbb{R}^p : \delta \neq 0, \left\|\delta_{J_0^c}\right\|_1 \leq c_0\left\|\delta_{J_0}\right\|_1\right\}. \tag{11}$$

Assumption 1 (RE(s, c_0) (Bickel et al. [10])). For some integer s such that $1 \leq s \leq p$ and a positive number c_0, the following condition holds:

$$\kappa(s, c_0) \triangleq \min_{J_0 \subseteq \{1,2,\ldots,p\}, |J_0| \leq s} \min_{\delta \neq 0, \left\|\delta_{J_0^c}\right\|_1 \leq c_0\left\|\delta_{J_0}\right\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\left\|\delta_{J_0}\right\|_2} > 0. \tag{12}$$

Bickel et al. [10] showed that the bounds of estimation error and prediction error are $C\|\beta\|_0(\log(p)/n)^{q/2}$ and

$C\log(p)\|\beta\|_0$, respectively, for both the Lasso and Dantzig selector, where C is a positive constant and $\|\beta\|_0$ is the sparsity level. Next, we describe the RE$\tau_2(s,c_0)$ assumption presented by Wang and Su [7], which is obtained by replacing $\|\delta\|_2$ by its upper bound $\|\delta\|_1$ in (10).

Assumption 2 (RE$\tau_2(s,c_0)$ (Wang and Su [7])). For some integer s such that $1 \le s \le p$ and a positive number c_0, the following condition holds:

$$\tau_2(s,c_0) \triangleq \min_{J_0\subseteq\{1,2,\ldots,p\},|J_0|\le s\delta\ne 0,\left\|\delta^c_{J_0}\right\|_1\le c_0\left\|\delta_{J_0}\right\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta\|_1} > 0. \quad (13)$$

The two conditions are very similar. The only difference is the ℓ_1- versus ℓ_2-norm of a part of δ in the denominator. The RE$\tau_2(s,c_0)$ condition is equivalent to RE(s,c_0); see [7, 13] for the discussion on equivalence. The results of [7, 13] are more precise for the bounds of estimation and prediction than those derived in Bickel et al. [10] and do not lie on the sparsity level $\|\beta\|_0$.

In order to obtain our regularity condition in this paper, we decompose δ into a set of vectors $\delta_{S_0}, \delta_{S_1}, \delta_{S_2}, \ldots, \delta_{S_K}$, such that S_0 corresponds to locations of the s largest coefficient of δ in absolute values, S_1 corresponds to locations of the next s largest coefficient of $\delta_{S_0^c}$ in absolute values, and so on. Hence, we have $S_0^c = \bigcup_{k=1}^{K} S_k$, where $K \ge 1$, $|S_k| = s$, for all $k = 1,\ldots,K-1$, and $|S_K| \le s$.

Now for each $j \ge 1$, we have

$$\left\|\delta_{S_j}\right\|_2 \le \sqrt{s}\left\|\delta_{S_j}\right\|_\infty \le \frac{1}{\sqrt{s}}\left\|\delta_{S_{j-1}}\right\|_1, \quad (14)$$

where vector $\|\cdot\|_\infty$ represents the largest entry in absolute value in the vector, and hence

$$\begin{aligned}\left\|\delta_{S_0^c}\right\|_2 &\le \sum_{k\ge 1}\left\|\delta_{S_k}\right\|_2 \\ &\le s^{-1/2}\left(\left\|\delta_{S_0}\right\|_1 + \left\|\delta_{S_1}\right\|_1 + \left\|\delta_{S_2}\right\|_1 + \cdots\right) \\ &\le s^{-1/2}\left(\left\|\delta_{S_0}\right\|_1 + \left\|\delta_{S_0^c}\right\|_1\right) = s^{-1/2}\|\delta\|_1.\end{aligned} \quad (15)$$

Replacing $\|\delta\|_1$ by $\sqrt{s}\|\delta_{S_0^c}\|_2$ in (13), we get the following assumption.

Assumption 3 (LR$\varphi_1(s,c_0)$). For some integer s such that $1 \le s \le p$ and a positive number c_0, the following condition holds:

$$\varphi_1(s,c_0) \triangleq \min_{J_0\subseteq\{1,2,\ldots,p\},|J_0|\le s\delta\ne 0,\left\|\delta^c_{J_0}\right\|_1\le c_0\left\|\delta_{J_0}\right\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\sqrt{s}\left\|\delta_{J_0^c}\right\|_2} > 0. \quad (16)$$

The inequality $\sqrt{s}\|\delta_{S_0^c}\|_2 \le \|\delta\|_1$ immediately implies that the assumption LR$\varphi_1(s,c_0)$ is weaker than the assumptions RE$\tau_2(s,c_0)$ and RE(s,c_0). Noting the norm $\|\delta_{J_0^c}\|_2$ in the denominator of (16), it makes the proof become more complicated. We need an equivalent condition of LR$\varphi_1(s,c_0)$ for the sake of simplicity, as similarly discussed on equivalence (cf. [7, 13]).

Assumption 4 (LR$\varphi_2(s,c_0)$). For some integer s such that $1 \le s \le p$ and a positive number c_0, the following condition holds:

$$\varphi_2(s,c_0) \triangleq \min_{J_0\subseteq\{1,2,\ldots,p\},|J_0|\le s\delta\ne 0,\left\|\delta^c_{J_0}\right\|_1\le c_0\left\|\delta_{J_0}\right\|_1} \frac{\|X\delta\|_2}{\sqrt{s}\sqrt{n}\left\|\delta_{J_0}\right\|_1} > 0. \quad (17)$$

The two conditions above can be used to solve all the problems of sparse recovery in high-dimensional regression. Due to technical reasons, we only give the results when the LR$\varphi_2(s,c_0)$ is satisfied.

4. Main Results of Sparse Recovery for Regression Model

In order to provide performance guarantees for ℓ_1-norm penalty applied to sparse linear models, it is sufficient to assume that the regularity conditions are satisfied. In this section, we show main results when the LR$\varphi_2(s,c_0)$ is satisfied. In particular, for convenience, we assume that all the diagonal elements of the matrix X^TX/n are equal to 1.

We firstly prove a type of approximate equivalence between the Lasso and the Dantzig selector. Similar results on equivalence can be found in [7, 10, 13]. It is expressed as closeness of the prediction losses $\|X\beta - X\beta_D\|_2^2$ and $\|X\beta - X\beta_L\|_2^2$ when the number of nonzero components of the Lasso or the Dantzig selector is small as compared to the sample size.

Theorem 5. *For linear model* (1)*, let* $W_i \sim N(0,\sigma^2)$ *be independent random variables with* $\sigma^2 > 0$*. Consider the Lasso estimator* $\widehat{\beta}_L$ *and Dantzig estimator* $\widehat{\beta}_D$ *defined by* (5) *and* (6) *with the same* λ*. If LR*$\varphi_2(s,c_0)$ *is satisfied, where* $c_0 > 0$*, then, with probability of at least* $1 - p^{1-A^2/8}$*, one has*

$$\left|\frac{1}{n}\|X\beta - X\beta_D\|_2^2 - \frac{1}{n}\|X\beta - X\beta_L\|_2^2\right| \le \frac{16\lambda^2}{s\varphi_2^2}. \quad (18)$$

Next, we get the bounds on the rate of convergence of Lasso and Dantzig selector.

Theorem 6. *For linear model* (1)*, let* $W_i \sim N(0,\sigma^2)$ *be independent random variables with* $\sigma^2 > 0$*. Consider the Lasso estimator* $\widehat{\beta}_L$ *defined by* (5) *with* $\lambda > 2\lambda_0 > 0$*. If LR*$\varphi_2(s,3)$ *is satisfied, where* $c_0 > 0$*, then, with probability of at least* $1 - p\exp(2\lambda^2 n/\sigma^2)$*, one has*

$$\left\|\widehat{\beta}_L - \beta\right\|_1 \le \frac{4\lambda}{s\varphi_2^2(s,c_0)}, \quad (19)$$

$$\left\|\widehat{\beta}_L - \beta\right\|_2 \le \frac{4\lambda}{s\varphi_2^2}\left(\frac{3}{5} + \frac{1}{\sqrt{s}}\right), \quad (20)$$

$$\frac{1}{n}\left\|X\left(\widehat{\beta}_L - \beta\right)\right\|_2^2 \le \frac{144\lambda^2}{25s\varphi_2^2(s,c_0)}, \quad (21)$$

where $\varphi_2 = \varphi_2(s,3)$.

Theorem 7. *For linear model* (1)*, let* $W_i \sim N(0,\sigma^2)$ *be independent random variables with* $\sigma^2 > 0$*. Consider the*

Dantzig selector $\widehat{\beta}_D$ *defined by* (6) *with* $\lambda > \lambda_0 > 0$. *If* $LR\varphi_2(s,1)$ *is satisfied, where* $c_0 > 0$, *then, with probability of at least* $1 - p\exp(\lambda^2 n/2\sigma^2)$, *one has*

$$\left\|\widehat{\beta}_D - \beta\right\|_1 \le \frac{8\lambda}{s\varphi_2^2}, \tag{22}$$

$$\left\|\widehat{\beta}_D - \beta\right\|_2 \le \frac{4\lambda}{s\varphi_2^2}\left(1 + \frac{2}{\sqrt{s}}\right), \tag{23}$$

$$\frac{1}{n}\left\|X\left(\widehat{\beta}_D - \beta\right)\right\|_2^2 \le \frac{16\lambda^2}{s\varphi_2^2}, \tag{24}$$

where $\varphi_2 = \varphi_2(s,1)$.

Remark 8. We have no conditions on the parameter λ. As in [10], we can rewrite λ in terms of another parameter A in order to clarify the notation:

$$\lambda = A\sigma\sqrt{\frac{\log p}{n}}, \quad A > \sqrt{2}. \tag{25}$$

Then, the results of Theorems 5–7 are as follows:

$$\left|\frac{1}{n}\left\|X\beta - X\beta_D\right\|_2^2 - \frac{1}{n}\left\|X\beta - X\beta_L\right\|_2^2\right| \le \frac{16A^2\sigma^2\log p}{n\varphi_2^2 s}, \tag{26}$$

$$\left\|\widehat{\beta}_L - \beta\right\|_1 \le \frac{4A\sigma}{s\varphi_2^2(s,3)}\sqrt{\frac{\log p}{n}}, \tag{9′}$$

$$\left\|\widehat{\beta}_L - \beta\right\|_2 \le \frac{4A\sigma}{s\varphi_2^2}\left(\frac{3}{5} + \frac{1}{\sqrt{s}}\right)\sqrt{\frac{\log p}{n}}, \tag{10′}$$

$$\frac{1}{n}\left\|X\left(\widehat{\beta}_L - \beta\right)\right\|_2^2 \le \frac{144A^2}{25\varphi_2^2(s,3)}\frac{\sigma^2\log p}{sn}, \tag{11′}$$

$$\left\|\widehat{\beta}_D - \beta\right\|_1 \le \frac{8A\sigma}{s\varphi_2^2(s,1)}\sqrt{\frac{\log p}{n}}, \tag{12′}$$

$$\left\|\widehat{\beta}_D - \beta\right\|_2 \le \frac{4A\sigma}{s\varphi_2^2}\left(1 + \frac{2}{\sqrt{s}}\right)\sqrt{\frac{\log p}{n}}, \tag{13′}$$

$$\frac{1}{n}\left\|X\left(\widehat{\beta}_D - \beta\right)\right\|_2^2 \le \frac{16A^2}{\varphi_2^2(s,1)}\frac{\sigma^2\log p}{sn}. \tag{14′}$$

The results of Theorems 7.1 and 7.2 in Bickel et al. [10] are

$$\begin{aligned}
&\left\|\widehat{\beta}_L - \beta\right\|_1 \le \frac{4A\sigma s}{\varphi_2^2(s,3)}\sqrt{\frac{\log p}{n}},\\
&\frac{1}{n}\left\|X\left(\widehat{\beta}_L - \beta\right)\right\|_2^2 \le \frac{144A^2}{25\varphi_2^2(s,3)}\frac{\sigma^2 s\log p}{n},\\
&\left\|\widehat{\beta}_D - \beta\right\|_1 \le \frac{8A\sigma s}{\varphi_2^2(s,1)}\sqrt{\frac{\log p}{n}},\\
&\frac{1}{n}\left\|X\left(\widehat{\beta}_D - \beta\right)\right\|_2^2 \le \frac{16A^2}{\varphi_2^2(s,1)}\frac{\sigma^2 s\log p}{n}.
\end{aligned} \tag{27}$$

Comparing the results above, our results greatly improve those in Bickel et al. [10].

Additionally, the similar results for Lasso can be found in Wang and Su [7]. They are

$$\begin{aligned}
&\left\|\widehat{\beta}_L - \beta\right\|_1 \le \frac{4A}{\tau_1^2(s,1)}\sigma\sqrt{\frac{\log p}{n}},\\
&\frac{1}{n}\left\|X\left(\widehat{\beta}_L - \beta\right)\right\|_2^2 \le \frac{144A^2}{25\tau_1^2(s,1)}\frac{\sigma^2\log p}{n}.
\end{aligned} \tag{28}$$

It is clear that our results are more precise than those in the existing results, for example, [7, 10].

Remark 9. The assumptions $LR\varphi_1(s,c_0)$ and $LR\varphi_2(s,c_0)$ are weaker than assumptions $RE\tau_2(s,c_0)$ and $RE(s,c_0)$, since $\sqrt{s}\|\delta_{S_0^c}\|_2 \le \|\delta\|_1$. Note that the inequality $\sqrt{s}\|\delta_{S_0^c}\|_2 \le \|\delta\|_1$ holds under the setting discussed in Section 3. That is, our weaker assumptions hold under certain condition, but they cannot be considered to be better than those in previous paper at any time.

5. Lemmas and the Proofs of the Results

In this section, we give three lemmas and the proofs of the theorems.

Lemma 10. *Let* $W_i \sim N(0,\sigma^2)$ *be independent random variables with* $\sigma^2 > 0$. *Then, for any* $\lambda_0 > 0$,

$$P\left(\frac{1}{n}\left|X^T W\right|_\infty \ge \lambda_0\right) \le p\exp\left(\frac{\lambda_0^2 n}{2\sigma^2}\right). \tag{29}$$

Proof. Since $W_i \sim N(0,\sigma^2)$, it immediately follows that

$$\begin{aligned}
&P\left(\frac{1}{n}\left|X^T W\right|_\infty \ge \lambda_0\right)\\
&\le \sum_j P\left(\frac{1}{n}\left|\sum_i x_{i,j}w_i\right| \ge \lambda_0\right)\\
&\le p\sum_j P\left(\left|n^{-1/2}\sigma^{-1}\sum_i x_{i,j}w_i\right| \ge \frac{\lambda_0\sqrt{n}}{(\sigma)}\right)\\
&\le p\sum_j P\left(|\eta| \ge \frac{\lambda_0\sqrt{n}}{(\sigma)}\right),\\
&\le p\exp\left(\frac{\lambda_0^2 n}{2\sigma^2}\right),
\end{aligned} \tag{30}$$

where $\eta \sim N(0,1)$. □

Lemma 11. *Let* $W_i \sim N(0,\sigma^2)$ *be independent random variables with* $\sigma^2 > 0$. *Let* $\widehat{\beta}_L$ *be the Lasso estimator defined*

by (5). *Then, with probability of at least* $1 - p\exp(2\lambda^2 n/\sigma^2)$, *one has, simultaneously for all* $\beta \in \mathbb{R}^p$ *and* $\lambda > 2\lambda_0$,

$$\frac{1}{n}\left\|X\widehat{\beta}_L - X\beta\right\|_2^2 + \lambda\sum_{j=1}^{p}\left|\beta_j - \widehat{\beta}_{L,j}\right| \le 4\lambda\sum_{j\in J(\beta)}\left|\beta_j - \widehat{\beta}_{L,j}\right|, \tag{31}$$

$$\left\|\frac{1}{n}X^T X\left(\beta - \widehat{\beta}_L\right)\right\|_\infty \le \frac{3\lambda}{2}. \tag{32}$$

Proof. By the definition of $\widehat{\beta}_L$,

$$\frac{1}{n}\left\|Y - X\widehat{\beta}_L\right\|_2^2 + 2\lambda\left\|\widehat{\beta}_L\right\|_1 \le \frac{1}{n}\|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1 \tag{33}$$

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

$$\frac{1}{n}\left\|X\widehat{\beta}_L - X\beta\right\|_2^2 + 2\lambda\left\|\widehat{\beta}_L\right\|_1 \le 2\lambda\|\beta\|_1 + \frac{2}{n}W^T X\left(\widehat{\beta}_L - \beta\right). \tag{34}$$

From Lemma 10, we have that

$$\begin{aligned}\frac{1}{n}\left\|X\widehat{\beta}_L - X\beta\right\|_2^2 &\le 2\lambda\sum_{j=1}^{p}\left|\beta_j\right| \\ &\quad - 2\lambda\sum_{j=1}^{p}\left|\widehat{\beta}_{L,j}\right| + \lambda\sum_{j=1}^{p}\left|\beta_j - \widehat{\beta}_{L,j}\right|\end{aligned} \tag{35}$$

holds with probability of at least $1 - p\exp(2\lambda^2 n/\sigma^2)$.

Adding the term $\sum_{j=1}^{p}\lambda|\widehat{\beta}_{j,L} - \beta_j|$ to both sides of this inequality, it yields that

$$\begin{aligned}&\frac{1}{n}\left\|X\widehat{\beta}_L - X\beta\right\|_2^2 + \lambda\sum_{j=1}^{p}\left|\beta_j - \widehat{\beta}_{L,j}\right| \\ &\le 2\lambda\sum_{j=1}^{p}\left|\beta_j\right| - 2\lambda\sum_{j=1}^{p}\left|\widehat{\beta}_{L,j}\right| \\ &\quad + 2\lambda\sum_{j=1}^{p}\left|\beta_j - \widehat{\beta}_{L,j}\right| \\ &\le 2\sum_{j=1}^{p}\lambda\left(\left|\beta_j - \widehat{\beta}_{L,j}\right| + \left|\beta_j\right| - \left|\widehat{\beta}_{L,j}\right|\right).\end{aligned} \tag{36}$$

Now, note that

$$\left|\widehat{\beta}_{j,L} - \beta_j\right| + \left|\beta_j\right| - \left|\widehat{\beta}_{j,L}\right| = 0 \tag{37}$$

since $j \notin J(\beta)$. So, we get that

$$\begin{aligned}&\frac{1}{n}\left\|X\widehat{\beta}_L - X\beta\right\|_2^2 + \lambda\sum_{j=1}^{p}\left|\beta_j - \widehat{\beta}_{L,j}\right| \\ &\le 2\lambda\sum_{j\in J(\beta)}\left(\left|\beta_j - \widehat{\beta}_{L,j}\right| + \left|\beta_j\right| - \left|\widehat{\beta}_{L,j}\right|\right) \\ &\le 4\lambda\sum_{j\in J(\beta)}\left|\beta_j - \widehat{\beta}_{L,j}\right|.\end{aligned} \tag{38}$$

To prove (32), it suffices to note that, from Lemma 10 and $\lambda > 2\lambda_0$, we have that

$$\left\|\frac{1}{n}X^T W\right\|_\infty \le \frac{\lambda}{2}. \tag{39}$$

Then

$$\begin{aligned}&\left\|\frac{1}{n}X^T X\left(\beta - \widehat{\beta}_L\right)\right\|_\infty \\ &= \left\|\frac{1}{n}X^T\left(Y - W - X\widehat{\beta}_L\right)\right\|_\infty \\ &\le \left\|\frac{1}{n}X^T\left(Y - X\widehat{\beta}_L\right)\right\|_\infty + \left\|\frac{1}{n}X^T W\right\|_\infty \\ &\le \lambda + \frac{\lambda}{2} = \frac{3\lambda}{2}.\end{aligned} \tag{40}$$

□

Lemma 12. *Let* $\beta \in \mathbb{R}^p$ *satisfy the Dantzig constraint*

$$\left\|\frac{1}{n}X^T X\left(\widehat{\beta}_D - \beta\right)\right\|_\infty \le 2\lambda \tag{41}$$

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

$$\left\|\delta_{J_0^c}\right\|_1 \le \left\|\delta_{J_0}\right\|_1. \tag{42}$$

Further, let the assumptions of Lemma 11 be satisfied. Then, with probability of at least $1 - p\exp(\lambda^2 n/2\sigma^2)$, *one has, for* $\lambda > \lambda_0$,

$$\left\|\frac{1}{n}X^T\left(X\beta - X\widehat{\beta}_D\right)\right\|_\infty \le 2\lambda. \tag{43}$$

Proof. Inequality (42) immediately follows from the definition of Dantzig selector.

Next, we prove (43). From Lemma 10 and analogously to (32), using the definition of Dantzig selector, we get that

$$\begin{aligned}&\left\|\frac{1}{n}X^T X\left(\beta - \widehat{\beta}_D\right)\right\|_\infty \\ &= \left\|\frac{1}{n}X^T\left(Y - W - X\widehat{\beta}_D\right)\right\|_\infty \\ &\le \left\|\frac{1}{n}X^T\left(Y - X\widehat{\beta}_D\right)\right\|_\infty \\ &\quad + \left\|\frac{1}{n}X^T W\right\|_\infty \le \lambda + \lambda = 2\lambda.\end{aligned} \tag{44}$$

□

Proof of Theorem 5. Set $\delta = \widehat{\beta}_L - \widehat{\beta}_D$. We start the calculation by simple matrix equality:

$$\begin{aligned}&\left\|X\beta - X\widehat{\beta}_D\right\|_2^2 - \left\|X\beta - X\widehat{\beta}_L\right\|_2^2 \\ &= 2\left(\widehat{\beta}_L - \widehat{\beta}_D\right)^T X^T\left(X\beta - X\widehat{\beta}_D\right) - \left\|X\left(\widehat{\beta}_L - \widehat{\beta}_D\right)\right\|_2^2 \\ &\le 2\|\delta\|_1\left\|X^T\left(X\beta - X\widehat{\beta}_D\right)\right\|_\infty - \|X\delta\|_2^2 \\ &\le 4n\lambda\|\delta\|_1 - \|X\delta\|_2^2,\end{aligned} \tag{45}$$

where the last inequality holds with probability of at least $1 - p\exp(\lambda^2 n/2\sigma^2)$ from (43).

By assumption LR$\varphi_2(s,1)$ and (42), we get that

$$\left\|X\beta - X\widehat{\beta}_D\right\|_2^2 - \left\|X\beta - X\widehat{\beta}_L\right\|_2^2 \leq 8n\lambda\left\|\delta_J\right\|_1 - \varphi_2^2 s\left\|\delta_J\right\|_1^2 \leq \frac{16n^2\lambda^2}{\varphi_2^2 s}. \tag{46}$$

From (32), a nearly identical argument yields that

$$\left\|X\beta - X\widehat{\beta}_L\right\|_2^2 - \left\|X\beta - X\widehat{\beta}_D\right\|_2^2 \leq 2\|\delta\|_1\left\|X^T X\left(\beta - \widehat{\beta}_L\right)\right\|_\infty - \left\|X\left(\beta - \widehat{\beta}_L\right)\right\|_2^2 \tag{47}$$

$$\leq 3n\lambda\|\delta\|_1 - \|X\delta\|_2^2 \tag{48}$$

$$\leq \frac{9n^2\lambda^2}{\varphi_2^2 s}. \tag{49}$$

This theorem follows from (46) and (49). □

Proof of Theorem 6. Set $\delta = \widehat{\beta}_L - \beta$. Using (31) with probability of at least $1 - p\exp(2\lambda^2 n/\sigma^2)$,

$$\frac{1}{n}\|X\delta\|_2^2 \leq 4\lambda\left\|\delta_{J_0}\right\|_1 - \lambda\|\delta\|_1. \tag{50}$$

From (48), we have

$$\frac{2}{n}\|X\delta\|_2^2 \leq 3\lambda\|\delta\|_1. \tag{51}$$

Then

$$\frac{1}{n}\|X\delta\|_2^2 \leq \frac{12}{5}\lambda\left\|\delta_{J_0}\right\|_1. \tag{52}$$

By assumption LR$\varphi_2(s,3)$, we obtain that

$$\varphi_2^2 s\left\|\delta_{J_0}\right\|_1^2 \leq \frac{1}{n}\|X\delta\|_2^2 \leq \frac{12}{5}\lambda\left\|\delta_{J_0}\right\|_1, \tag{53}$$

where $\varphi_2 = \varphi_2(s,3)$. Thus,

$$\left\|\delta_{J_0}\right\|_1 \leq \frac{12\lambda}{5s\varphi_2^2(s,c_0)}, \tag{54}$$

$$\frac{1}{n}\|X\delta\|_2^2 \leq \frac{144\lambda}{25s\varphi_2^2(s,c_0)}. \tag{55}$$

From (50), we have that

$$\lambda\|\delta\|_1 \leq 4\lambda\left\|\delta_{J_0}\right\|_1 - \frac{1}{n}\|X\delta\|_2^2 \leq \frac{4\lambda^2}{s\varphi_2^2(s,c_0)}. \tag{56}$$

Thus,

$$\|\delta\|_1 \leq \frac{4\lambda}{s\varphi_2^2(s,c_0)}. \tag{57}$$

Inequalities (55) and (57) coincide with (19) and (21), respectively.

Finally, to prove (20) we decompose δ into a set of vectors $\delta_{J_0}, \delta_{J_1}, \delta_{J_2}, \ldots, \delta_{J_K}$, such that J_0 corresponds to locations of the s largest coefficient of δ in absolute values, J_1 corresponds to locations of the next s largest coefficient of $\delta_{J_0^c}$ in absolute values, and so on. Hence we have that $J_0^c = \cup_{k=1}^K J_k$, where $K \geq 1$, $|J_k| = s$, for all $k = 1, \ldots, K-1$, and $|J_K| \leq s$.

It immediately follows that

$$\left\|\delta_{J_0^c}\right\|_2 \leq \frac{1}{\sqrt{s}}\|\delta\|_1. \tag{58}$$

On the other hand, from (54), we have that

$$\left\|\delta_{J_0}\right\|_2 \leq \left\|\delta_{J_0}\right\|_1 \leq \frac{12\lambda}{5s\varphi_2^2(s,c_0)}. \tag{59}$$

Therefore,

$$\|\delta\|_2 \leq \left\|\delta_{J_0}\right\|_2 + \left\|\delta_{J_0^c}\right\|_2 \leq \frac{12\lambda}{5s\varphi_2^2(s,c_0)} + \frac{1}{\sqrt{s}}\|\delta\|_1 \leq \frac{12\lambda}{5s\varphi_2^2(s,c_0)} + \frac{4\lambda}{\sqrt{s}s\varphi_2^2(s,c_0)} \leq \frac{4\lambda}{s\varphi_2^2(s,c_0)}\left(\frac{3}{5} + \frac{1}{\sqrt{s}}\right), \tag{60}$$

and the theorem follows. □

Proof of Theorem 7. Set $\delta = \widehat{\beta}_D - \beta$. Using (42) and (43), with probability of at least $1 - p\exp(\lambda^2 n/2\sigma^2)$, we have that

$$\frac{1}{n}\|X\delta\|_2^2 = \frac{1}{n}\delta^T X^T X\delta \leq \frac{1}{n}\left\|X^T X\delta\right\|_\infty\|\delta\|_1 \leq 2r\left(\left\|\delta_{J_0}\right\|_1 + \left\|\delta_{J_0^c}\right\|_1\right) \leq 4r\left\|\delta_{J_0}\right\|_1. \tag{61}$$

From assumption LR$\varphi_2(s,1)$, we get that

$$\frac{1}{n}\|X\delta\|_2^2 \geq s\varphi_2^2\left\|\delta_{J_0}\right\|_1^2, \tag{62}$$

where $\varphi_2 = \varphi_2(s,1)$. This and (61) yield that

$$\frac{1}{n}\|X\delta\|_2^2 \leq \frac{16\lambda^2}{s\varphi_2^2}, \qquad \left\|\delta_{J_0}\right\|_1 \leq \frac{4\lambda}{s\varphi_2^2}. \tag{63}$$

The first inequality in (63) implies (24). Next, (22) is straightforward in view of the second inequality in (63) and of relation (42). The proof of (23) follows from (20) in Theorem 6. From (22) and (58), we get that

$$\left\|\delta_{J_0^c}\right\|_2 \leq \frac{1}{\sqrt{s}}\frac{8\lambda}{s\varphi_2^2}. \tag{64}$$

Then

$$\begin{aligned}\|\delta\|_2 &\le \left\|\delta_{J_0}\right\|_2 + \left\|\delta_{J_0^c}\right\|_2 \\ &\le \frac{4\lambda}{s\varphi_2^2} + \frac{1}{\sqrt{s}}\frac{8\lambda}{s\varphi_2^2} \le \frac{4\lambda}{s\varphi_2^2}\left(1 + \frac{2}{\sqrt{s}}\right),\end{aligned} \tag{65}$$

where the second inequality holds from the second inequality in (63) and the inequality $\|\delta_{J_0}\|_2 \le \|\delta_{J_0}\|_1$. □

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

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

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] 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.

[4] E. Gautier and A. B. Tsybakov, "High-dimensional instrumental variables regression and confidence sets," Working Paper, 2011.

[5] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*, Springer Series in Statistics, Springer, New York, NY, USA, 2011.

[6] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[7] S. Q. Wang and L. M. Su, "The oracle inequalities on simultaneous Lasso and Dantzig selector in high-dimensional nonparametric regression," *Mathematical Problems in Engineering*, vol. 2013, Article ID 571361, 6 pages, 2013.

[8] S. A. van de Geer and P. Buhlmann, "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[9] S. Q. Wang and L. M. Su, "Recovery of high-dimensional spares signals via ℓ_1-minimization," *Journal of Applied Mathematics*, vol. 2013, Article ID 636094, 6 pages, 2013.

[10] 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.

[11] S. van de Geer, *The Deterministic Lasso*, Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland, 2007.

[12] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[13] S. Q. Wang and L. M. Su, "Simultaneous lasso and dantzig selector in high dimensional nonparametric regression," *International Journal of Applied Mathematics and Statistics*, vol. 42, no. 12, pp. 103–118, 2013.

[14] S. Q. Wang and L. M. Su, "New bounds of mutual incoherence property on sparse signals recovery," *International Journal of Applied Mathematics and Statistics*, vol. 47, no. 17, pp. 462–477, 2013.

[15] P. Alquier and M. Hebiri, "Generalization of L_1 constraints for high dimensional regression problems," *Statistics and Probability Letters*, vol. 81, no. 12, pp. 1760–1765, 2011.

[16] P. Zhao and B. Yu, "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, vol. 7, no. 12, pp. 2541–2563, 2006.

[17] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approximation*, vol. 34, no. 1, pp. 61–88, 2011.

[18] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008.

[19] R. G. Baraniuk, R. A. DeVore, and M. B. Davenport, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

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

[21] Y. de Castro, "A remark on the lasso and the Dantzig selector," *Statistics and Probability Letters*, vol. 83, no. 1, pp. 304–314, 2013.