Weaker Regularity Conditions and Sparse Recovery in High-Dimensional Regression

Shiqing Wang, Yan Shi, and Limin Su

College of Mathematics and Information Sciences, North China University of Water Resources and Electric Power, Zhengzhou 450045, China
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

Study of high-dimensional regression has attracted a great deal of attention. One concrete instance of a high-dimensional inference problem concerns the standard linear regression model:

\[ y = X\beta + W, \]

(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 \( \hat{\beta} \in \mathbb{R}^p \) 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 \( \beta \) is an archetypal example of such structure; that is, only a few components of \( \beta \) are different from zero, say \( s \)-sparsity; \( \beta \) 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 \( \ell_1 \)-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 \( \ell_1 \)-based relaxations to reliably estimate \( \beta \); 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 settings.

Thus, in the setting of high-dimensional linear regression, the interesting question is accurately estimating the regression vector \( \beta \) 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 $\ell_1$-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 what 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

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

(2)

the number of nonzero coordinates of $\beta$, where $I_{|\beta_j| \neq 0}$ denotes the indicator function

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

(3)

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

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

(4)

for the $\ell_q$-norm of the vector of $\beta$. Moreover, a vector $\beta$ 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 $\Delta_J$ the vector in $\mathbb{R}^p$ that has the same coordinates as $\Delta$ on $J$ and zero coordinates on the complement of $J$.

For linear regression model (1), regularized estimation with the $\ell_1$-norm penalty, also known as the Lasso [1] or the basis pursuit [2], refers to the following convex optimization problem:

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

(5)

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

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

(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 $\beta_{J_s} \in \mathbb{R}^{|J_s|}$ be a subvector of $\beta$ in $\mathbb{R}^p$ confined to $J_s$. One of the common properties of the Lasso and the Dantzig selector is that, for an appropriately chosen $\lambda$ and a vector $\delta = \tilde{\beta} - \beta$, where $\tilde{\beta}$ is the solution from either the Lasso or the Dantzig selector, it holds with high probability (cf. Lemmas 11 and 12):

$$\|\delta_{J_s}\|_1 \leq c_0 \|\delta_{J_s}\|_1,$$

(7)

where $c_0 = 1$ 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

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

(8)

is the set of nonzero coefficients of the true parameter $\beta$ of the model.

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

$$\Psi_n = \frac{1}{n} X^TX,$$

(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 $\beta$ has coefficients that are mostly 0. Clearly, the matrix $\Psi_n$ is degenerate, and ordinary least squares does not work in this case, since it requires that

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

(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 $\ell_1$-norm of only a part of $\delta$. Note that the role of (7) is to restrict set of vectors $\{\delta \in \mathbb{R}^p : \delta \neq 0\}$ to

$$\{\delta \in \mathbb{R}^p : \delta \neq 0, \|\delta_{J_s}\|_1 \leq c_0 \|\delta_{J_s}\|_1\}.$$  

(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:

$$\kappa(s, c_0) \equiv \min_{J_s \subseteq \{1, \ldots, p\}} \min_{1 \leq s, \delta \neq 0} \min_{\delta_{J_s \setminus \{s\}}} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta\|_2} > 0.$$  

(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
$C \log(p)\|\beta\|_n$, respectively, for both the Lasso and Dantzig selector, where $C$ is a positive constant and $\|\beta\|_n$ 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 $1 \leq s \leq p$ and a positive number $c_0$, the following condition holds:

$$\tau_2 \left( s, c_0 \right) \triangleq \min_{\|L\|_2=s} \min_{\|\delta\|_1 \leq c_0} \frac{\|X\delta\|_2}{\sqrt{n} \|\delta\|_1} > 0. \quad (13)$$

The two conditions are very similar. The only difference is the $\ell_1$- versus $\ell_2$-norm of a part of $\delta$ 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 $\|\delta\|_1$.

In order to obtain our regularity conditions in this paper, we decompose $\delta$ into a set of vectors $\delta_{s_0}, \delta_{s_1}, \delta_{s_2}, \ldots, \delta_{s_K}$, such that $S_{s_0}$ corresponds to locations of the $s$ largest coefficient of $\delta$ in absolute values, $S_1$ corresponds to locations of the next $s$ largest coefficient of $\delta_{s_0}$ in absolute values, and so on. Hence, we have $S_{k} = \bigcup_{i=1}^{k} S_k$, where $K \geq 1, |S_k| = s$, for all $k = 1, \ldots, K - 1$, and $|S_K| \leq s$.

For each $j \geq 1$, we have

$$\|\delta_j\|_2 \leq \sqrt{s}\|\delta_j\|_{\infty} \leq \frac{1}{\sqrt{s}} \|\delta_j\|_1, \quad (14)$$

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

$$\|\delta_{s_0}\|_2 \leq \sum_{k \geq 1} \|\delta_{s_k}\|_2 \leq s^{-1/2} \left( \|\delta_{s_1}\|_1 + \|\delta_{s_2}\|_1 + \ldots \right) \quad (15)$$

$$\leq s^{-1/2} \left( \|\delta_{s_1}\|_1 + \|\delta_{s_2}\|_1 \right) = s^{-1/2} \|\delta\|_1.$$

Replacing $\|\delta\|_1$ by $\sqrt{s}\|\delta_{s_0}\|_2$ in (13), we get the following assumption.

**Assumption 3** (LR$\phi_3(s,c_0)$). For some integer $s$ such that $1 \leq s \leq p$ and a positive number $c_0$, the following condition holds:

$$\phi_3 \left( s, c_0 \right) \triangleq \min_{\|L\|_2=s} \min_{\|\delta\|_1 \leq c_0} \frac{\|X\delta\|_2}{\sqrt{n} \|\delta\|_1} > 0. \quad (16)$$

The inequality $\sqrt{s}\|\delta_{s_0}\|_2 \leq \|\delta\|_1$ immediately implies that the assumption LR$\phi_3(s,c_0)$ is weaker than the assumptions RE$\tau_2(s,c_0)$ and RE$(s,c_0)$. Noting the norm $\|\delta\|_2$ in the denominator of (16), it makes the proof become more complicated. We need an equivalent condition of LR$\phi_3(s,c_0)$ for the sake of simplicity, as similarly discussed on equivalence (cf. [7, 13]).

**Assumption 4** (LR$\phi_3(s,c_0)$). For some integer $s$ such that $1 \leq s \leq p$ and a positive number $c_0$, the following condition holds:

$$\phi_3 \left( s, c_0 \right) \triangleq \min_{\|L\|_2=s} \min_{\|\delta\|_1 \leq c_0} \frac{\|X\delta\|_2}{\sqrt{n} \|\delta\|_1} > 0. \quad (17)$$

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$\phi_3(s,c_0)$ is satisfied.

**4. Main Results of Sparse Recovery for Regression Model**

In order to provide performance guarantees for $\ell_1$-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$\phi_3(s,c_0)$ is satisfied. In particular, for convenience, we assume that all the diagonal elements of the matrix $X^T X/n$ 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 $\|X\beta - \hat{X}_D\|_2$ and $\|X\beta - X\beta_L\|_2$ 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 $\hat{\beta}_L$ and Dantzig estimator $\hat{\beta}_D$ defined by (5) and (6) with the same $\lambda$. If LR$\phi_3(s,c_0)$ is satisfied, where $c_0 > 0$, then, with probability of at least $1 - p^{-1 - K/8}$, one has

$$\frac{1}{n} \|X\beta - X\beta_D\|_2^2 \leq \frac{16\lambda^2}{sq^2}. \quad (18)$$

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 $\hat{\beta}_L$ defined by (5) with $\lambda > 2\lambda_0 > 0$. If LR$\phi_3(s,3)$ is satisfied, where $c_0 > 0$, then, with probability of at least $1 - p \exp(2\lambda n/\sigma^2)$, one has

$$\|\hat{\beta}_L - \beta\|_2 \leq \frac{4\lambda}{sq^2} \left( s, c_0 \right), \quad (19)$$

$$\|\hat{\beta}_L - \beta\|_2 \leq \frac{4\lambda}{sq^2} \left( s, c_0 \right), \quad (20)$$

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

where $\phi_2 = \phi_2(s,3)$.

**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 \( \hat{\beta}_D \) 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

\[
\| \hat{\beta}_D - \beta \|_1 \leq \frac{8\lambda}{\sqrt{s}^2},
\]

\[
\| \hat{\beta}_D - \beta \|_2 \leq \frac{4\lambda}{\sqrt{s}^2} \left( 1 + \frac{2}{\sqrt{5}} \right),
\]

\[
\frac{1}{n} \| X (\hat{\beta}_D - \beta) \|_2^2 \leq \frac{16\lambda^2}{s \varphi_2^2},
\]

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

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

\[
\lambda = A\sigma \sqrt{\frac{\log p}{n}}, \quad A > \sqrt{\Sigma}.
\]

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

\[
\frac{1}{n} \| X\beta - X\hat{\beta}_D \|_2^2 - \frac{1}{n} \| X\beta - X\hat{\beta}_L \|_2^2 \leq \frac{16\lambda^2\sigma^2}{n\varphi_2^2} \log p,
\]

\[
\| \hat{\beta}_L - \beta \|_1 \leq \frac{4A\sigma}{\sqrt{s}^2} \left( \frac{3}{5} + \frac{1}{\sqrt{5}} \right) \sqrt{\frac{\log p}{n}},
\]

\[
\frac{1}{n} \| X (\hat{\beta}_L - \beta) \|_2^2 \leq \frac{144A^2}{25\varphi_2^2} \sigma^2 \log p,
\]

\[
\| \hat{\beta}_D - \beta \|_1 \leq \frac{8A\sigma}{\varphi_2^2} \sqrt{\frac{\log p}{n}},
\]

\[
\frac{1}{n} \| X (\hat{\beta}_D - \beta) \|_2^2 \leq \frac{16A^2\sigma^2}{\varphi_2^2} \log p,
\]

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

\[
\| \hat{\beta}_L - \beta \|_1 \leq \frac{4A}{r_1^2 (s, 1)} \sigma \sqrt{\frac{\log p}{n}},
\]

\[
\frac{1}{n} \| X (\hat{\beta}_L - \beta) \|_2^2 \leq \frac{144A^2 \sigma^2}{25r_1^2 (s, 1)} \log p.
\]

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 \( RE_2(s, c_0) \) and \( RE(s, c_0) \), since \( \sqrt{s}\| \delta \|_2 \leq \| \delta \|_1 \). Note that the inequality \( \sqrt{s}\| \delta \|_2 \leq \| \delta \|_1 \) 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 \( \lambda_0 > 0 \),

\[
P \left( \frac{1}{n} \| X^TW \|_\infty \geq \lambda_0 \right) \leq p \exp \left( \frac{\lambda_0^2 n}{2\sigma^2} \right).
\]

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

\[
P \left( \frac{1}{n} \| X^TW \|_\infty \geq \lambda_0 \right) \leq \sum_j P \left( \left| \frac{1}{n} \sum_i x_{i,j}w_i \right| \geq \lambda_0 \right)
\]

\[
\leq p \sum_j \left( \frac{n^{-1/2} \sum x_{i,j}w_i}{(\sigma)} \right) \geq \lambda_0 \sqrt{n} \left( \frac{n^{-1/2} \sum x_{i,j}w_i}{(\sigma)} \right)
\]

\[
\leq p \sigma \sum_j \left( \frac{|\eta|}{(\sigma)} \right) \geq \lambda_0 \sqrt{n} \left( \frac{|\eta|}{(\sigma)} \right)
\]

\[
\leq p \exp \left( \frac{\lambda_0^2 n}{2\sigma^2} \right),
\]

where \( \eta \sim N(0, 1) \). \( \square \)

Lemma 11. Let \( W_i \sim N(0, \sigma^2) \) be independent random variables with \( \sigma^2 > 0 \). Let \( \hat{\beta}_L \) be the Lasso estimator defined
by (5). Then, with probability of at least $1 - p \exp(2\lambda^2n/\sigma^2)$, one has, simultaneously for all $\beta \in \mathbb{R}^p$ and $\lambda > 2\lambda_0$,

$$\frac{1}{n} \|X\tilde{\beta}_L - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j - \tilde{\beta}_{L,j}| \leq 4\lambda \sum_{j \notin J(\beta)} |\beta_j - \tilde{\beta}_{L,j}|,$$

(31)

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

$$\frac{1}{n} \|X^TW\|_\infty \leq \frac{\lambda}{2}.$$  (39)

Then

$$\frac{1}{n} \|X^T \left(\beta - \tilde{\beta}_L\right)\|_\infty$$

$$= \frac{1}{n} \|X^T \left(\beta - \tilde{\beta}_D\right)\|_\infty$$

$$\leq \frac{1}{n} \|X^T \left(Y - X\tilde{\beta}_L\right)\|_\infty + \frac{1}{n} \|X^T W\|_\infty$$

$$\leq \lambda + \frac{\lambda}{2} = \frac{3\lambda}{2}.$$  (40)

Proof. By the definition of $\tilde{\beta}_L$, for all $\beta \in \mathbb{R}^p$, which is equivalent to

$$\frac{1}{n} \|X\tilde{\beta}_L - X\beta\|_2^2 + 2\lambda \|\tilde{\beta}_L\|_1 \leq \frac{1}{n} \|Y - X\beta\|_2^2 + 2\lambda \|\beta\|_1$$

(33)

From Lemma 10, we have that

$$\frac{1}{n} \|X\tilde{\beta}_L - X\beta\|_2^2 \leq 2\lambda \sum_{j=1}^{p} |\beta_j|$$

$$- 2\sum_{j=1}^{p} |\beta_{L,j}| + \lambda \sum_{j=1}^{p} |\beta_j - \tilde{\beta}_{L,j}|$$

(35)

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

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

$$\frac{1}{n} \|X\tilde{\beta}_L - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j - \tilde{\beta}_{L,j}|$$

$$\leq 2\lambda \sum_{j=1}^{p} |\beta_j| - 2\sum_{j=1}^{p} |\beta_{L,j}|$$

$$+ 2\lambda \sum_{j=1}^{p} |\beta_j - \tilde{\beta}_{L,j}|$$

(36)

$$\leq 2\lambda \left(|\beta_j - \tilde{\beta}_{L,j}| + |\beta_j| - |\tilde{\beta}_{L,j}|\right).$$

Now, note that

$$|\tilde{\beta}_{L,j} - \beta_j| + |\beta_j| - |\tilde{\beta}_{L,j}| = 0$$

(37)

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

$$\frac{1}{n} \|X\tilde{\beta}_L - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j - \tilde{\beta}_{L,j}|$$

$$\leq 2\lambda \sum_{j \notin J(\beta)} \left(|\beta_j - \tilde{\beta}_{L,j}| + |\beta_j| - |\tilde{\beta}_{L,j}|\right)$$

$$\leq 4\lambda \sum_{j \notin J(\beta)} |\beta_j - \tilde{\beta}_{L,j}|,$$

(38)

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

$$\|X\beta - X\tilde{\beta}_D\|_2^2 - \|X\beta - X\tilde{\beta}_L\|_2^2$$

$$= 2(\tilde{\beta}_L - \tilde{\beta}_D)^T X^T (X\beta - X\tilde{\beta}_D) - \|X(\tilde{\beta}_L - \tilde{\beta}_D)\|_2^2$$

$$\leq 2\|\delta\|_1 \|X^T (X\beta - X\tilde{\beta}_D)\|_\infty - \|X\delta\|_2^2$$

$$\leq 4n\lambda \|\delta\|_1 - \|X\delta\|_2^2,$$  (45)
where the last inequality holds with probability of at least $1 - p \exp(\lambda^2 n/2\sigma^2)$ from (43).

By assumption $LR\phi_2(s, 1)$ and (42), we get that
\[
\|X\beta - X\hat{\beta}_D\|_2^2 - \|X\beta - X\hat{\beta}_L\|_2^2 \leq 8n\lambda\|\delta_j\|_1 - \varphi_2^2\|\delta_j\|_1^2 \leq \frac{16n^2\lambda^2}{\varphi_2^4}.
\]  
(46)

From (32), a nearly identical argument yields that
\[
\|X\beta - X\hat{\beta}_L\|_2^2 - \|X\beta - X\hat{\beta}_L\|_2^2 \\
\leq 2\|\delta\|_1\|X^T X (\beta - \hat{\beta}_L)\|_{\infty} - \|X (\beta - \hat{\beta}_L)\|_2^2 \\
\leq 3n\lambda\|\delta\|_1 - \|X\delta\|_2^2 \\
\leq \frac{9n^2\lambda^2}{\varphi_2^4}.
\]  
(47)

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

**Proof of Theorem 6.** Set $\delta = \hat{\beta}_L - \beta$. Using (31) with probability of at least $1 - p \exp(2\lambda^2 n/\sigma^2)$,
\[
\frac{1}{n}\|X\delta\|_2^2 \leq 4\lambda\|\delta_j\|_1 - \lambda\|\delta\|_1.
\]  
(50)

From (48), we have
\[
\frac{2}{n}\|X\delta\|_2^2 \leq 3\lambda\|\delta\|_1.
\]  
(51)

Then
\[
\frac{1}{n}\|X\delta\|_2^2 \leq \frac{12}{5}\lambda\|\delta_j\|_1.
\]  
(52)

By assumption $LR\phi_2(s, 3)$, we obtain that
\[
\varphi_2^2\|\delta_j\|_1^2 \leq \frac{1}{n}\|X\delta\|_2^2 \leq \frac{12}{5}\lambda\|\delta_j\|_1.
\]  
(53)

where $\varphi_2 = \varphi_2(s, 3)$. Thus,\[
\|\delta_j\|_1 \leq \frac{12\lambda}{5s\varphi_2^2(s, c_0)}.
\]  
(54)

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

From (50), we have that
\[
\lambda\|\delta\|_1 \leq 4\lambda\|\delta_j\|_1 - \frac{1}{n}\|X\delta\|_2^2 \leq \frac{4\lambda^2}{s\varphi_2^2(s, c_0)}.
\]  
(56)

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

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

Finally, to prove (20) we decompose $\delta$ into a set of vectors $\delta_j, \delta_j, \delta_j, \ldots, \delta_j$, such that $J_0$ corresponds to locations of the $s$ largest coefficient of $\delta$ in absolute values, $J_1$ corresponds to locations of the $s$ largest coefficient of $\delta_j$, in absolute values, and so on. Hence we have that $\hat{J}_0 = \cup_{k=1}^K J_k$, where $K \geq 1, |J_k| = s$, for all $k = 1, \ldots, K - 1$, and $|J_K| \leq s$.

It immediately follows that
\[
\|\delta_j\|_2 \leq \frac{1}{\sqrt{s}}\|\delta\|_1.
\]  
(58)

On the other hand, from (54), we have that
\[
\|\delta_j\|_2 \leq \|\delta_j\|_1 \leq \frac{12\lambda}{5s\varphi_2^2(s, c_0)}.
\]  
(59)

Therefore,
\[
\|\delta\|_2 \leq \|\delta_j\|_2 + \|\delta_j\|_2 \\
\leq \frac{12\lambda}{5s\varphi_2^2(s, c_0)} + \frac{1}{\sqrt{s}}\|\delta\|_1
\]  
(60)

and the theorem follows.

**Proof of Theorem 7.** Set $\delta = \hat{\beta}_D - \beta$. Using (42) and (43), with probability of at least $1 - p \exp(\lambda^2 n/2\sigma^2)$, we have that
\[
\frac{1}{n}\|X\delta\|_2^2 = \frac{1}{n}\delta^T X^T X\delta \\
\leq \frac{1}{n}\|X^T X\delta\|_{\infty}\|\delta\|_1 \\
\leq 2r(\|\delta_j\|_1 + \|\delta_j\|_1) \leq 4r\|\delta_j\|_1
\]  
(61)

From assumption $LR\phi_2(s, 1)$, we get that
\[
\frac{1}{n}\|X\delta\|_2^2 \geq s\varphi_2^2\|\delta_j\|_1^2
\]  
(62)

where $\varphi_2 = \varphi_2(s, 1)$. This and (61) yield that
\[
\frac{1}{n}\|X\delta\|_2^2 \leq \frac{16\lambda^2}{s\varphi_2^2}, \|\delta_j\|_1 \leq \frac{4\lambda}{s\varphi_2^2}
\]  
(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
\[
\|\delta_j\|_2 \leq \frac{1}{\sqrt{s}}\frac{8\lambda}{s\varphi_2^2}.
\]  
(64)
Then
\[ \|\delta\|_2 \leq \|\delta_{j_0}\|_2 + \|\delta_{j_0}'\|_2 \]
\[ \leq \frac{4\lambda}{s\varphi_2} + \frac{8\lambda}{\sqrt{3}s\varphi_2} \leq \frac{4\lambda}{s\varphi_2} \left( 1 + \frac{2}{\sqrt{3}} \right), \tag{65} \]
where the second inequality holds from the second inequality in (63) and the inequality \( \|\delta_{j_0}\|_2 \leq \|\delta_{j_0}'\|_1 \). □

**Conflict of Interests**

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

**References**


