Journal of Machine Learning Research 11 (2010) 3519-3540 Submitted 10/09; Revised 8/10; Published 12/10

Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q Loss in ℓ_r Balls

Fei Ye FYE@DRW.COM
Department of Quantitative Research
DRW Trading Group
Chicago, IL 60661-2555, USA

Cun-Hui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics and Biostatistics
Rutgers University
Piscataway, NJ 08854-8019, USA

Editor: Hui Zou

Abstract

We consider the estimation of regression coefficients in a high-dimensional linear model. For regression coefficients in ℓ_r balls, we provide lower bounds for the minimax ℓ_q risk and minimax quantiles of the ℓ_q loss for all design matrices. Under an ℓ_0 sparsity condition on a target coefficient vector, we sharpen and unify existing oracle inequalities for the Lasso and Dantzig selector. We derive oracle inequalities for target coefficient vectors with many small elements and smaller threshold levels than the universal threshold. These oracle inequalities provide sufficient conditions on the design matrix for the rate minimaxity of the Lasso and Dantzig selector for the ℓ_q risk and loss in ℓ_r balls, $0 \le r \le 1 \le q \le \infty$. By allowing $q = \infty$, our risk bounds imply the variable selection consistency of threshold Lasso and Dantzig selectors.

Keywords: variable selection, estimation, oracle inequality, minimax, linear regression, penalized least squares, linear programming

1. Introduction

As modern information technologies relentlessly generate voluminous and complex data, penalized high-dimensional regression methods have been a focus of intense research activities in machine learning and statistics in recent years. In many statistical and engineering applications, the number p of design variables (features, covariates) can be larger or even of larger order than the sample size n, but the number of important variables may still be smaller than the sample size. In such cases, one seeks a parsimonious model that fits the data well. In linear regression, a popular approach for achieving this goal is to impose a suitable penalty on the empirical loss.

This paper considers the estimation of a sparse vector of regression coefficients in a linear model. Specifically, we are interested in the rate minimaxity of the Lasso and Dantzig selector under the ℓ_q loss for the estimation of regression coefficients in ℓ_r balls. This requires lower bounds of the minimax ℓ_q risk and minimax quantiles of the ℓ_q loss over all estimators as well as matching upper bounds for the Lasso and Dantzig selector.

Let $y \in \mathbb{R}^n$ be a response vector and $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ be a design matrix. The Lasso (Tibshirani, 1996) is the ℓ_1-penalized estimator

$$\widehat{\beta}^{(L)}(\lambda) = \arg\min_b \left\{ \|y - Xb\|^2/(2n) + \lambda \|b\|_1 \right\} \tag{1}$$

for the regression coefficients. In the signal processing literature, the Lasso is known as basis pursuit (Chen and Donoho, 1994). The Lasso has the interpretation as boosting (Freund and Schapire, 1996; Friedman, Hastie, and Tibshirani, 2000) and is computationally feasible for high-dimensional data (Osborne, Presnell, and Turlach, 2000a,b; Efron, Hastie, Johnstone, and Tibshirani, 2004). More recently, Candes and Tao (2007) proposed an ℓ_1-minimization method called the Dantzig selector,

$$\widehat{\beta}^{(D)}(\lambda) = \arg\min_b \left\{ \|b\|_1 : |x_j'(y - Xb)/n| \le \lambda, \forall j \right\}. \tag{2}$$

A focus of recent studies of high-dimensional linear regression has been on the performance of the Lasso and Dantzig selector for the estimation of the regression coefficients. Candes and Tao (2007) derived an elegant probabilistic upper bound for the ℓ_2 loss of the Dantzig selector under a condition on the number of nonzero coefficients and a uniform uncertainty principle on the Gram matrix. Efron, Hastie, and Tibshirani (2007) questioned whether a similar performance bound holds for the Lasso estimator as well. Upper bounds for the ℓ_q loss of the Lasso estimator has being studied by Bunea, Tsybakov, and Wegkamp (2007) and van de Geer (2008) for $q = 1$, Zhang and Huang (2008) for $q \in [1,2]$, Meinshausen and Yu (2009) for $q = 2$, Bickel, Ritov, and Tsybakov (2009) for $q \in [1,2]$ with a parallel analysis of the Dantzig selector, and Zhang (2009b) for $q \ge 1$. Under different sets of regularity conditions on the Gram matrix and the sparsity of regression coefficients $\beta \in \mathbb{R}^p$, these results provide error bounds of the form $\|\widehat{\beta} - \beta\|_q \le O(k^{1/q}\lambda)$, where k is the number of nonzero entries of a target vector of regression coefficients or an intrinsic dimensionality of the sparse estimation problem. For $N(0, \sigma^2)$ errors and standardized designs with $\|x_j\| = \sqrt{n}$, these studies require a universal penalty level $\lambda_{univ} = \sigma\sqrt{(2/n)\log p}$ or greater for the Dantzig selector and a penalty level λ greater by a constant factor than λ_{univ} for the Lasso. Different sets of regularity conditions lead to different forms of constant factors in the error bounds so that the existing error bounds are typically not directly comparable mathematically.

This paper contributes to high-dimensional regression in several ways. We provide lower bounds for the minimax ℓ_q risk in ℓ_r balls and the minimax quantiles of the ℓ_q loss for all designs X. We derive sufficient conditions on X for the Lasso and Dantzig selector to attain the rate of the minimax ℓ_q risk and the minimax quantiles of the ℓ_q loss. We provide oracle inequalities for the ℓ_q loss of the Lasso and Dantzig selector which sharpen, unify and extend the existing results and allow the penalty level λ to be of smaller order than the universal penalty level.

The rest of the paper is organized as follows. In Section 2, we describe lower bounds for the minimax risk and loss in ℓ_r balls. In Section 3, we provide oracle inequalities for the Lasso and Dantzig selector under the ℓ_0 sparsity of regression coefficients. We compare these oracle inequalities with existing ones and describe their implications in variable selection and rate minimaxity in ℓ_0 balls. In Section 4, we provide more general oracle inequalities to allow many small regression coefficients and penalty levels of smaller order than $\sigma\sqrt{(2/n)\log p}$. These oracle inequalities are used to establish the rate minimaxity for the ℓ_q loss in ℓ_r balls. In Section 5, we make a few remarks. An appendix contains all proofs.

We use the following notation throughout the paper. For vectors $v = (v_1, \ldots, v_p)'$, $\|v\|_0 = \#\{j : v_j \neq 0\}$ and $\|v\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|v\| = \|v\|_2$ and the usual extension to $q = \infty$. Functions are applied to vectors in individual components, $f(v) = (f(v_1), \ldots, f(v_p))'$. For matrices M and $0 \leq a, b \leq \infty$, $\|M\|_{a,b} = \max\{\|Mv\|_b : \|v\|_a = 1\}$ is the operator norm from ℓ_a to ℓ_b. For subsets A and B of $\{1, \ldots, p\}$, $X_A = (x_j, j \in A)$, $\Sigma_{A,B} = X_A' X_B/n$, $\Sigma_{A,*} = X_A' X/n$, $\Sigma_A = \Sigma_{A,A}$, and P_A is the projection from $\mathbb{R}^n$ to the linear span of $\{x_j : j \in A\}$. For real x, $x_+ = \max(x, 0)$, $1/x_+ = \infty$ for $x \leq 0$, and $\lceil x \rceil$ is the largest integer upper bound of x. For real numbers a_n and b_n, $a_n \approx b_n$ means $a_n = (1 + o(1))b_n$, $a_n \lessapprox b_n$ means $a_n \leq (1 + o(1))b_n$, and $a_n \lesssim b_n$ means $a_n = O(b_n)$. For simplicity, the dependence of estimators on the penalty level λ is suppressed unless otherwise stated.

2. Lower Bounds for the Estimation Risk and Loss

Consider the linear model

$$y = X\beta + \varepsilon = \sum_{j=1}^{p} \beta_j x_j + \varepsilon. \tag{3}$$

Throughout the sequel, $P_{\beta,X}$ is the probability measure given $\{\beta, X\}$ under which $\varepsilon \sim N(0, \sigma^2 I_n)$. For simplicity, we also assume $\|x_j\|^2 = n$ whenever $P_{\beta,X}$ is referred to. Define

$$\Theta_{r,R} = \{v \in \mathbb{R}^p : \|v\|_r \leq R\},\ r > 0, \quad \Theta_{0,k} = \{v \in \mathbb{R}^p : \|v\|_0 \leq k\}, \tag{4}$$

as the ℓ_r and ℓ_0 balls respectively. Let $\sigma_n = \sigma/\sqrt{n}$ and

$$\lambda_{univ} = \sigma_n \sqrt{2 \log p},\ \lambda_{mm} = \sigma_n \left\{ 2 \log \left(\frac{\sigma_n^r p}{R^r} \right) \right\}^{1/2},\ \lambda_{mm} = \sigma_n \sqrt{2 \log(p/k)}, \tag{5}$$

be respectively the universal (univ) penalty level (Donoho and Johnstone, 1994) and the minimax (mm) penalty levels for the ℓ_r and ℓ_0 balls. The dependence of λ_{mm} on $\{r, R\}$ or $\{0, k\}$ is suppressed since λ_{mm} is always used in association with a specific ball in (4).

Donoho and Johnstone (1994) proved that for $0 < r < q$ and based on a p-vector $\tilde{y} \sim N(\beta, \sigma_n^2 I_p)$, the minimax ℓ_q risk in the ℓ_r ball $\Theta_{r,R}$ is approximately

$$\inf_{\delta} \sup_{\beta \in \Theta_{r,R}} E_{\beta,X} \left\| \delta(\tilde{y}) - \beta \right\|_q^q = (1 + o(1)) R^r \lambda_{mm}^{q-r}$$

and is achieved within an infinitesimal fraction by threshold estimators at the threshold level λ_{mm}, provided that $\lambda_{mm}/\sigma_n \to \infty$ and $R^r/\lambda_{mm}^r \to \infty$. Here the infimum is taken over all Borel mappings δ of proper dimensions. The following theorem extends their result to the estimation of regression coefficients under $P_{\beta,X}$. For any class of vectors $\Theta \subset \mathbb{R}^p$, the minimax ℓ_q risk is $\mathscr{R}_q(\Theta; X) = \inf_\delta \sup_{\beta \in \Theta} E_{\beta,X} \|\delta(X, y) - \beta\|_q^q$.

Theorem 1 *Let $\Theta_{r,R}$ and $\Theta_{0,k}$ be as in (4) and λ_{mm} as in (5) with $R > 0$ and $q \geq r > 0$. Suppose $R^r/\lambda_{mm}^r \to \infty$, $k \to \infty$ and $\lambda_{mm}/\sigma_n \to \infty$. Then,*

$$\frac{\mathscr{R}_q(\Theta_{r,R}; X)}{R^r \lambda_{mm}^{q-r}} \geq (1 + o(1)), \quad \frac{\mathscr{R}_q(\Theta_{0,k}; X)}{k \lambda_{mm}^q} \geq (1 + o(1)). \tag{6}$$

Moreover, for either $\Theta = \Theta_{r,R}$ with $k = R^r/\lambda_{mm}^r$ or $\Theta = \Theta_{0,k}$,

$$\inf_X \inf_\delta \sup_{\beta \in \Theta} P_{\beta,X} \left\{ \|\delta(X, y) - \beta\|_q^q \geq (1 - \varepsilon) k \lambda_{mm}^q \right\} \geq \frac{\varepsilon + o(1)}{3^q},\ 0 < \varepsilon < 1. \tag{7}$$

Remark 2 *The value k in (7) can be viewed as an intrinsic lower bound for the number of parameters to be estimated for the minimaxity under the ℓ_q loss. By (5), $\lambda_{mm} \le \lambda_{univ}$ iff $R \ge \sigma_n$ for $r > 0$. For $\lambda_{univ} \asymp \lambda_{mm}$ and respectively for $r = 0$ and $0 < r \le 1$, Corollary 4 in Section 3.1 and Theorem 17 in Section 4.2 provide conditions on X for the Lasso and Dantzig selector to match the rate of the minimax lower bound $k\lambda_{mm}^q$ in (7). For $\lambda_{mm} \ll \lambda_{univ}$, Theorem 19 in Section 4.2 gives the rate minimaxity of the Lasso.*

During the revision of this paper, we became aware of the technical report of Raskutti, Wainwright, and Yu (2009). The lower bounds in Theorem 1 are identical for all design matrices X and thus are sharp only up to a constant factor under certain conditions on X. For example, the minimax risk in a parameter class $\Theta \ni 0$ is no smaller than the radius of the null set $\Theta \cap \{b : Xb = 0\}$. This and some other aspect of the design matrix have been used to derive sharper lower bounds in Raskutti, Wainwright, and Yu (2009). In our technical report (Ye and Zhang, 2009) and in Zhang (2009a), Theorem 1 only covers the case $r > 0$.

3. Oracle Inequalities under ℓ_0 Sparsity and Variable Selection

We discuss in three subsections oracle inequalities for the ℓ_q loss, related work, and variable selection. We focus on coefficient vectors with a relatively small number of nonzero entries here. The more complicated ℓ_r-sparse case will be considered in Section 4.

3.1 Oracle Inequalities and Rate Minimaxity under ℓ_0 Sparsity

For $\xi \ge 0$ and $J \subset \{1, \ldots, p\}$, define cone invertibility factors (CIF)

$$CIF_{q,\ell}(\xi, J) = \inf \Big\{ \frac{|J|^{1/q} \|\Sigma u\|_\infty}{\|u_A\|_q} : u \in \mathscr{C}(\xi, J), |A \setminus J| \le \ell \Big\} \tag{8}$$

with cones $\mathscr{C}(\xi, J) = \{u : \|u_{J^c}\|_1 \le \xi \|u_J\|_1 \ne 0\}$. Note that $CIF_{q,\ell}(\xi, J) = CIF_{q,p-|J|}(\xi, J)$ for $\ell \ge p - |J|$ and since the infimum is attained when u_A takes the ℓ largest elements of u outside J, $CIF_{q,\ell}(\xi, J)$ is decreasing in ℓ. Similarly, define sign-restricted cone invertibility factors (SCIF)

$$SCIF_{q,\ell}(\xi, J) = \inf \Big\{ \frac{|J|^{1/q} \|\Sigma u\|_\infty}{\|u_A\|_q} : u \in \mathscr{C}_-(\xi, J), |A \setminus J| \le \ell \Big\}. \tag{9}$$

with sign-restricted cones $\mathscr{C}_-(\xi, J) = \{u \in \mathscr{C}(\xi, J) : u_j \Sigma_{j,*} u \le 0 \ \forall j \notin J\}$. We first present oracle inequalities in terms of the (sign-restricted) CIF.

Theorem 3 *Let $\widehat{\beta}^{(D)}$ and $\widehat{\beta}^{(L)}$ be the Lasso and Dantzig selector in (1) and (2), $1 \le q \le \infty$, $\beta^* \in \mathbb{R}^p$, $J = \{j : \beta_j^* \ne 0\}$ and $z_\infty^* = \|X'(y - X\beta^*)/n\|_\infty$.*
(i) Let $CIF_{q,\ell}(1, J)$ and $SCIF_{q,\ell}(\xi, J)$ be as in (8) and (9). In the event $z_\infty^* \le \lambda$,

$$\|\widehat{\beta}^{(D)} - \beta^*\|_q \le \frac{|J|^{1/q}(\lambda + z_\infty^*)}{CIF_{q,p}(1, J)} \le \frac{2|J|^{1/q}\lambda}{CIF_{q,p}(1, J)}, \tag{10}$$

and in the event $z_\infty^* \le \lambda(\xi - 1)/(\xi + 1)$

$$\|\widehat{\beta}^{(L)} - \beta^*\|_q \le \frac{|J|^{1/q}(\lambda + z_\infty^*)}{SCIF_{q,p}(\xi, J)} \le \frac{\{2\xi/(\xi + 1)\}|J|^{1/q}\lambda}{SCIF_{q,p}(\xi, J)}. \tag{11}$$

(ii) For $\lambda = \sigma\sqrt{(2/n)\log(p/\varepsilon)}$ *and* $\lambda = \sigma\sqrt{(2/n)\log(p/\varepsilon)}(\xi+1)/(\xi-1)$ *respectively, (10) and (11) hold with* $\beta^* = \beta$ *and at least probability* $1-\varepsilon/\sqrt{\pi\log(p/\varepsilon)}$ *under* $P_{\beta,X}$.

Theorem 3 immediately provides a sufficient condition on the design X for the rate minimaxity of the Lasso and Dantzig selector in the quantiles of the ℓ_q loss in ℓ_0 balls, in the sense of (7) of Theorem 1. We state this result as the following corollary. Define

$$CIF_q^*(\xi,k) = \inf\Big\{k^{1/q}\|\Sigma u\|_\infty/\|u\|_q : 0 < \|u\|_1 \le (1+\xi)\max_{|J|=k}\|u_J\|_1\Big\}. \tag{12}$$

Corollary 4 *Suppose* $k\wedge(p/k)\to\infty$ *for certain* $(p,k) = (p_n,k_n)$.
(i) Suppose $M_{q,k}^{(D)} = 2\sqrt{\log p}/\{\sqrt{\log(p/k)}CIF_q^*(1,k)\} \le M_* < \infty$. *Then, the Dantzig selector* $\widehat\beta = \widehat\beta^{(D)}$ *with* $\lambda = \lambda_{univ}$ *in (5) is rate minimax in the sense of*

$$\sup_{\beta\in\Theta_{0,k}} P_{\beta,X}\big\{\|\widehat\beta-\beta\|_q \ge M_* k^{1/q}\lambda_{mm}\big\} \to 0. \tag{13}$$

(ii) Suppose $M_{q,k,\xi}^{(L)} = \max_{|J|\le k} 2\{\xi/(\xi-1)\}\sqrt{\log p}/\{\sqrt{\log(p/k)}SCIF_{q,p}(\xi,J)\} \le M_* < \infty$. *Then, (13) holds for the Lasso* $\widehat\beta = \widehat\beta^{(L)}$ *with* $\lambda = \lambda_{univ}(\xi+1)/(\xi-1)$.

The proof of Theorem 3 is simple since the (sign-restricted) CIF are exactly what we need. Let $h^{(D)} = \beta^{(D)}(\lambda)-\beta^*$, $h^{(L)} = \beta^{(L)}(\lambda)-\beta^*$ and $h = h^{(D)}$ or $h^{(L)}$ throughout the sequel. It follows from the feasibility of β^* for the constraint in (2) and the Karush-Kuhn-Tucker conditions for the Lasso (1) respectively that for $z_\infty^* \le \lambda$,

$$h^{(D)} \in \mathscr{C}(1,J),\ v'\Sigma h^{(L)} \le (z_\infty^*+\lambda)\|v_J\|_1 + (z_\infty^*-\lambda)\|v_{J^c}\|_1, \tag{14}$$

for all vectors v satisfying $\mathrm{sgn}(v_{J^c}) = \mathrm{sgn}(h^{(L)}_{J^c})$. With $v = h^{(L)}$ or $v_J = 0$ in (14), it follows that $h^{(L)} \in \mathscr{C}_-(\xi,J)$ for $\xi \ge (\lambda+z_\infty^*)/(\lambda-z_\infty^*)$. Theorem 3 then follows from

$$\|\Sigma h\|_\infty \le z_\infty^* + \lambda. \tag{15}$$

The lower bounds for the (sign-restricted) CIF in the following proposition facilitate more explicit results and connections to existing approaches. For given $\{J,\ell\}$ define

$$\phi_{q,\ell}(u,A,r,w,B;\xi,J) = 1-\xi|J|^{1-1/q}(a_r/\ell)^{1-1/r}\|\Sigma_{B,A}w_A\|_{r/(r-1)} \tag{16}$$

in the following domain: $u\in\mathscr{C}(\xi,J)$, $A = \arg\max_{|B\setminus J|\le\ell}\|u_B\|_1$ (determined by u up to ties), $r\ge 1$, $w_A'\Sigma_A u_A = \|u_A\|_q$, $B\cap A = \emptyset$, $|B| = \lceil \ell/a_r\rceil$ and $a_r = (1-1/r)/r^{1/(r-1)}$.

Proposition 5 *Let* $\{\xi,q,r\}\subset[1,\infty]$, $1\le s\le q$ *and* $0<\ell\le p-|J|$. *Then,*

$$\max\Big\{\frac{CIF_{q,\ell}(\xi,J)}{CIF_{s,p}(\xi,J)}, \frac{SCIF_{q,\ell}(\xi,J)}{SCIF_{s,p}(\xi,J)}\Big\} \le C_{s,q}(\xi,|J|/\ell), \tag{17}$$

where $C_{s,q}(\xi,t) = (1+\xi)^{(1/s-1/q)/(1-1/q)}\{1+\xi^q(a_q t)^{q-1}\}^{(1-1/s)/(q-1)}$;

$$CIF_{q,\ell}(\xi,J) = \inf_{u\in\mathscr{C}(\xi,J)}\inf_{|A\setminus J|=\ell}\sup_{v\ne0}\frac{|J|^{1/q}v'\Sigma u}{\|v\|_1\|u_A\|_q} \ge \phi_{q,\ell}^*(\xi,J), \tag{18}$$

where $\phi^*_{q,\ell}(\xi,J) = \min_{u,A}\max_{r,w}\min_B \phi_{q,\ell}(u,A,r,w,B;\xi,J)|J|^{1/q}/\|w_A\|_1$*; and*

$$SCIF_{q,\ell}(\xi,J) \geq \inf_{u\in\mathscr{C}_-(\xi,J)}\inf_{|A\setminus J|=\ell}\sup_{v\in\mathscr{Q}(u,J)}\frac{|J|^{1/q}v'\Sigma u}{\|v_J\|_1\|u_A\|_q} \geq \phi^{**}_{q,\ell}(\xi,J), \tag{19}$$

where $\phi^{**}_{q,\ell}(\xi,J) = \min_{u,A}\max_{r,w}\min_B \phi_{q,\ell}(u,A,r,w,B;\xi,J)|J|^{1/q}/\|w_J\|_1$ *with* $\mathscr{Q}(u,J) = \{v : sgn(v_{J^c}) = sgn(u_{J^c})\}$ *and vectors* $u\in\mathscr{C}_-(\xi,J)$ *and* $w\in\mathscr{Q}(u,J)$.

Remark 6 *Taking* $w = u\|u_A\|/(u'_A\Sigma_A u_A)$ *in (16) for the* ℓ_2 *loss, we find*

$$\begin{aligned}\widetilde{\phi}^*_{2,\ell}(\xi,J) &= \min_{u,A}\max_r\min_B\big\{u'_A\Sigma_A u_A - \xi|J|^{1/2}(a_r/\ell)^{1-1/r}\|\Sigma_{B,A}u_A\|_{r/(r-1)}\big\},\\ &\leq \min\big\{\phi^{**}_{2,\ell}(\xi,J), \{(1+\ell/|J|)^{1/2}\wedge(1+\xi)\}\phi^*_{2,\ell}(\xi,J)\big\}\end{aligned} \tag{20}$$

with $\|u_A\| = 1$ *and* $\{A,r,B\}$ *in (16), due to* $\|u_A\|_1 \leq \{(1+\ell/|J|)^{1/2}\wedge(1+\xi)\}|J|^{1/2}$ *and* $\|w_J\|_1 \leq |J|^{1/2}/(u'_A\Sigma_A u_A)$*. However, for* $q>2$*,* $w\propto u$ *in (16) does not lead to a right normalization since* $\|u_A\|/|A|^{1/2}$ *does not control* $\|u_A\|_q/|A|^{1/q}$*. For general* $q\in[1,\infty]$*,* $\|\Sigma_A w_A\|_{q/(q-1)} = w'_A\Sigma_A u_A/\|u_A\|_q = 1$ *gives a properly length normalized lower bound:*

$$\phi^*_{q,\ell}(\xi,J) \geq \min_{A,B}\big\{1-\xi|J|^{1-1/q}(a_r/\ell)^{1-1/r}\|\Sigma_A^{-1}\Sigma_{A,B}\|_{r,q}\big\}|J|^{1/q}/\|\Sigma_A^{-1}\|_{\infty,q}. \tag{21}$$

For $(|A|,|B|,\|u\|,\|v\|_r) = (\lceil a\rceil,\lceil b\rceil,1,1)$ with $A\cap B=\emptyset$, define

$$\delta^\pm_a = \max_{A,u}\Big\{\pm\Big(\|\Sigma_A u\|-1\Big)\Big\},\ \delta_a = \delta^+_a\vee\delta^-_a,\ \theta^{(r)}_{a,b} = \max_{A,B,u,v} v'\Sigma_{A,B}u. \tag{22}$$

Oracle inequalities for the ℓ_q loss, $1\leq q\leq 2$, have been given in terms of the quantities in (22). The quantities $1\pm\delta^\pm_a$ give the maximum and minimum sparse eigenvalues of the Gram matrix Σ up to dimension $\lceil a\rceil$. For $\|u_A\|=1$, we have $u'_A\Sigma_A u_A \geq 1-\delta^-_{|A|}$, $\|\Sigma_{B,A}u_A\|_{r/(r-1)}\leq\theta^{(r)}_{|B|,|A|}$ and $\|\Sigma_{B,A}u_A\|^2\leq(1+\delta^+_{|B|})u'_A\Sigma_A u_A$. Thus, Theorem 3, (17) with $C_{q,2}(\xi,k/\ell) = (1+\xi)^{2/q-1}\{1+\xi^2 k/(4\ell)\}^{1-1/q}$, and (20) with $r\in\{2,\infty\}$ yield the following corollary.

Corollary 7 *Let* $\|\beta^*\|_0 = |J| = k$ *and* $1\leq q\leq 2$*. Then, for* $z^*_\infty = \|X'(y-X\beta^*)/n\|_\infty\leq\lambda$*,*

$$\|\widehat{\beta}^{(D)}(\lambda)-\beta^*\|_q \leq \frac{2|J|^{1/q}\lambda}{CIF_{q,p}(1,J)} \leq \frac{2^{2/q}(1+k/(4\ell))^{1-1/q}(2\wedge\sqrt{1+\ell/k})k^{1/q}\lambda}{(1-\delta^-_{k+\ell})_+\{1-\min(\widetilde{F}_1,\widetilde{F}_2,\widetilde{F}_3)\}_+}, \tag{23}$$

where $1\leq\ell\leq p-k$*,* $\widetilde{F}_1 = (k^{1/2}/\ell)\theta^{(\infty)}_{\ell,k+\ell}/(1-\delta^-_{k+\ell})_+$*,* $\widetilde{F}_2 = \sqrt{k/(4\ell)}\,\theta^{(2)}_{4\ell,k+\ell}/(1-\delta^-_{k+\ell})_+$ *and* $\widetilde{F}_3 = \{(k/(4\ell))(1+\delta^+_{4\ell})/(1-\delta^-_{k+\ell})_+\}^{1/2}$*. Moreover, for* $z^*_\infty\leq\lambda(\xi-1)/(\xi+1)$*,*

$$\|\widehat{\beta}^{(L)}(\lambda)-\beta^*\|_q \leq \frac{\{2\xi/(\xi+1)\}|J|^{1/q}\lambda}{SCIF_{q,p}(\xi,J)} \leq \frac{(1+\xi)^{2/q-2}(1+\xi^2 k/(4\ell))^{1-1/q}2\xi k^{1/q}\lambda}{(1-\delta^-_{k+\ell})_+\{1-\xi\min(\widetilde{F}_1,\widetilde{F}_2,\widetilde{F}_3)\}_+}. \tag{24}$$

3.2 Related Work

Here we discuss the connections between the results in Section 3.1 and related work. We show that the (sign-restricted) CIF-based oracle inequalities in Theorem 3 and Corollary 7 sharpen, unify and extend existing performance bounds and offer a clear explanation of the differences among existing analyses.

The CIF and SCIF in (8) and (9) are related to the restricted eigenvalues (RE)

$$RE_{q,\ell}(\xi,J) = \inf\Big\{ \frac{|J|^{1/q}\|Xu\|}{|nJ|^{1/2}\|u_A\|_q} : u \in \mathscr{C}(\xi,J), |A\setminus J| \le \ell \Big\}, \tag{25}$$

since they all conveniently provide factors required in proofs of the same type of oracle inequalities. The quantity (25) includes as special cases the ℓ_2 version $RE_{2,p}(\xi,J)$ of Bickel, Ritov, and Tsybakov (2009) and Koltchinskii (2009) and the constant factor $RE_{1,0}(3,J)$ for the compatibility condition of van de Geer (2007). van de Geer and Bühlmann (2009) called $RE_{1,0}(\xi,J)$ the restricted ℓ_1-eigenvalue.

Both types of quantities in (8), (9) and (25) involve minimum ratios of seminorms over cones. The main differences between them are the quantities used to bound the estimation error and the sign-restriction for $SCIF_{q,\ell}(\xi,J)$. Let $\{h^{(L)},h^{(D)},h\}$ be as in (14) and (15). While the RE in (25) uses the prediction regret $h'\Sigma h$ to bound the estimation error h, the (sign-restricted) CIF in (8) and (9) uses the ℓ_∞ norm of the gradient to bound the estimation error via (15). The use of the gradient bound in (8) and (9) provides sharper oracle inequalities by allowing the flexibility with the choice of v in (18) and (19).

For the ℓ_2 loss of the Dantzig selector, the oracle inequality of Bickel, Ritov, and Tsybakov (2009) and Theorem 3 can be compared as follows: for $z^* \le \lambda$

$$\|h^{(D)}\| \le \frac{2|J|^{1/2}\lambda}{CIF_{2,p}(1,J)} \le \frac{4|J|^{1/2}\lambda}{RE_{1,0}(1,J)RE_{2,p}(1,J)} \le \frac{4(1+\sqrt{|J|/\ell})|J|^{1/2}\lambda}{RE^2_{2,\ell}(1,J)}. \tag{26}$$

The right-hand side of (26) is the upper bound in (7.6) of Theorem 7.1 of Bickel, Ritov, and Tsybakov (2009). An application of the upper bound of van de Geer and Bühlmann (2009) on $\|h^{(D)}_J\|_1$ in a modification of the proof of Bickel, Ritov, and Tsybakov (2009) yields the intermediate upper bound. Theorem 3 provides the sharpest upper bound in (26). The second inequality in (26) follows from (18) with $\{\ell,v\} = \{p-|J|,u\}$. The third inequality in (26) follows from $RE_{1,0}(\xi,J) \ge RE_{2,\ell}(\xi,J)$ as in van de Geer and Bühlmann (2009) and $(1+\xi|J|/\ell)^{1/2}RE_{2,p}(\xi,J) \ge RE_{2,\ell}(\xi,J)$ by the shifting inequality in Candes and Tao (2007). The shifting inequality of Cai, Wang, and Xu (2010) can be used to reduce the factor $(1+\sqrt{|J|/\ell})$ to $\sqrt{1+|J|/(4\ell)}$ in (26), but (26) still holds. Similarly, for $z^* \le \lambda/2$,

$$\|h^{(L)}\| \le \frac{(3/2)|J|^{1/2}\lambda}{SCIF_{2,p}(3,J)} \le \frac{(3/2)|J|^{1/2}\lambda}{RE_{1,0}(3,J)RE_{2,p}(3,J)} \le \frac{4(1+3\sqrt{|J|/\ell})|J|^{1/2}\lambda}{RE^2_{2,\ell}(3,J)}, \tag{27}$$

with (7.10) of Theorem 7.2 of Bickel, Ritov, and Tsybakov (2009) on the right-hand side. The differences of the upper bounds in (26) and (27) could be nontrivial since van de Geer and Bühlmann (2009) showed by example the possibility of $RE_{2,0}(\xi,J)/RE_{1,0}(\xi,J) \to 0$.

A significant difference between the (sign-restricted) CIF- and RE- based oracle inequalities is their relationships to the oracle inequalities of Candes and Tao (2007) and Zhang (2009b). While

Theorem 3 is sharper than these inequalities, the RE-based oracle inequalities seem not to have this property.

The flexibility with the choice of v in (18) and (19) is a significant advantage for (8) and (9). The square of the potentially small $(\|\Sigma_A^{1/2}u_A\| - \|\Sigma_{A^c}^{1/2}u_{A^c}\|)_+$ has been used to bound $u'\Sigma u = \|Xu\|^2/n$ from below in the proofs of RE-based oracle inequalities. This is not needed in Candes and Tao (2007) and Zhang (2009b) since their arguments correspond to using vectors v with $v_{A^c} = 0$ in (18) and (19). For example, (23) with $\widetilde{F}_3$ is at least by a factor $\{1 - \min(\widetilde{F}_1, \widetilde{F}_2, \widetilde{F}_3)\}_+$ sharper than inserting Lemma 4.1 (ii) into (7.6) in Bickel, Ritov, and Tsybakov (2009), even after an application of sharper shifting inequalities to their Lemma 4.1 (ii). For $q > 2$, the ratio in (18) is not properly length normalized with $v = u$, as discussed in Remark 6. Thus, direct extensions of (25) with $u'\Sigma u$ in the numerator may not yield performance bounds of the right order.

Corollary 7 sharpens the oracle inequality of Candes and Tao (2007). The upper bound (23) with $\widetilde{F}_1$ is of the sharpest form among the three for large ℓ due to the factor $1/\ell$, compared with $\sqrt{1/(4\ell)}$ for $\widetilde{F}_2$ and $\widetilde{F}_3$. For $\ell = k/4$ and $q = 2$, (23) with $\widetilde{F}_2$ yields

$$z^* \le \lambda \;\Rightarrow\; \|\widehat{\beta}^{(D)}(\lambda) - \beta^*\| \le \frac{2\sqrt{k}\lambda}{CIF_{2,p}(1,J)} \le \frac{\sqrt{10k}\lambda}{(1 - \delta^-_{1.25k} - \theta^{(2)}_{k,1.25k})_+}, \tag{28}$$

a slightly sharper version of Cai, Wang, and Xu (2010) improvement of Candes and Tao (2007) result due to $\delta^-_{1.25k} \le \delta_{1.25k}$. Similarly, (24) with $\xi = \sqrt{2}$ and $\ell = k/2$ yields

$$z^* \le \lambda \;\Rightarrow\; \|\widehat{\beta}^{(L)}\big((1+\sqrt{2})^2\lambda\big) - \beta^*\| \le \frac{(1+\sqrt{2})\sqrt{8k}\lambda}{SCIF_{2,p}(\sqrt{2},J)} \le \frac{(1+\sqrt{2})4\sqrt{k}\lambda}{(1 - \delta^-_{1.5k} - \theta^{(2)}_{2k,1.5k})_+}. \tag{29}$$

The right-hand sides of (28) and (29) are directly comparable with the Candes and Tao (2007) inequality $\|\widehat{\beta}^{(D)}(\lambda) - \beta^*\| \le 4\sqrt{k}\lambda/(1 - \delta_{2k} - \theta^{(2)}_{k,2k})_+$ in the same event.

Another option is to apply (17), (18) and (21) with $(r,s) = (\infty, q)$ to (10), resulting in

$$\|\widehat{\beta}^{(D)}(\lambda) - \beta^*\|_q \le \frac{2|J|^{1/q}\lambda}{CIF_{q,p}(1,J)} \le \max_{A,B} \frac{2\lambda\|\Sigma_A^{-1}\|_{\infty,q}(1 + (a_q k/\ell)^{q-1})^{1/q}}{\{1 - (k^{1-1/q}/\ell)\|\Sigma_A^{-1}\Sigma_{A,B}\|_{\infty,q}\}_+} \tag{30}$$

for all $1 \le q \le \infty$ in the event $z^*_\infty \le \lambda$, where $k = |J|$ and A and B are as in (16). Inequality (30) and the combination of (11) and (19) with $w_A \propto \mathrm{sgn}(u_A)|u_A|^{q-1}$ are related to the results of Zhang (2009b). For $\|\beta^*\|_0 = k$, his oracle inequality for the Lasso can be written as

$$\big\|\widehat{\beta}^{(L)}(\lambda/t^*) - \beta^*\big\|_q \le \frac{32(1+F^*)G^*}{\widetilde{C}(1-F^*)_+^2} \quad \text{for } z^*_\infty \le \lambda,\ 1 \le q \le \infty,\ \ell \ge k, \tag{31}$$

with $F^* = \max_{A,B,w}(k^{1-1/q}\|\Sigma_{B,A}w_A\|_1/\ell)$, $G^* = \widetilde{C}\lambda(k+\ell)^{1/q}\max_{A,w}\|w_A\|_{q/(q-1)}$ for $w \propto \Sigma_A^{-1}\mathrm{sgn}(u_A)|u_A|^{q-1}$, $G^* = \widetilde{C}\lambda k^{1/q}\max_{A,w}\|w_A\|_{q/(q-1)}$ for $w_A \propto \mathrm{sgn}(u_A)|u_A|^{q-1}$, $\widetilde{C} = (1 + 1/t^*)$ and $t^* = (1-F^*)/\{4(1+F^*)\}$, where $w_A'\Sigma_A u_A = \|u_A\|_q = 1$ and $\{A, B, u_A\}$ are as in (16). It turns out that the combination of (11) and (19) with the same choice of w and $(r, a_r) = (\infty, 1)$ is at least by a factor of $5/12$ sharper than (31), even with the suboptimal $a_q = 1$. For small k/ℓ, Zhang (2009b) pointed out the smaller order $k^{1/q}\lambda$ of G^* for $w_A \propto \mathrm{sgn}(u_A)|u_A|^{q-1}$ as an advantage of the Lasso, compared with the order $(k+\ell)^{1/q}\lambda$. The cost of this advantage is the square of $(1-F^*)_+$ in the

denominator of (31), compared with (23) and (30) for the Dantzig selector. Moreover, the error bound in (23) for the Dantzig selector is also of the order $k^{1/q}\lambda$ for $q \le 2$ with much smaller constant factors, and the difference between $k^{1/q}$ and $(k+\ell)^{1/q}$ diminishes for large q as in Theorem 8. Thus, the advantage of the Lasso in this aspect has some limitations.

We have proved that Theorem 3 sharpens and unifies a number of existing oracle inequalities for the Lasso and Dantzig estimators, so that they can all be viewed as (possibly more explicit) upper bounds for the right-hand sides of (10) and (11). The choice $r = \infty$ in (16), and consequently in (18), (19), (20) and (21), typically gives oracle inequalities of the sharpest form involving the dimension-normalized $\|\cdot\|_{\infty,q}$ norm as in (30), compared with the typical $\|\cdot\|_{q,q}$ norm in the literature. Oracle inequalities for $\beta^*_{J^c} \ne 0$ are given in Section 4. Although (10) and (11) are of the same format, the Dantzig selector requires smaller λ and smaller $\xi = 1$. This theoretical advantage of the Dantzig selector for $z^*_\infty \le \lambda$ reverses if $\|\widehat{\beta}^{(D)}(\lambda)\|_1 \le \|\beta^*\|_1$ is replaced by $\|\widehat{\beta}^{(D)}(\lambda)\|_1 \le \|\widehat{\beta}^{(L)}(\lambda)\|_1$ for $z^*_\infty > \lambda$.

3.3 Vaiable Selection

Variable selection is fundamental for the interpretation of models with hign-dimensional data. Meinshausen and Bühlmann (2006), Tropp (2006), Zhao and Yu (2006), and Wainwright (2009) proved that the Lasso is variable selection consistent under a strong irrepresentable condition and some other regularity conditions on X and β. Candes and Plan (2009) proved the selection consistency of the Lasso under random permutation and sign-change of regression coefficients and a mild condition on the maximum absolute correlation among design vectors. Consistent variable selection in linear regression can be achieved with a concave penalty (Fan and Li, 2001; Zhang, 2010) or adaptive Lasso (Zou, 2006; Huang, Ma, and Zhang, 2008), without requiring the strong irrepresentable condition.

The ℓ_∞ error bounds in Theorem 3 (i) and their more explicit versions, for example, (30) with $q = \infty$, naturally lead to variable selection by thresholding the Lasso or Dantzig selector. We focus on the Dantzig selector here although parallel results can be obtained for the Lasso in the same way. Zhang (2009b) studied the selection consistency of thresholding the Lasso through his upper bounds for $\|\widehat{\beta}^{(L)} - \beta^*\|_\infty$. Lounici (2008) considered thresholding either the Lasso or the Dantzig selector under the stronger condition $\max_{j\ne k}|(\Sigma)_{jk}| \le 1/\{\alpha(1+2\xi)|J|\}$ with $\alpha > 1$, $(\Sigma)_{jj} = 1$, $\xi = 3$ for the Lasso, and $\xi = 1$ for the Dantzig selector.

Candes and Tao (2007) considered the Gauss-Dantzig selector

$$\widehat{\beta}^{(GD)} = \arg\min_b \Big\{ \|y - Xb\| : |\widehat{\beta}_j| \le \lambda' \Rightarrow b_j = 0, \forall j, \widehat{\beta} = \widehat{\beta}^{(D)}(\lambda) \Big\}. \tag{32}$$

For threshold functions $t(x;\lambda)$ satisfying $\{x : t(x;\lambda) = 0\} = \{x : |x| \le \lambda\}$ and $xt(x;\lambda) \ge 0$, define the threshold Dantzig selector as

$$\widehat{\beta}^{(TD)} = t\big(\widehat{\beta}^{(D)}(\lambda);\lambda'\big). \tag{33}$$

This includes the hard $t(x;\lambda) = xI\{|x| > \lambda\}$ and the soft $t(x;\lambda) = \mathrm{sgn}(x)(|x| - \lambda)_+$. Define

$$\widehat{\beta}^{(oracle)} = \arg\min_b \Big\{ \|y - Xb\| : \beta_j = 0 \Rightarrow b_j = 0, \forall j \Big\}.$$

Theorem 8 *Suppose (3) holds with $J = \{j : \beta_j \neq 0\}$. Let $\widehat{\beta}^{(GD)}$ and $\widehat{\beta}^{(TD)}$ be as in (32) and (33) respectively with the universal penalty level $\lambda = \lambda_{univ} = \sigma\sqrt{(2/n)\log p}$ and a threshold level λ' satisfying $2\lambda_{univ}/CIF_{\infty,p}(1,J) \le \lambda' < \min_{\beta_j\neq 0}|\beta_j|/2$. Then,*

$$P_{\beta,X}\Big\{ sgn(\widehat{\beta}^{(TD)}) \neq sgn(\beta) \text{ or } \widehat{\beta}^{(GD)} \neq \widehat{\beta}^{(oracle)} \Big\} \le 1/\sqrt{\pi\log p} \to 0. \tag{34}$$

Remark 9 *If we use (21) in Theorem 8, $(|J|/\ell)\max_{A,B}\|\Sigma_A^{-1}\Sigma_{A,B}\|_{\infty,\infty} < 1$ becomes a basic condition for (34). Meanwhile, the strong irrepresentable condition for the selection consistency of the Lasso without post-thresholding is $\|\Sigma_{J^c,J}\Sigma_J^{-1}\|_{\infty,\infty} < 1$. Compared with Lounici (2008), we improve the factor $1+2\xi$ to $1+\xi$ via*

$$CIF_{\infty,p}(\xi,J) \ge \min_{j\in J}(\Sigma)_{jj} - (|J|-1+\xi|J|)\max_{j\neq k}|(\Sigma)_{jk}|.$$

4. Upper Bounds for the ℓ_q Loss in ℓ_r Balls

We divide this section into two subsections. The first subsection provides non-probabilistic oracle inequalities: conditions on the data (X,y) and a target coefficient vector β^* for upper bounds of $\|\widehat{\beta}-\beta^*\|_q$ for the Lasso and Dantzig selector. The second subsection provides sufficient condition on the design X for the rate minimaxity for the ℓ_q loss in ℓ_r balls under $P_{\beta,X}$.

4.1 Oracle Inequalities

The oracle inequalities here differ from Theorem 3 (i) by allowing target vectors with many small entries and smaller penalty levels.

Our first theorem deals with the usual $\lambda \ge z_\infty^*$ and allows targets β^* with small $\|\beta^*_{J^c}\|_1$ for a certain set $J \subset \{1,\ldots,p\}$. The effect of the elements of β^* in J^c is controlled by

$$M_q(\lambda,\rho) = \sup\Big\{\|u\|_q : \|\Sigma u\|_\infty \le \lambda, \|u\|_1 \le \rho\Big\}. \tag{35}$$

Theorem 10 *Let β^* be a target vector, $q \in [1,\infty]$, $J \subset \{1,\ldots,p\}$ and $\rho_J = \|\beta^*_{J^c}\|_1$. Then,*

$$\|\widehat{\beta}^{(D)} - \beta^*\|_q \le \max\Big\{\frac{2|J|^{1/q}\lambda}{CIF_{q,p}(\xi,J)}, 2M_q\Big(\lambda, \frac{\xi+1}{\xi-1}\rho_J\Big)\Big\},\ \forall \xi > 1, \tag{36}$$

in the event $z_\infty^* = \|X'(y-X\beta^*)/n\|_\infty \le \lambda$. Moreover, for $z_\infty^* \le \lambda(\xi_0-1)/(\xi_0+1)$,

$$\|\widehat{\beta}^{(L)} - \beta^*\|_q \le \max\Big\{\frac{\xi_1|J|^{1/q}\lambda}{CIF_{q,p}(\xi,J)}, M_q\big(\xi_1\lambda, \xi_2\rho_J\big)\Big\},\ \forall \xi > \xi_0, \tag{37}$$

where $CIF_{q,\ell}(\xi,J)$ is as in (8), $\xi_1 = 2\xi_0/(\xi_0+1)$ and $\xi_2 = (1+\xi_0)(1+\xi)/(\xi-\xi_0)$.

Remark 11 *The first component of (36) and (37) can be viewed as the cost of estimating the large components β^* in J without knowing J, and the second component the cost of having potentially many small elements of β^* in J^c. Since $M_q(\lambda,\rho) \le \rho$ for $q \ge 1$, $\beta^*_{J^c}$ does not contribute to the order of the error bounds in (36) and (37) when $\rho_J \lesssim |J|^{1/q}\lambda/CIF_{q,p}(\xi,J)$. In Proposition 15 below, we provide conditions for $M_q(\lambda,\rho) \lesssim \lambda(\rho/\lambda)^{1/q}$, as if $\Sigma = I$.*

Our next theorem deals with smaller penalty levels satisfying $z^*_{2,d} \le \lambda < z^*_\infty$, where

$$z^*_{q,d} = \max_{|A|=d} z^*_{q,A},\ z^*_{q,A} = \|X'_A(y - X\beta^*)/n\|_q/|A|^{1/q}, \quad z^*_\infty = z^*_{\infty,1}. \tag{38}$$

Since $z^*_{q,d}$ is the length normalized ℓ_q norm of the d largest absolute values of the elements of $z = X'(y - X\beta^*)/n$, $z^*_{q,d}$ is increasing in q and decreasing in d, and $z^*_{q,d} \le z^*_\infty$. Let

$$\mathscr{C}_{relax}(\xi, J, d) = \Big\{ u : \|u_{J^c}\|_1 \le \xi d^{1/2} \max_{|A|=d, A\supseteq J} \|u_A\| \Big\} \tag{39}$$

as a relaxed cone, and define the corresponding relaxed CIF as

$$CIF_{q,relax}(\xi, J, d) = \inf_{u\in\mathscr{C}_{relax}(\xi,J,d)} \Big\{ \frac{d^{1/q}\|\Sigma_{A,*}u\|}{d^{1/2}\|u\|_q} : A = \operatorname*{arg\,max}_{|B|=d, B\supseteq J} \|u_B\| \Big\}. \tag{40}$$

The following quantity plays the role of (35) for relaxed cones:

$$M_{q,relax}(\lambda, \rho, J, d) = \sup_{\|u\|_1\le\rho} \Big\{ \|u\|_q : \|\Sigma_{A,*}u\| \le d^{1/2}\lambda, A = \operatorname*{arg\,max}_{|B|=d, B\supseteq J} \|u_B\| \Big\}. \tag{41}$$

Since $\|u_J\|_1 \le |A|^{1/2}\|u_A\|$ for $A \supseteq J$, the relaxed cone (39) is larger than the cone in (8). Moreover, since $\|\Sigma_{A,*}u\|/|A|^{1/2} \le \|\Sigma u\|_\infty$,

$$CIF_{q,relax}(\xi, J, d) \le (d/|J|)^{1/q} CIF_{q,p}(\xi, J),\ M_{q,relax}(\lambda, \rho, J, d) \ge M_q(\lambda, \rho).$$

Theorem 12 *Let β^* be a target vector, $q \in [1,\infty]$, $J \subset \{1,\dots,p\}$ with $(4|J|/3) \vee 1 \le d \le p$, $\rho_J = \|\beta^*_{J^c}\|_1$ and $z^*_{2,d}$ be as in (38). Then, for $z^*_{2,d} \le \lambda(\xi_0 - 1)/(\xi_0 + 1)$,*

$$\|\widehat{\beta}^{(L)} - \beta^*\|_q \le \max\Big\{ \frac{\xi_1 d^{1/q}\lambda}{CIF_{q,relax}(\xi, J, d)}, M_{q,relax}\Big(\xi_1\lambda, \xi_2\rho_J, J, d\Big) \Big\}, \tag{42}$$

where $\xi_1 = 2\xi_0/(\xi_0+1)$ and $\xi_2 = (\xi_0+1)(\xi+1)/(\xi-\xi_0)$.

Remark 13 *By (22) and (38), $(z^*_{2,d})^2/\{\sigma^2(1+\delta^+_d)\}$ is no greater than the maximum of $\binom{p}{d}$ χ^2_d variables under $P_{\beta^*,X}$ in (3), so that for certain $\lambda \asymp \sigma\{(1+\delta^+_d)(2/n)\log(p/d)\}^{1/2}$, $z^*_{2,d} \le \lambda(\xi_0 - 1)/(\xi_0+1)$ with large probability. Thus, for $|J| = k \asymp d$ and $\log(p/d) \ll \log p$, Theorem 12 allows $\lambda \ll \lambda_{univ} = \sigma\{(2/n)\log p\}^{1/2}$. Zhang (2010) derived similar oracle inequalities for the Lasso, MC+, and other concave penalized least squares estimators at the same λ under the sparse Riesz condition $|J| \le d/\{(1+\delta^+_d)/(1-\delta^-_d) + 1/2\}$.*

Remark 14 *Since $\|u_A\|/d^{1/2}$ does not control $\|u_A\|_q/d^{1/q}$ for $q > 2$ and large $|A| = d$, the relaxed constant factors in (40) and (41) are not properly normalized for $q > 2$ and $\Sigma = I$. Thus, Theorem 12 is most useful when $1 \le q \le 2$, although it is valid for all $1 \le q \le \infty$.*

We use the following quantities to bound the constant factors in Theorems 10 and 12:

$$\begin{aligned} \eta_{q,d} &= \max_{|A|=d} \|\Sigma_A^{-1}\|_{\infty,q}/d^{1/q},\ \eta^*_{q,d} = \max_A \|\Sigma_A^{-1}\|_{q,q}, \\ \kappa_{q,d,\ell} &= \max_{|A|=d} \min_{r\ge1} \max_{|B|=\ell} \ell^{1-1/q}(a_r/\ell)^{1-1/r}\|\Sigma_A^{-1}\Sigma_{A,B}\|_{r,q}. \end{aligned} \tag{43}$$

Proposition 15 *Let $CIF_{q,relax}(\xi,J,\ell)$, $M_q(\lambda,\rho)$ and $M_{q,relax}(\lambda,\rho,J,\ell)$ be as in (40), (35) and (41) respectively. Let $\{\eta_{q,d},\eta^*_{q,d},\kappa_{q,d,\ell}\}$ be as in (43) and $a_q=(1-1/q)/q^{1/(q-1)}$. Then,*

$$M_q(\lambda,\rho)\le \eta_{q,k}k^{1/q}\lambda+(\kappa^q_{q,k,k}+a_q^{q-1})^{1/q}k^{1/q-1}\rho,\ \forall\, k\ge 1, 1\le q\le\infty, \tag{44}$$

$$M_{q,relax}(\lambda,\rho,J,d)\le \eta^*_{2,d}d^{1/q}\lambda+\{(d/\ell)^{1-q/2}\kappa^q_{2,d,\ell}+a_q^{q-1}\}^{1/q}\ell^{1/q-1}\rho,\ 1\le q\le 2, \tag{45}$$

with $\ell=d-|J|$, and with $C_{q,2}(\xi,t)$ and $\widetilde{\phi}^*_{2,\ell}(\xi,J)$ as in (17) and (20),

$$CIF_{q,p}(\xi,J)\ge CIF_{q,relax}(\xi,J,\ell)\ge \frac{(d/|J|)^{1/q-1/2}\widetilde{\phi}^*_{2,\ell}(\xi\sqrt{d/|J|},J)}{C_{q,2}\big(\xi(d/|J|)^{1/2},|J|^2/(d\ell)\big)}. \tag{46}$$

Remark 16 *Suppose that the quantities in (43) are bounded whenever invoked. For $\rho/\lambda\asymp k\asymp\ell$, (44) and (45) give the rate $M_q(\lambda,\rho)\lesssim\rho(\lambda/\rho)^{1-1/q}$ and $M_{q,relax}(\lambda,\rho,J,d)\lesssim\rho(\lambda/\rho)^{1-1/q}$, the same as the simplest case $\Sigma=I$. Since (46) is of the form (20), Corollary 7 can be automatically extended under the setting of Theorem 12.*

4.2 Rate ℓ_q Minimax Estimation in ℓ_r Balls

We present sufficient conditions for the rate minimaxity of the Lasso and Dantzig selector in ℓ_r balls in (4) in the sense of (6) and (7). Let λ_{univ} and λ_{mm} be as in (5). We first consider the ℓ_q risk.

Theorem 17 *Let $q\ge 1\ge r>0$. Suppose $(\log p)/n=O(1)$ and $R^r/\lambda^r_{mm}\asymp d\le n\wedge p$ for some integer $d\to\infty$ satisfying $(\log d)/\log p\le c_0<1$. Let $0<\alpha_0<1$ and $\widehat{\beta}$ be either the Lasso or Dantzig selector with $\lambda=\lambda_{univ}/\alpha_0$. Suppose $p^{1-(\alpha_1/\alpha_0)^2}(n^q/d+d^{q/r-1})\le 1$ and $p^{1-(\alpha/\alpha_0)^2}d^{q/r-q}\le 1$ for certain $\{\alpha,\alpha_1\}\subset(\alpha_0,1)$. For the Dantzig selector, let $\xi>1$, $\xi^*=1$ for $r=1$ and $\xi^*=(\xi+1)/(\xi-1)$ for $r<1$. For the Lasso, let $\xi>(1+\alpha)/(1-\alpha)$ $\xi^*=1/(1-\alpha_1)$ for $r=1$ and $\xi^*=(\xi+1)/\{\xi-1-\alpha(\xi+1)\}$ for $r<1$. Then,*

$$\frac{\sup_{\beta\in\Theta_{r,R}}E_{\beta,X}\|\widehat{\beta}-\beta\|_q^q}{R^r\lambda_{mm}^{q-r}}\lessapprox\Big[\max\Big\{\frac{C_1I\{r<1\}}{CIF_q^*(\xi,d)},C_1M_q\Big(\frac{1}{d^{1/q}},\frac{\xi^*C_2}{d^{1/q-1}}\Big)\Big\}\Big]^q, \tag{47}$$

where $C_1=2(d\lambda^r_{mm}/R^r)^{1/q}/(\alpha_0\sqrt{1-c_0})$, $C_2=\alpha_0\sqrt{1-c_0}R/(d^{1/r}\lambda_{mm})$, $CIF_q^*(\xi,d)$ is as in (12) and $M_q(\lambda,\rho)$ is as in (35). Consequently, if either $\eta_{q,d}+\kappa_{q,d,d}=O(1)$ with the $\{\eta_{q,d},\kappa_{q,d,d}\}$ in (43) or $M_q(d^{-1/q},d^{1-1/q})+I\{r<1\}/CIF_q^*(\xi,d)=O(1)$, then

$$\sup_{\|\beta\|_r\le R}E_{\beta,X}\|\widehat{\beta}-\beta\|_q^q\lesssim\inf_{\delta}\sup_{\|\beta\|_r\le R}E_{\beta,X}\|\delta(X,y)-\beta\|_q^q. \tag{48}$$

Remark 18 *For $q=2$, $\eta_{q,k}+\kappa_{q,k,k}=O(1)$ for some $k\asymp d$ if the sparse Riesz condition holds Zhang and Huang (2008), that is, $1/(1-\delta_d^-)+\delta_d^+=O(1)$ for the $\delta_d^\pm$ in (22). For $p\gg n$, random matrix theory can be applied to validate such conditions up to $d\asymp n/\log(p/n)$.*

Theorem 17 differs from existing results by directly comparing the ℓ_q risk of estimators with the minimax risk, instead of finding upper bounds for the ℓ_q loss. It is based on the oracle inequality for $\lambda>\lambda_{univ}$ in Theorem 10. However, in practice, a penalty level $\lambda<\lambda_{univ}$ is often empirically the best choice. As we mentioned in Remark 2, $\lambda_{mm}<\lambda_{univ}$ iff $R>\sigma/\sqrt{n}$. For $\lambda_{mm}/\lambda_{univ}=o(1)$,

oracle inequalities requiring penalty levels $\lambda \geq \lambda_{univ}$ do not match the order of the minimax lower bounds in Theorem 1. For example, when $p = n\log n$ and $d \asymp R^r/\lambda_{mm}^r \asymp n/\log\log n$, $\lambda_{mm}/\lambda_{univ} \to 0$ as $n \to \infty$ and the regularity conditions on X may still hold. Theorem 19 below closes this gap by providing the rate minimaxity of the Lasso in the quantiles of the ℓ_q loss with $\lambda \asymp \lambda_{mm} = o(\lambda_{univ})$. Define

$$CIF^*_{q,relax}(\xi,k,d) = \inf_{|A|=d}\Big\{\frac{d^{1/q}\|\Sigma_{A,*}u\|}{d^{1/2}\|u\|_q} : \min_{|A\setminus J|=d-k}\|u_{J^c}\|_1 < \xi d^{1/2}\|u_A\|\Big\}, \tag{49}$$

$$M^*_{q,relax}(\lambda,\rho,k,d) = \sup_{|A|=d}\Big\{\|u\|_q : \|\Sigma_{A,*}u\| \leq d^{1/2}\lambda, \min_{|A\setminus J|=d-k}\|u_{J^c}\|_1 \leq \rho\Big\}. \tag{50}$$

Theorem 19 *Let* $\lambda = \min\big(\lambda_{univ}, (1+\varepsilon_0)(1+\delta_d^+)^{1/2}\lambda_{mm}\big)/\alpha$ *with* $0 < \varepsilon_0 \leq \alpha < 1$ *and* $\{\lambda_{mm}, \lambda_{univ}, \delta_d^+\}$ *in (5) and (22). Let* $0 < r \leq 1 \leq q \leq 2$. *Suppose* $n \wedge p \geq d \asymp R^r/\lambda_{mm}^r \to \infty$, $\delta_d^+ = O(1)$ *and* $\lambda_{mm}n^{1/2}/\sigma \to \infty$. *Suppose that for certain* $k+\ell = d$ *with* $k \asymp \ell$,

$$\max\Big\{1/CIF^*_{q,relax}(\xi,k,d), M^*_{q,relax}(d^{-1/2}, d^{1/2}, k, d)\Big\} = O(1).$$

Then, the Lasso is rate minimax in ℓ_r *balls in the sense that for all* $\varepsilon > 0$,

$$\begin{aligned} &\inf\Big[t : \sup_{\|\beta\|_r \leq R} P_{\beta,X}\Big\{\|\widehat{\beta}^{(L)} - \beta\|_q^q \geq t^q R^r \lambda_{mm}^{q-r}\Big\} \leq \varepsilon\Big] \\ \lesssim\ &\inf\Big[t : \inf_{\delta}\sup_{\|\beta\|_r \leq R} P_{\beta,X}\Big\{\|\delta(X,y) - \beta\|_q^q \geq t^q R^r \lambda_{mm}^{q-r}\Big\} \leq \varepsilon\Big] < \infty. \end{aligned} \tag{51}$$

In particular, (51) holds if $1/(1-\delta_d^-) + \delta_d^+ = O(1)$ *as in Remark 18.*

The quantities (8), (12), (35), (40), (41), (49) and (50) are best understood by comparisons with functions of (22) and (43) via Propositions 5 and 15. These quantities also facilitate comparisons between our and existing upper bounds on the loss as in the derivation of Corollary 7. In such comparisons, the Hölder inequality and (22) give

$$\eta_{q,d} \leq \eta^*_{q,d},\ \ \kappa_{q,d,\ell} \leq \kappa^*_{q,d,\ell},\ \ \eta^*_{2,d} = 1/(1-\delta_d^-),\ \ \kappa^*_{2,d,\ell} \leq \theta_{d,\ell}\eta^*_{2,d},$$

where $\kappa^*_{q,d,\ell} = a_q^{1-1/q}\max_{A,B}\|\Sigma_A^{-1}\Sigma_{A,B}\|_{q,q}$ with $|A| = d$, $|B| = \ell$ and $A \cap B = \emptyset$.

5. Discussion

Although this paper focuses on the estimation of regression coefficients, the estimation of $X\beta^*$ (prediction) is an important problem (Greenshtein and Rotiv, 2004). Similar to the proof of (11), (9), (14) and (15) imply

$$\|X\widehat{\beta}^{(L)} - X\beta^*\|^2/n + 2\lambda\|(\widehat{\beta} - \beta^*)_J\|_1/(\xi+1) \leq \frac{\{2\xi/(\xi+1)\}|J|\lambda^2}{SCIF_{1,0}(\xi,J)} \tag{52}$$

in the event $z^*_\infty = \|X'(y - X\beta^*)/n\|_\infty \leq \lambda(\xi-1)/(\xi+1)$. Since $SCIF_{1,0}(\xi,J) \geq RE_{1,0}(\xi,J)$, (52) implies Lemma 2.1 of van de Geer and Bühlmann (2009) for $(z^*_\infty, \xi) = (0,1)$.

As we have explained in Section 3.2, the use of $\|\Sigma u\|_\infty$ in the numerator of (8) and (9) seems necessary to ensure the dominance of Theorem 3 over the oracle inequalities of the type (30). However, if we make the numerator quadratic in u, Corollary 7 still holds up to a constant factor with the following weak CIF:

$$CIF^w_{q,\ell}(\xi,J) = \inf_{u\in\mathscr{C}(\xi,J)} \frac{|J|^{1/q}u'_A\Sigma_{A,*}u}{\|u_J\|_1\|u_A\|_q},\ SCIF^w_{q,\ell}(\xi,J) = \inf_{u\in\mathscr{C}_-(\xi,J)} \frac{|J|^{1/q}u'_A\Sigma_{A,*}u}{\|u_J\|_1\|u_A\|_q}, \tag{53}$$

where $A = \arg\max_{|A\setminus J|\le\ell}\|u_A\|$. For example, (26), (27) and (28) are still consequences of Theorem 3 when (53) is used instead of (8) and (9).

Since the oracle inequalities in this paper apply directly to data points (X,y) and target vectors β^*, the normality assumption on the error in (3) is not crucial for the upper bounds for the estimation risk and loss (not even the condition $E_{\beta,X}y = X\beta$). For example, for the estimation of a target β^* with $X\beta^* \approx Ey$, the upper bounds in Theorem 19 are valid for $\|\widehat{\beta}^{(L)} - \beta^*\|_q^q$ with large probability under P and $\sigma = \sigma_1 + \sigma_2$, provided that

$$\begin{cases} E\exp(v'X'(y-Ey)) \le \exp(-n\sigma_1^2 v'\Sigma v/2), \\ \max_{|A|=\ell}\|P_A(Ey - X\beta^*)\| \le \sigma_2\sqrt{2\ell\log(p/\ell)}. \end{cases}$$

For design matrices X with iid sub-Gaussian rows, our results can be extended to β in ℓ_r balls with $1 < r \le 2$ due to $\sigma_2 \le O(\|\beta_{J^c}\|)$ for $\beta^*_J = \beta_J$ and $\beta^*_{J^c} = 0$.

The proofs in this paper do not completely deal with the most difficult case of $q > 2$ and $\lambda_{mm} = o(\lambda_{univ})$. For example, an extension of Theorem 12 to $q > 2$ seems to require sharp upper bounds for $z^*_{q,d}$ in (38).

For $\lambda < \lambda_{univ}$, the proof of Theorem 12 can be extended to the Dantzig selector with the feasibility of β^* replaced by the feasibility of $\widehat{\beta}^{(L)}$. This would yield slightly worse error bounds than those in Theorem 12. However, if we modify the Dantzig selector as

$$\widetilde{\beta} = \arg\min_b \left\{\|b\|_1 : \max_{|A|=d}\|X'_A(y - Xb)\| \le \sqrt{d}\lambda\right\}, \tag{54}$$

the feasibility of β^* would be guaranteed in the event $z^*_{2,d} \le \lambda$ even for $\lambda = o(\lambda_{univ})$ as in Theorems 12 and 19. This will provide sharper error bounds for the smaller λ and $q \le 2$. We omit this modification since the computational issues with (54) is not clear for $d > 1$.

Acknowledgments

This project is partially supported by the National Science Foundation (NSF grants DMS-0604571, DMS-0804626, DMS-0906420) and the National Security Agency (NSA grant H98230-09-1-0006).

Appendix A. Proofs

We provide all the proofs here. Lemms are stated and proved as needed.

Proof of Theorem 1. Let $\Theta = \Theta_{r,R}$ and $k = R^r/\lambda_{mm}^r$ for $r > 0$ and $\Theta = \Theta_{0,k}$ for $r = 0$. By (5), $\lambda_{mm}/\sigma_n = \sqrt{2\log(p/k) - r\log(\lambda_{mm}/\sigma_n)}$ for $r > 0$. Since $\lambda_{mm}/\sigma_n \to \infty$,

$$\lambda_{mm} = (1+o(1))\sigma_n\sqrt{2\log(p/k)},\ \min(k, p/k) \to \infty, \quad \forall\, r \geq 0. \tag{55}$$

Let $P_{\mu,w}$ be a (prior) probability distribution under which (z_j, β_j) are iid vectors with

$$z_j|\beta_j \sim N(\beta_j, \sigma_n^2),\ P_{\mu,w}\{\beta_j = \mu\} = w = 1 - P_{\mu,w}\{\beta_j = 0\},$$

where $\mu = \lambda_{mm}(1-\varepsilon)$ and $w = (1-\varepsilon)k/p$. Since $\widetilde{z}_j = x_j'(y - \sum_{i\neq j}\beta_i x_i)/n$ is sufficient for β_j given $(X, y, \beta_i, i \neq j)$ and $\widetilde{z}_j|\beta \sim z_j|\beta$, the minimum Bayes risk is bounded by

$$\begin{aligned}
\inf_{\widehat{\beta}} E_{\mu,w}E_{\beta,X}\|\widehat{\beta} - \beta\|_q^q &\geq E_{\mu,w}\sum_{j=1}^p \min_t E_{\beta,X}\Big[|t-\beta_j|^q\Big|X, y\Big] \\
&\geq E_{\mu,w}\sum_{j=1}^p \min_t E_{\beta,X}\Big[|t-\beta_j|^q\Big|X, y, \beta_i, i \neq j\Big] \\
&= E_{\mu,w}\sum_{j=1}^p \min_t E_{\beta,X}\Big[|t-\beta_j|^q\Big|z_j\Big] \\
&= (1+o(1))k\lambda_{mm}^q
\end{aligned} \tag{56}$$

as $k \wedge (p/k) \to \infty$ and then $\varepsilon \to 0$. The last step above is by Donoho and Johnstone (1994).

Let $N = \#\{j : \beta_j \neq 0\}$. Under $P_{\mu,w}$, $\|\beta\|_r^r = N\mu^r$ and $\|\beta\|_q^q = N\mu^q$, so that

$$\beta \in \Theta \ \Leftrightarrow\ N \leq k/(1-\varepsilon)^r \ \Rightarrow\ \|\beta\|_q^q \leq k\lambda_{mm}^q,\ \forall\, r \geq 0, \tag{57}$$

due to $R^r/\mu^r = k/(1-\varepsilon)^r$ and $\{k/(1-\varepsilon)^r\}\mu^p/\lambda_{mm}^p = k(1-\varepsilon)^{q-r} \leq k$ for $r > 0$. Let

$$\delta^* = \arg\min_\delta E_{\mu,w}E_{\beta,X}\big[\|\delta(X,y) - \beta\|_q^q\big|X, y, \beta \in \Theta\big].$$

Since the conditional Bayes risk of δ^* is no greater than the minimax risk in Θ,

$$\begin{aligned}
& (1+o(1))k\lambda_{mm}^q \\
\leq\ & E_{\mu,w}E_{\beta,X}\big[\|\delta^* - \beta\|_q^q\big|\beta \in \Theta\big] + E_{\mu,w}E_{\beta,X}\|\delta^* - \beta\|_q^q I\{\beta \notin \Theta\} \\
\leq\ & \mathscr{R}(\Theta; X) + 2^{(q-1)_+}E_{\mu,w}E_{\beta,X}\big(\|\delta^*\|_q^q + \|\beta\|_q^q\big)I\{\beta \notin \Theta\}.
\end{aligned} \tag{58}$$

Since $E_{\mu,w}E_{\beta,X}[\|\delta^* - \beta\|_q^q|X, y, \beta \in \Theta] \leq E_{\mu,w}[\|\beta\|_q^q|X, y, \beta \in \Theta] \leq k\lambda_{mm}^q$ by (57), $\|\delta^*\|_q^q \leq 2^{(q-1)_+}k\lambda_{mm}^q$ a.s. Thus, since $N \sim \text{Binomial}(p, w)$ with $pw = (1-\varepsilon)k \to \infty$,

$$\begin{aligned}
& E_{\mu,w}E_{\beta,X}\big(\|\delta^*\|_q^q + \|\beta\|_q^q\big)I\{\beta \notin \Theta\} \\
\leq\ & 2^{(q-1)_+}k\lambda_{mm}^q P_{\mu,w}\{N > wp/(1-\varepsilon)\} + \mu^q E_{\mu,w}NI\{N > wp/(1-\varepsilon)\} \\
=\ & o(1)k\lambda_{mm}^q
\end{aligned} \tag{59}$$

by (57). The combination of (58) and (59) gives (6).

Now consider the loss $L(\delta, \beta) = I\{\|\delta - \beta\|_q > ck^{1/q}\lambda_{mm}\}$ in (7). Define

$$\widehat{\beta} = \delta(X,y)I\Big\{\|\delta(X,y)\|_q \leq (1+c)k^{1/q}\lambda_{mm}\Big\}.$$

By (57), $\beta \in \Theta$ implies $\|\beta\|_q^q \le k\lambda_{mm}^q$ and

$$\begin{aligned}\|\widehat{\beta}-\beta\|_q^q &\le c^q k\lambda_{mm}^q I\Big\{\|\delta-\beta\|_q \le ck^{1/q}\lambda_{mm}\Big\} \\ &\quad +\Big(\|\beta\|_q+(1+c)k^{1/q}\lambda_{mm}\Big)^q I\Big\{\|\delta-\beta\|_q > ck^{1/q}\lambda_{mm}\Big\}.\end{aligned}$$

Since $\|\beta\|_q^q = N\mu^q \le N\lambda_{mm}^q$, it follows that

$$\begin{aligned}E_{\mu,w}E_{\beta,X}\|\widehat{\beta}-\beta\|_q^q &\le c^q k\lambda_{mm}^q + (2+c)^q k\lambda_{mm}^q \max_{\beta\in\Theta} E_{\beta,X}L(\delta(X,y),\beta) \\ &\quad +2^{q-1}\lambda_{mm}^q E_{\mu,w}\Big(N+(1+c)^q k\Big)I\{\beta\notin\Theta\}. \qquad (60)\end{aligned}$$

Since $E_{\mu,w}\Big(N+(1+c)^q k\Big)I\{\beta\notin\Theta\} = o(1)k$ by (57), (56) and (60) yield

$$\sup_{\beta\in\Theta} E_{\beta,X}L(\delta(X,y),\beta) \ge \frac{1-c^q+o(1)}{(2+c)^q}.$$

Since the $o(1)$ is uniform in the choice of $\delta(X,y)$, we find

$$\inf_{\delta}\sup_{\beta\in\Theta} P_{\beta,X}\Big\{\|\delta(X,y)-\beta\|_q^q > (1-\varepsilon)k\lambda_{mm}^q\Big\} \ge \frac{\varepsilon+o(1)}{3^q},\ \forall 0<\varepsilon<1.$$

This gives (7) and completes the proof. ■

Proof of Theorem 3. Part (i) follows from (14) and (15) as briefly explained in the paragraph below the statement of the theorem. For the Dantzig selector, $z^* \le \lambda$ implies (15) and the feasibility of β^* for the ℓ_∞ constraint in (2), and the feasibility of β^* implies (14). For the Lasso estimator, (14) and (15) follow from the Karush-Kuhn-Tucker conditions

$$\|x_j(y-X\widehat{\beta})/n\|_\infty \le \lambda, \quad \widehat{\beta}_j \ne 0 \ \Rightarrow\ x_j(y-X\widehat{\beta})/n = \mathrm{sgn}(\widehat{\beta}_j)\lambda.$$

Part (ii) follows from $P_{\beta,X}\{z_\infty^* > t\sigma/\sqrt{n}\} \le 2pP\{N(0,1)>t\}$. ■

In this paper and those cited in Section 3.2, tails of ℓ_q norms or inner products are bounded by shifting inequalities (Candes and Tao, 2007, Lemma 3.1). The following lemma combines and extends the sharp shifting inequalities of Cai, Wang, and Xu (2010) for $q=2$ and Ye and Zhang (2009) for $w'h$ with $q=\infty$.

Lemma 20 *Let $1 \le q \le \infty$ and $a_q = (1-1/q)/q^{1/(q-1)}$ with $a_\infty = 1$. Let $h \in \mathbb{R}^p$, $J \subset \{1,\ldots,p\}$ and A be the union of J and the indices of the ℓ largest $|h_j|$ with $j \notin J$, $1 \le \ell \le p-|J|$. Then, $\|h_{A^c}\|_q \le (a_q/\ell)^{1-1/q}\|h_{J^c}\|_1$. Moreover, for any vector $w \in \mathbb{R}^p$,*

$$\sum_{j\notin A} w_j h_j \le \|h_{J^c}\|_1\Big(\frac{a}{\ell}\vee\frac{a_q}{\ell}\Big)^{1-1/q}\max\Big\{\|w_B\|_{q/(q-1)} : B\cap A=\emptyset, |B|\le\lceil\ell/a\rceil\Big\}. \qquad (61)$$

Proof. We first prove that for all decreasing functions $h(t) \ge 0$,

$$\sum_{m=0}^{\infty}\Big(\int_{\ell+m\ell/a}^{\ell+(m+1)\ell/a} h^q(t)dt\Big)^{1/q} \le \max\big\{1,(a_q/a)^{1-1/q}\big\}\frac{a^{1-1/q}}{\ell^{1-1/q}}\int_0^\infty h(t)dt. \qquad (62)$$

With $x = at/\ell - m$ and possibly different $h \downarrow 0$, the above inequality is a consequence of

$$\max\Big\{\int_a^{1+a} h^q(x)dx : \int_0^1 h(x)dx = 1\Big\} \le \max\big\{1, (a_q/a)^{q-1}\big\} \tag{63}$$

It suffices to consider $0 < a \le a_q$. Since $\int_a^{1+a} h^q(x)dx$ is convex in h, it suffices to consider $h(x) = v + (u-v)I\{x \le w\}$ for certain $u \ge 1 \ge v$ and $a \le w \le 1$. Since $\int_0^1 h(x)dx = 1$, $v = (1-uw)/(1-w)$. Thus, for fixed w, $\int_a^{1+a} h^q(x)dx$ is convex in u, and its maximum is attained at the extreme points $u \in \{1, 1/w\}$. For $u = 1/w$, we have $v = 0$ and $\int_a^{1+a} h^q(x)dx = u^{q-1} - au^q$, so that the optimal u satisfies $au = (q-1)/q$, resulting in the maximum $\{(q-1)/(qa)\}^{q-1}/q = (a_q/a)^{q-1}$. This yields (63) since $\int_a^{1+a} h^q(x)dx = 1$ at the other extreme $u = 1$. Thus, $\|h_{A^c}\|_q \le (a_q/\ell)^{1-1/q}\|h_{J^c}\|_1$ by (62), and (61) follows with an application of the Hölder inequality to $\int_{\ell+m\ell/a}^{\ell+(m+1)\ell/a} w(t)h(t)dt$. ■

Proof of Proposition 5. Let $k = |J|$. Since $\|u_{A^c}\|_q \le (a_q/\ell)^{1-1/q}\|u_{J^c}\|_1$ by Lemma 20 and $\|u_{J^c}\|_1 \le \xi\|u_J\|_1 \le \xi k^{1-1/q}\|u_A\|_q$, $\|u\|_q \le C_{q,q}(\xi, k/\ell)\|u_A\|_q$. By the Hölder inequality $\|u\|_s \le \|u\|_1^{(1/s-1/q)/(1-1/q)}\|u\|_q^{(1-1/s)/(1-1/q)} = \|u\|_1^{1/s_1}\|u\|_q^{1-1/s_1}$ with $s_1 = (1-1/q)/(1/s-1/q) \ge s$. Since $\|u\|_1 \le (1+\xi)\|u_J\|_1 \le (1+\xi)k^{1-1/q}\|u_A\|_q$,

$$\|u\|_s k^{1/q-1/s}/\|u_A\|_q \le (1+\xi)^{1/s_1} C_{q,q}^{1-1/s_1}(\xi, k/\ell) = C_{s,q}(\xi, k/\ell).$$

This gives (17). Since $v'_{J^c}\Sigma_{J^c,*}u \le 0$ for $u \in \mathscr{C}_-(\xi, J)$ and $v \in \mathscr{Q}(u, J)$, the first parts of (18) and (19) follow from (8) and (9). For $h = u$ and the choice of A in Lemma 20, an application of (61) with $w = (v'_A\Sigma_{A,*})'$ yields

$$\begin{aligned} v'_A\Sigma_{A,*}u &= v'_A\Sigma_A u_A + v'_A\Sigma_{A,A^c}u_{A^c} \\ &\ge v'_A\Sigma_A u_A - \max_B \|\Sigma_{B,A}v_A\|_{r/(r-1)}(a_r/\ell)^{1-1/r}\|u_{J^c}\|_1. \end{aligned} \tag{64}$$

Since $\|u_{J^c}\|_1 \le \xi\|u_J\|_1 \le \xi k^{1-1/q}\|u_A\|_q$, (64) and (16) imply

$$\begin{aligned} \frac{v'_A\Sigma_{A,*}u}{\|v_A\|_1\|u_A\|_q/k^{1/q}} &\ge \frac{v'_A\Sigma_A u_A - \xi\|u_A\|_q k^{1-1/q}(a_r/\ell)^{1-1/r}\max_B\|\Sigma_{B,A}v_A\|_{r/(r-1)}}{\|v_A\|_1\|u_A\|_q/k^{1/q}} \\ &= \phi_{q,\ell}(u, A, r, w, B; \xi, J)k^{1/q}/\|w_A\|_1 \end{aligned}$$

with $w = v\|u_A\|_q/(v'_A\Sigma_A u_A)$. This gives the second parts of (18) and (19). ■

Proof of Theorem 8. This theorem is a direct consequence of (10) with $q = \infty$, since $\|\widehat{\beta} - \beta\|_\infty \le \lambda' < \min_{\beta_j \ne 0}|\beta_j|/2$ guarantees $\{j : |\widehat{\beta}_j| > \lambda'\} = \{j : \beta_j \ne 0\}$. ■

Proof of Theorem 10. Let $h = \widehat{\beta} - \beta^*$ for either estimator. As in (14) and (15),

$$\|\Sigma h\|_\infty \le \xi_1\lambda,\ \|h_{J^c}\|_1 \le \xi_0\|h_J\|_1 + (\xi_0+1)\rho_J, \tag{65}$$

in the given events, with $\{\xi_0, \xi_1\} = (1, 2)$ for the Dantzig selector and the $\{\xi_0, \xi_1\}$ in (37) for the Lasso. It follows from Theorem 3 that (36) and (37) hold for $\|h_{J^c}\|_1 \le \xi\|h_J\|_1$, or equivalently $h \in \mathscr{C}(\xi, J)$. By (65), it remains to consider

$$\xi\|h_J\|_1 \le \|h_{J^c}\|_1 \le \xi_0\|h_J\|_1 + (\xi_0+1)\rho_J.$$

Since $\xi > \xi_0$, $\|h_J\|_1 \le (\xi_0+1)\rho_J/(\xi-\xi_0)$ in this case, so that $\|h\|_1 \le (1+\xi_0)\|h_J\|_1 + (\xi_0+1)\rho_J \le \rho_J(1+\xi_0)(1+\xi)/(\xi-\xi_0) = \xi_2\rho_J$. Thus, (35) gives $\|h\|_q \le M_q(\xi_1\lambda, \xi_2\rho_J)$. ■

Proof of Theorem 12. Let $h = \widehat{\beta}^{(L)} - \beta^*$, $z = X'(y - X\beta^*)/n$, $k = |J|$, $\ell = d - k$, and $A = \arg\max_{|B|=d, B\supset J} \|h_B\|$ as in (40). Since $k \le 3d/4$, $4\ell \ge d$. Lemma 20 with $\{w, q, a_q\} = \{z, 2, 1/4\}$ yields $h'_{A^c} z_{A^c} \le \|h_{J^c}\|_1 z^*_{2,4\ell} \le \|h_{J^c}\|_1 z^*_{2,d}$, so that

$$\begin{aligned} \|Xh\|^2/n &= h'_A z_A + h'_{A^c} z_{A^c} - h'X'(y - X\widehat{\beta}^{(L)})/n \\ &\le \sqrt{d}\|h_A\| z^*_{2,d} + \|h_{J^c}\|_1 z^*_{2,d} - \lambda\|\widehat{\beta}^{(L)}\|_1 + \lambda\|\beta^*\|_1. \end{aligned}$$

Since $-\lambda\|\widehat{\beta}^{(L)}\|_1 + \lambda\|\beta^*\|_1 \le -\lambda\|h_{J^c}\|_1 + \lambda\|h_J\|_1 + 2\lambda\rho_J$ and $\|h_J\|_1 \le \sqrt{3d/4}\|h_A\|$,

$$\|Xh\|^2/n + (\lambda - z^*_{2,d})\|h_{J^c}\|_1 \le (\lambda + z^*_{2,d})\sqrt{d}\|h_A\| + 2\lambda\rho_J.$$

Since $\Sigma_{A,*}h = z_A - X'_A(y - X\widehat{\beta}^{(L)})/n$, $\|\Sigma_{A,*}h\| \le (z^*_{2,d} + \lambda)\sqrt{d}$. Thus, as in (65),

$$\|\Sigma_{A,*}h\| \le \xi_1\sqrt{d}\lambda,\ \|h_{J^c}\|_1 \le \xi_0\sqrt{d}\|h_A\| + (\xi_0 + 1)\rho_J. \tag{66}$$

For $\|h_{J^c}\|_1 \le \xi\sqrt{d}\|h_A\|$, (42) follows from

$$d^{1/2-1/q}\|h\|_q CIF_{q,relax}(\xi, J, d) \le \|\Sigma_{A,*}h\| \le \xi_1\sqrt{d}\lambda.$$

For $\xi\sqrt{d}\|h_A\| \le \|h_{J^c}\|_1$, the second inequality of (66) gives $\sqrt{d}\|h_A\| \le (\xi_0 + 1)\rho_J/(\xi - \xi_0)$ and then $\|h\|_1 \le (1 + \xi_0)(\sqrt{d}\|h_A\| + \rho_J) \le \xi_2\rho_J$. Thus, $\|h\|_q \le M_{q,relax}(\xi_1\lambda, \xi_2\rho_J, J, d)$ by (41), in view of the first inequality of (66). ■

Proof of Proposition 15. Let A be the index set of the k largest u_j and w satisfy $\|\Sigma_A w_A\|_{q/(q-1)} = w'_A\Sigma_A u_A/\|u_A\|_q = 1$. By Lemma 20, $w'_A\Sigma_{A,A^c}u_{A^c} \le \kappa_{q,k,k}k^{1/q-1}\rho$, so that $\|u_A\|_q = w'_A\Sigma_A u_A \le \|w_A\|_1\|\Sigma u\|_\infty + \kappa_{q,k,k}k^{1/q-1}\rho \le \eta_{q,k}k^{1/q}\lambda + \kappa_{q,k,k}k^{1/q-1}\rho$. This and $\|u\|_q \le (\|u_A\|_q^q + (a_q/k)^{q-1}\rho^q)^{1/q}$ from Lemma 20 yields (44).

Let A be as in (41) and $w_A = \Sigma_A^{-1}u_A/\|u_A\|$. For $\|\Sigma_{A,*}u\| \le \sqrt{d}\lambda$ and $\|u\|_1 \le \rho$, $\|u_A\| = w'_A\Sigma_A u_A \le \|w_A\|d^{1/2}\lambda + w_A\Sigma_{A,A^c}u_{A^c} \le \eta^*_{2,d}d^{1/2}\lambda + \kappa_{2,d,\ell}\ell^{-1/2}\rho$, so that (45) follows from $\|u\|_q^q \le (d^{1/q-1/2}\|u_A\|)^q + (a_q/\ell)^{q-1}\rho^q$.

Let u and A be as in (40) and $k = |J|$. Similar to the proof of Proposition 5, we have $\|u\|^2 \le \{1 + \xi^2 k/(4\ell)\}\|u_A\|^2$ and $\|u\|_1 \le \sqrt{k}\|u_A\| + \|u_{J^c}\|_1 \le (1 + \xi\sqrt{d/k})k^{1/2}\|u_A\|$. Thus, for $1 \le q \le 2$, $\|u\|_q \le \|u\|_1^{2/q-1}\|u\|^{2-2/q} \le C_{q,2}(\xi(d/k)^{1/2}, k^2/(d\ell))k^{1/q-1/2}\|u_A\|$. Since $\|u_{J^c}\|_1 \le \xi d^{1/2}\|u_A\|$ and (64) holds for $v = u$, $u'_A\Sigma_{A,*}u \ge \phi_{2,\ell}(u, A, r, w, B; \xi', J)/\|w_A\|$ for $\|u_A\| = 1$, where $\xi' = \xi\sqrt{d/k}$ and $w = u/(u'_A\Sigma_A u_A)$. Thus, (20) gives (46). ■

The proof of Theorem 17 requires the following lemma.

Lemma 21 *Let $\widehat{\beta}$ be either the Dantzig or the Lasso estimator at penalty level λ. Suppose $\|\beta\|_r \le R$ with $0 < r \vee 1 \le q$. For any event Ω_0 with $t_* = \sqrt{2\log(1/P_{\beta,X}(\Omega_0))} \ge 1$,*

$$E_{\beta,X}\|\widehat{\beta} - \beta\|_q^q I_{\Omega_0} \le 2^{q-1}P_{\beta,X}(\Omega_0)\Big\{\frac{\Gamma(2q+1)}{(t_*^2 n\lambda/\sigma^2)^q} + \Big(\frac{(t_* + \sqrt{n})^2}{n\lambda/\sigma^2} + 2p^{(1-1/r)_+}R\Big)^q\Big\}. \tag{67}$$

In particular, if $(\log p)/n + \sigma^2/(n\lambda^2) + R^r/(n\lambda^r) + \lambda^r/R^r = O(1)$, then

$$E_{\beta,X}\|\widehat{\beta} - \beta\|_q^q I_{\Omega_0} = o(1)R^r\lambda^{q-r}, \tag{68}$$

provided that $P_{\beta,X}(\Omega_0)(\lambda^r/R^r)\{(\sigma/\lambda)^{2q} + p^{q(1-1/r)_+}(R/\lambda)^q\} = o(1)$.

Remark 22 *Since the unit sphere $S^{n-1} \subset \mathbb{R}^n$ is covered by $(2/\varepsilon+1)^n$ ε-balls for all $\varepsilon > 0$, a certain ε ball contains at least m unit vectors $x_j/\|x_j\|$ for $\varepsilon = (\log(p/m))/(2n)$. It follows that the set of design vectors x_j contains some highly correlated clusters when $(\log p)/n \geq 2$. Thus, the condition $(\log p)/n = O(1)$ is natural for the estimation of β.*

Proof. Let $\widehat{\beta}$ be the Lasso estimator. Since $\widehat{\beta}$ minimizes the penalized loss, $\lambda\|\widehat{\beta}\|_1 \leq \|\varepsilon\|^2/(2n) + \lambda\|\beta\|_1$, so that

$$\|\widehat{\beta}\|_1 + \|\beta\|_1 \leq \frac{\|\varepsilon\|^2}{2n\lambda} + 2\|\beta\|_1 \leq \frac{(\|\varepsilon\|/\sigma - t_* - \sqrt{n})_+^2}{n\lambda/\sigma^2} + \frac{(t_* + \sqrt{n})^2}{n\lambda/\sigma^2} + 2p^{(1-1/r)_+}R.$$

Since $\|\varepsilon/\sigma\|$ is a Lipschitz(1) function of $\varepsilon/\sigma \sim N(0, I_n)$ and $E_{\beta,X}\|\varepsilon\|/\sigma \leq \sqrt{n}$, the Gaussian isoperimetric theorem gives $P_{\beta,X}\{\|\varepsilon\|/\sigma - \sqrt{n} > t\} \leq e^{-t^2/2}$, so that

$$\begin{aligned} E_{\beta,X}(\|\varepsilon\|/\sigma - t_* - \sqrt{n})_+^{2q} &\leq \int_0^\infty P_{\beta,X}\{\|\varepsilon\|/\sigma - t_* - \sqrt{n} > t\} dt^{2q} \\ &\leq \int_0^\infty e^{-t_*^2/2 - t_* t} dt^{2q} = P_{\beta,X}(\Omega_0)\Gamma(2q+1)/t_*^{2q}. \end{aligned}$$

The above inequalities yield (67) due to $\|\widehat{\beta} - \beta\|_q^q \leq (\|\widehat{\beta}\|_1 + \|\beta\|_1)^q$ for $q \geq 1$.

It follows from (67) that

$$\begin{aligned} & E_{\beta,X}\|\widehat{\beta} - \beta\|_q^q I_{\Omega_0}/\{R^r\lambda^{q-r}\} \\ = \; & O(\lambda^r/R^r) P_{\beta,X}(\Omega_0)\Big\{O(1) + (t_*^2/n+1)^q(\sigma/\lambda)^{2q} + p^{q(1-1/r)_+}R^q/\lambda^q\Big\}. \end{aligned}$$

Since the right-hand side is of no greater order than $P_{\beta,X}(\Omega_0)\{(t_*^2+n)^q + p^q n^{q/r}\} = o(1)$ for $t_*^2/(n \vee \log p) \to \infty$, it suffices to consider the case $t_*^2/n = O(1)$. Hence, (68) holds under the specified conditions. The same conclusions hold for the Dantzig selector, since $\|\widehat{\beta}^{(D)}\|_1 \leq \|\widehat{\beta}^{(L)}\|_1$. ■

Proof of Theorem 17. We first bound $\lambda_{univ}/\lambda_{mm}$ and the expected loss for large $z_\infty^* = \|X'\varepsilon/n\|_\infty$. Let $\sigma_n = \sigma/\sqrt{n}$. Since $R^r/\lambda_{mm}^r \asymp d$, (5) and (55) give

$$2\sigma_n^2 \log(p/d) \approx \lambda_{mm}^2. \tag{69}$$

Since $(\log d)/\log p \leq c_0$, $(1-c_0)\lambda_{univ}^2 = (1-c_0)2\sigma_n^2 \log p \leq 2\sigma_n^2 \log(p/d) \approx \lambda_{mm}^2$ and

$$C_1 \approx C_1^* = 2(\lambda/\lambda_{mm})(d\lambda_{mm}^r/R^r)^{1/q},\ C_2 \approx C_2^* = R/(\lambda d^{1/r}).$$

Let $\Omega_0 = \{z_\infty^*/\lambda > \alpha_1\}$. Since z_∞^* is the maximum of p variables from $N(0, \sigma_n^2)$, $P_{\beta,X}\{\Omega_0\} \ll p\exp(-n(\alpha_1\lambda)^2/(2\sigma^2)) \leq p^{1-(\alpha_1/\alpha_0)^2}$ for large n. Thus, due to $\lambda^2/\sigma_n^2 \asymp \log p$ and $n \geq d \asymp R^r/\lambda_{mm}^r \asymp R^r/\lambda^r \to \infty$, we have

$$P_{\beta,X}(\Omega_0)(\lambda^r/R^r)\{(\sigma/\lambda)^{2q} + (R/\lambda)^q\} = o(1)p^{1-(\alpha_1/\alpha_0)^2}(n^q/d + d^{q/r-1}) = o(1).$$

Since $0 < r \leq 1$, Lemma 21 gives $E_{\beta,X}\|h\|_q^q I\{z_\infty^*/\lambda > \alpha_1\} = o(R^r\lambda_{mm}^{q-r})$, where $h = \widehat{\beta} - \beta$.

Next we prove $E_{\beta,X}\|h\|_q^q I\{z_\infty^*/\lambda \leq \alpha_1\} = O(R^r\lambda_{mm}^{q-r})$. Consider $z_\infty^* \leq \alpha_1\lambda$. By (65) with $J = \emptyset$, $\|h\|_q \leq 2M_q(\lambda, \xi'\|\beta\|_1)$ with $\xi' = \xi^*$ for $r = 1$. Since $M_q(\lambda, \rho) = cM_q(\lambda/c, \rho/c)$,

$$\|h\|_q/(R^{r/q}\lambda_{mm}^{1-r/q}) \leq C_1^* M_q(d^{-1/q}, \xi_2' C_2^* d^{1/r-1/q}).$$

This gives (47) for $r = 1$ and $E_{\beta,X}\|h\|_q^q I\{\alpha\lambda < z^*_\infty \le \alpha_1\lambda\} = o(p^{1-(\alpha/\alpha_0)^2} d^{q(1/r-1)}) = o(1)$ for $r < 1$. Let $J = \arg\max_{|A|=d} \|\beta_A\|_1$. For $\beta \in \Theta_{r,R}$, $\|\beta_{J^c}\|_\infty \le (R^r/d)^{1/r}$, so that $\rho_J/(\lambda d) \le (R^r/d)^{(1-r)/r} R^r/(\lambda d) = C_2^*$. Thus, in the event $z^*_\infty \le \alpha\lambda$, Theorem 10 gives

$$\|h\|_q/(R^{r/q}\lambda_{mm}^{1-r/q}) \le \max\left\{C_1^*/CIF_q^*(\xi,d), C_1^* M_q(d^{-1/q}, \xi^* C_2^* d^{1-1/q})\right\},\ r < 1.$$

It remains to prove (48) under $\eta_{q,d} + \kappa_{q,d,d} = O(1)$. In fact, by (44) it suffices to prove $1/CIF_q^*(\xi,k) = O(1)$ for any $k + \ell = d$ with $k \asymp d$. This follows from (21), since $CIF_{q,p}(\xi,|J|) \gtrsim \{1 - \xi(k/\ell)^{1-1/q}\kappa_{q,d,d}\}/\eta_{q,d} > 0$ uniformly for $|J| = k$ and small k/ℓ. ■

The proof of Theorem 19 requires the following simpler version of Lemma 2 in Zhang (2010).

Lemma 23 *Let $\widetilde{p}_\ell$ be the positive number satisfying $2\log\widetilde{p}_\ell - 1 - \log(2\log\widetilde{p}_\ell) = (2/\ell)\log\binom{p}{\ell}$. Suppose $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ under probability P. Then,*

$$P\Big\{\max_{|A|=\ell}\|P_A\varepsilon\| \ge \sigma\sqrt{2\ell\log\widetilde{p}_\ell}\Big\} \le \frac{1}{2\sqrt{\log\widetilde{p}_\ell}} \le \frac{1}{\sqrt{2}},$$

where $P_A = X_A(X_A'X_A)^{-1}X_A'$ is the projection to the linear span of $\{x_j, j \in A\}$.

Proof of Theorem 19. Let $\gamma_d = (1+\delta_d^+)^{1/2}$ There are two cases. We omit the proof in the case of $\lambda = \lambda_{univ}/\alpha$ since it is identical to the second half of the proof of Theorem 17. It remains to consider the case $\lambda < \lambda_{univ}/\alpha$, that is, $(1+\varepsilon_0)\gamma_d\lambda_{mm} < \lambda_{univ}$.

For $|A| = d$, $\|X_A'\varepsilon/n\| \le \gamma_d\|P_A\varepsilon\|/\sqrt{n}$, so that by (38) and Lemma 23

$$P_{\beta,X}\Big\{z^*_{2,d} \le \gamma_d\sigma\sqrt{(2/n)\log\widetilde{p}_d}\Big\} \ge 1 - 1/(2\sqrt{\log\widetilde{p}_d}) \to 1.$$

Since $R^r/\lambda_{mm}^r \asymp d$ and $\lambda_{mm}n^{1/2}/\sigma \to \infty$, $\lambda_{mm} = (1+o(1))\sigma\sqrt{(2/n)\log(p/d)}$ by (69). By Stirling, $\log\binom{p}{d} = (1+o(1))d\log(p/d)$ for $p/d \to \infty$, so that $\lambda_{mm} = (1+o(1))\sigma\sqrt{(2/n)\log\widetilde{p}_d}$. Thus, $\gamma_d\sigma\sqrt{(2/n)\log\widetilde{p}_d} \le \alpha\lambda$ and $P_{\beta,X}\{z^*_{2,d} \le \alpha\lambda\} \to 1$.

Consider the event $z^*_{2,d} \le \alpha\lambda$. Since $\xi > (1+\alpha)/(1-\alpha)$, Theorem 12 asserts that for $J = \arg\max_{|A|=k}\|\beta_A\|_1$ and certain constants $\{\xi_1, \xi_2\}$,

$$\|h\|_q \le \max\Big\{\frac{\xi_1 d^{1/q}\lambda}{CIF_{q,relax}(\xi,J,d)}, M_{q,relax}\Big(\xi_1\lambda, \xi_2\rho_J, J, d\Big)\Big\}.$$

The rest of the proof is similar to the proof of Theorem 17 and omitted. ■

References

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, 58:1300–1308, 2010.

E. Candes and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37: 2145–2177, 2009.

E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404, 2007.

S. Chen and D. L. Donoho. On basis pursuit. Technical report, Department of Statistics, Stanford University, 1994.

D. L. Donoho and I. Johnstone. Minimax risk over ℓ_p–balls for ℓ_q–error. *Probability Theory and Related Fields*, 99:277–303, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32:407–499, 2004.

B. Efron, T. Hastie, and R. Tibshirani. Discussion: The dantzig selector: statistical estimation when *p* is much larger than *n*. *Annals of Statistics*, 35:2358–2364, 2007.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156. Morgan Kauffmann, San Francisco, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28:337–407, 2000.

E. Greenshtein and Y. Rotiv. Persistence in high–dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high–dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009.

K. Lounici. Sup–norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso–type recovery of sparse representations for high–dimensional data. *Annals of Statistics*, 37:246–270, 2009.

M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000a.

M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000b.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high–dimensional linear regression over ℓ_q–balls. Technical report, University of California, Berkeley, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996.

J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006.

S. van de Geer. The deterministic lasso. Technical Report 140, ETH Zurich, Switzerland, 2007.

S. van de Geer. High–dimensional generalized linear models and the lasso. *Annals of Statistics*, 36: 614–645, 2008.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

M. J. Wainwright. Sharp thresholds for noisy and high–dimensional recovery of sparsity using ℓ_1–constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55: 2183–2202, 2009.

F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig estimators. Technical report, Department of Statistics and Biostatistics, Rutgers University, 2009.

C.-H. Zhang. Least squares estimation and variable selection under minimax concave penalty. In *Mathematisches Forschungsintitut Oberwolfach: Sparse Recovery Problems in High Dimensions*, 3 2009a.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942, 2010.

C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high–dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009b.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.