

DFG-SNF Research Group FOR916

Statistical Regularization and Qualitative Constraints

Shuheng Zhou

Thresholded Lasso for high dimensional variable selection and statistical estimation

Preprint FOR916 10-09

Preprint-Series of the Research Group FOR916

Thresholded Lasso for high dimensional variable selection and statistical estimation *

Shuheng Zhou

Seminar für Statistik, Department of Mathematics, ETH Zürich, CH-8092, Switzerland

February 8, 2010

Abstract

Given n noisy samples with p dimensions, where $n \ll p$, we show that the multi-step thresholding procedure based on the Lasso – we call it the *Thresholded Lasso*, can accurately estimate a sparse vector $\beta \in \mathbb{R}^p$ in a linear model $Y = X\beta + \epsilon$, where $X_{n \times p}$ is a design matrix normalized to have column ℓ_2 norm \sqrt{n} , and $\epsilon \sim N(0, \sigma^2 I_n)$. We show that under the restricted eigenvalue (RE) condition (Bickel-Ritov-Tsybakov 09), it is possible to achieve the ℓ_2 loss within a logarithmic factor of the ideal mean square error one would achieve with an *oracle* while selecting a sufficiently sparse model – hence achieving *sparse oracle inequalities*; the oracle would supply perfect information about which coordinates are non-zero and which are above the noise level. In some sense, the Thresholded Lasso recovers the choices that would have been made by the ℓ_0 penalized least squares estimators, in that it selects a sufficiently sparse model without sacrificing the accuracy in estimating β and in predicting $X\beta$. We also show for the Gauss-Dantzig selector (Candès-Tao 07), if X obeys a uniform uncertainty principle and if the true parameter is sufficiently sparse, one will achieve the sparse oracle inequalities as above, while allowing at most s_0 irrelevant variables in the model in the worst case, where $s_0 \leq s$ is the smallest integer such that for $\lambda = \sqrt{2 \log p/n}$, $\sum_{i=1}^p \min(\beta_i^2, \lambda^2 \sigma^2) \leq s_0 \lambda^2 \sigma^2$. Our simulation results on the Thresholded Lasso match our theoretical analysis excellently.

Keyword. Linear regression, Lasso, Gauss-Dantzig Selector, ℓ_1 regularization, ℓ_0 penalty, multiple-step procedure, ideal model selection, oracle inequalities, restricted orthonormality, statistical estimation, thresholding, linear sparsity, random matrices

1 Introduction

In a typical high dimensional setting, the number of variables p is much larger than the number of observations n . This challenging setting appears in linear regression, signal recovery, covariance selection

*A preliminary version of this paper with title: Thresholding Procedures for High Dimensional Variable Selection and Statistical Estimation, has appeared in Proceedings of Advances in Neural Information Processing Systems 22, (NIPS 2009). This research was supported by the Swiss National Science Foundation (SNF) Grant 20PA21-120050/1.

in graphical modeling, and sparse approximations. In this paper, we consider recovering $\beta \in \mathbb{R}^p$ in the following linear model:

$$Y = X\beta + \epsilon, \quad (1.1)$$

where X is an $n \times p$ design matrix, Y is a vector of noisy observations and ϵ is the noise term. We assume throughout this paper that $p \geq n$ (i.e. high-dimensional), $\epsilon \sim N(0, \sigma^2 I_n)$, and the columns of X are normalized to have ℓ_2 norm \sqrt{n} . Given such a linear model, two key tasks are to identify the relevant set of variables and to estimate β with bounded ℓ_2 loss. In particular, recovery of the sparsity pattern $S = \text{supp}(\beta) := \{j : \beta_j \neq 0\}$, also known as variable (model) selection, refers to the task of correctly identifying the support set (or a subset of “significant” coefficients in β) based on the noisy observations.

Even in the noiseless case, recovering β (or its support) from (X, Y) seems impossible when $n \ll p$. However, a line of recent research shows that when β is sparse: when it has a relatively small number of nonzero coefficients and when the design matrix X is also sufficiently nice, it becomes possible [Candès et al. \(2006\)](#); [Candès and Tao \(2005, 2006\)](#); [Donoho \(2006a\)](#). One important stream of research, which we also adopt here, requires computational feasibility for the estimation methods, among which the Lasso and the Dantzig selector are both well studied and shown with provable nice statistical properties; see for example [Bickel et al. \(2009\)](#); [Candès and Tao \(2007\)](#); [Greenshtein and Ritov \(2004\)](#); [Meinshausen and Bühlmann \(2006\)](#); [Meinshausen and Yu \(2009\)](#); [Ravikumar et al. \(2008\)](#); [van de Geer \(2008\)](#); [Wainwright \(2009b\)](#); [Zhao and Yu \(2006\)](#). For a chosen penalization parameter $\lambda_n \geq 0$, regularized estimation with the ℓ_1 -norm penalty, also known as the Lasso ([Tibshirani, 1996](#)) or Basis Pursuit ([Chen et al., 1998](#)) refers to the following convex optimization problem

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \quad (1.2)$$

where the scaling factor $1/(2n)$ is chosen by convenience; The Dantzig selector ([Candès and Tao, 2007](#)) is defined as,

$$(DS) \quad \arg \min_{\hat{\beta} \in \mathbb{R}^p} \|\hat{\beta}\|_1 \quad \text{subject to} \quad \left\| \frac{1}{n} X^T (Y - X\hat{\beta}) \right\|_{\infty} \leq \lambda_n. \quad (1.3)$$

Our goal in this work is to recover S as accurately as possible: we wish to obtain $\hat{\beta}$ such that $|\text{supp}(\hat{\beta}) \setminus S|$ (and sometimes $|S \triangle \text{supp}(\hat{\beta})|$ also) is small, with high probability, while at the same time $\|\hat{\beta} - \beta\|_2^2$ is bounded within logarithmic factor of the ideal mean square error one would achieve with an oracle which would supply perfect information about which coordinates are non-zero and which are above the noise level (hence achieving the *oracle inequality* as studied in [Candès and Tao \(2007\)](#); [Donoho and Johnstone \(1994\)](#)); We deem the bound on ℓ_2 -loss as a natural criteria for evaluating a sparse model when it is not exactly S . Let $s = |S|$.

Given $T \subseteq \{1, \dots, p\}$, let us define X_T as the $n \times |T|$ submatrix obtained by extracting columns of X indexed by T ; similarly, let $\beta_T \in \mathbb{R}^{|T|}$, be a subvector of $\beta \in \mathbb{R}^p$ confined to T . Formally, we propose and study a **Multi-step Procedure**: First we obtain an initial estimator β_{init} using the Lasso as in (1.2) or the Dantzig selector as in (1.3), with $\lambda_n = d\sigma\sqrt{2\log p/n}$, for some constant $d > 0$.

1. We then threshold the estimator β_{init} with t_0 , with the general goal such that, we get a set I_1 with cardinality at most $2s$; in general, we also have $|I_1 \cup S| \leq 2s$, where $I_1 = \{j \in \{1, \dots, p\} : \beta_{j,\text{init}} \geq t_0\}$ for some t_0 to be specified. Set $I = I_1$.
2. We then feed (Y, X_I) to either the Lasso estimator as in (1.2) or the ordinary least squares (OLS) estimator to obtain $\hat{\beta}$, where we set $\hat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$ and $\hat{\beta}_{I^c} = 0$.
3. Possibly threshold $\hat{\beta}_{I_1}$ with $t_1 = 4\lambda_n \sqrt{|I_1|}$ to obtain I_2 , and repeat step 2 with $I = I_2$ to obtain $\hat{\beta}_I$; set other coordinates to zero and return $\hat{\beta}$.

Our algorithm is constructive in that it relies neither on the unknown parameters s and $\beta_{\min} := \min_{j \in S} |\beta_j|$, nor the exact knowledge of those that characterize the incoherence conditions on X ; instead, our choice of λ_n and thresholding parameters only depends on σ, n , and p , and some crude estimation of certain parameters, which we will explain in later sections. In our experiments, we apply only the first two steps with the Lasso as an initial estimator, which we refer to as the *Thresholded Lasso* estimator; the Gauss-Dantzig selector is a two-step procedure with the Dantzig selector as β_{init} Candès and Tao (2007). We apply the third step only when β_{\min} is sufficiently large, so as to get a very sparse model $I \supset S$ (cf. Theorem 1.1). We now formally define some incoherence conditions in Section 1.1 and elaborate on our goals in Section 1.2, where we also outline the rest of this section.

1.1 Incoherence conditions

For a matrix A , let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ denote the smallest and the largest eigenvalues respectively. We refer to a vector $v \in \mathbb{R}^p$ with at most s non-zero entries, where $s \leq p$, as a **s -sparse** vector. Occasionally, we use $\beta_T \in \mathbb{R}^{|T|}$, where $T \subseteq \{1, \dots, p\}$, to also represent its 0-extended version $\beta' \in \mathbb{R}^p$ such that $\beta'_{T^c} = 0$ and $\beta'_T = \beta_T$; for example in (1.10) below. We assume

$$\Lambda_{\min}(2s) \triangleq \min_{v \neq 0; 2s\text{-sparse}} \frac{\|Xv\|_2^2}{n \|v\|_2^2} > 0, \quad (1.4)$$

where $n \geq 2s$ is necessary, as any submatrix with more than n columns must be singular. In general, we also assume that

$$\Lambda_{\max}(2s) \triangleq \max_{v \neq 0; 2s\text{-sparse}} \frac{\|Xv\|_2^2}{n \|v\|_2^2} < \infty. \quad (1.5)$$

Candès and Tao (2005) define the s -restricted isometry constant δ_s of X to be the smallest quantity such that for all $T \subseteq \{1, \dots, p\}$ with $|T| \leq s$ and coefficients sequences $(v_j)_{j \in T}$, it holds that

$$(1 - \delta_s) \|v\|_2^2 \leq \|X_T v\|_2^2 / n \leq (1 + \delta_s) \|v\|_2^2; \quad (1.6)$$

The (s, s') -restricted orthogonality constant $\theta_{s,s'}$ is the smallest quantity such that for all disjoint sets $T, T' \subseteq \{1, \dots, p\}$ of cardinality $|T| \leq s$ and $|T'| \leq s'$,

$$\frac{|\langle X_T c, X_{T'} c' \rangle|}{n} \leq \theta_{s,s'} \|c\|_2 \|c'\|_2 \quad (1.7)$$

holds, where $s + s' \leq p$. Note that $\theta_{s,s'}$ and δ_s are non-decreasing in s, s' and small values of $\theta_{s,s'}$ indicate that disjoint subsets covariates in X_T and $X_{T'}$ span nearly orthogonal subspaces (See Lemma 5.4 for a general bound on $\theta_{s,s'}$.) For δ_s , it holds that $1 - \delta_s \leq \Lambda_{\min}(s) \leq \Lambda_{\max}(s) \leq 1 + \delta_s$. Hence $\delta_{2s} < 1$ implies that condition (1.4) holds. As a consequence of these definitions, for any subset I , we have

$$\Lambda_{\max}(|I|) \geq \Lambda_{\max}(X_I^T X_I/n) \geq \Lambda_{\min}(X_I^T X_I/n) \geq \Lambda_{\min}(|I|) \quad (1.8)$$

where $\Lambda_{\min}(|I|) \geq \Lambda_{\min}(2s) > 0$ and $\Lambda_{\max}(|I|) \leq \Lambda_{\max}(2s)$ for $|I| \leq 2s$. We next introduce some conditions on the design, namely, the Restricted Eigenvalue (RE) condition by Bickel et al. (2009) and the Uniform Uncertainty Principle by Candès and Tao (2007) which we use throughout this paper.

Assumption 1.1. (Restricted Eigenvalue Condition $RE(s, k_0, X)$ (Bickel et al., 2009)) For some integer $1 \leq s \leq p$ and a number $k_0 > 0$, it holds for all $v \neq 0$,

$$\frac{1}{K(s, k_0)} \triangleq \min_{\substack{J_0 \subseteq \{1, \dots, p\}, \\ |J_0| \leq s}} \min_{\|v_{J_0^c}\|_1 \leq k_0 \|v_{J_0}\|_1} \frac{\|Xv\|_2}{\sqrt{n} \|v_{J_0}\|_2} > 0. \quad (1.9)$$

Assumption 1.2. (A Uniform Uncertainty Principle) (Candès and Tao, 2007) For some integer $1 \leq s < n/3$, assume $\delta_{2s} + \theta_{s,2s} < 1$, which implies that $\lambda_{\min}(2s) > \theta_{s,2s}$ given that $1 - \delta_{2s} \leq \Lambda_{\min}(2s)$.

If $RE(s, k_0, X)$ is satisfied with $k_0 \geq 1$, then (1.4) must hold; Bounds on prediction loss and ℓ_p loss, where $1 \leq p \leq 2$, for estimating the parameters are derived for both the Lasso and the Dantzig selector in both linear and nonparametric regression models; see Bickel et al. (2009). We now define oracle inequalities in terms of ℓ_2 loss as explored in Candès and Tao (2007), where they show such inequalities hold for the Dantzig selector under the UUP (cf. Proposition 4.1).

1.2 Oracle inequalities

Consider the least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$, where $|I| \leq s$. Consider the *ideal* least-squares estimator β^\diamond

$$\beta^\diamond = \arg \min_{I \subseteq \{1, \dots, p\}, |I| \leq s} \mathbf{E} \left\| \beta - \widehat{\beta}_I \right\|_2^2 \quad (1.10)$$

which minimizes the expected mean squared error. It follows from Candès and Tao (2007) that for $\Lambda_{\max}(s) < \infty$

$$\mathbf{E} \|\beta - \beta^\diamond\|_2^2 \geq \min(1, 1/\Lambda_{\max}(s)) \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n). \quad (1.11)$$

Now we check if for $\Lambda_{\max}(s) < \infty$, it holds with high probability that

$$\|\widehat{\beta} - \beta\|_2^2 = O(\log p) \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n), \quad \text{so that} \quad (1.12)$$

$$\|\widehat{\beta} - \beta\|_2^2 = O(\log p) \max(1, \Lambda_{\max}(s)) \mathbf{E} \|\beta^\diamond - \beta\|_2^2 \quad (1.13)$$

holds in view of (1.11). These bounds are meaningful since

$$\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n) = \min_{I \subseteq \{1, \dots, p\}} \|\beta - \beta_I\|_2^2 + \frac{|I|\sigma^2}{n}$$

represents the squared bias and variance. Define s_0 as the smallest integer such that

$$\sum_{i=1}^p \min(\beta_i^2, \lambda^2 \sigma^2) \leq s_0 \lambda^2 \sigma^2, \text{ where } \lambda = \sqrt{2 \log p/n}. \quad (1.14)$$

A consequence of this definition is: $|\beta_j| < \lambda \sigma$ for all $j > s_0$, if we order $|\beta_1| \geq |\beta_2| \dots \geq |\beta_p|$ (cf. (4.7)). We define a quantity $\lambda_{\sigma, a, p}$ for each $a > 0$, by which we bound the maximum correlation between the noise and covariates of X , which we only apply to X with column ℓ_2 norm bounded by \sqrt{n} ; For each $a \geq 0$, let

$$\mathcal{T}_a := \left\{ \epsilon : \|X^T \epsilon/n\|_\infty \leq \lambda_{\sigma, a, p}, \text{ where } \lambda_{\sigma, a, p} = \sigma \sqrt{1+a} \sqrt{2 \log p/n} \right\}, \quad (1.15)$$

we have (see Candès and Tao (2007)) $\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p} a)^{-1}$.

The main theme of our paper is to explore oracle inequalities of the thresholding procedures under conditions as described above. For the Lasso estimator and the Dantzig selector, under the sparsity constraint, such oracle results have been obtained in a line of recent work for either the prediction error or the ℓ_p loss, where $1 \leq p \leq 2$; see for example Bickel et al. (2009); Bunea et al. (2007a,b,c); Cai et al. (2009); Candès and Plan (2009); Candès and Tao (2007); Koltchinskii (2009a,b); van de Geer and Bühlmann (2009); van de Geer et al. (2010); van de Geer (2008); Zhang and Huang (2008); Zhang (2009) under conditions stated above, or other variants.

Along this line, we prove new results for both the Lasso as an initial estimator and for the thresholded estimators. In Section 1.3 and 1.4, we show oracle results for the Thresholded Lasso and the Gauss-Dantzig selector in terms of achieving the *sparse oracle inequalities* which we shall formally define in Section 1.4. While the focus of the present paper is on variable selection and oracle inequalities in terms of ℓ_2 loss, prediction errors are also explicitly derived in Section 1.5; there we introduce the oracle inequalities in terms of prediction error and show a natural interpretation for the Thresholded Lasso estimator when relating to the ℓ_0 penalized least squares estimators, in particular, ones that have been studied by Foster and George (1994); see also Barron et al. (1999); Birge and Massart (1997, 2001) for subsequent developments. In Section 1.6, we discuss recovery of a subset of strong signals.

1.3 Variable selection under the RE condition

Our first result in Theorem 1.1 shows that consistent variable selection is possible under the RE condition. We do not impose any extra constraint on s besides what is allowed in order for (1.9) to hold. Note that when $s > n/2$, it is impossible for the restricted eigenvalue assumption to hold as X_I for any I such that $|I| = 2s$ becomes singular in this case. Hence our algorithm is especially relevant if one would like to estimate a parameter β such that s is very close to n ; See Section 2 for such examples. Our analysis builds upon the

rate of convergence bounds for β_{init} derived in [Bickel et al. \(2009\)](#). The first implication of this work and also one of the motivations for analyzing the thresholding methods is: under [Assumption 1.1](#), one can obtain consistent variable selection for very significant values of s , if only a few extra variables are allowed to be included in the estimator $\widehat{\beta}$. Note that we did not optimize the lower bound on s as we focus on cases when the support S is large.

Theorem 1.1. *Suppose that $RE(s, k_0, X)$ holds with $K(s, k_0)$, where $k_0 = 1$ for the Dantzig selector and $= 3$ for the Lasso. Suppose $\lambda_n \geq f\lambda_{\sigma, a, p}$ for $\lambda_{\sigma, a, p}$ as in [\(1.15\)](#), where $f = 1$ for the Dantzig selector, and $= 2$ for the Lasso. Let $s \geq K^4(s, k_0)$. Suppose $\beta_{\min} := \min_{j \in S} |\beta_j| \geq B_4 \lambda_n \sqrt{s}$, where $B_4 = 4\sqrt{2} \max(K(s, k_0), 1) + \max(4K^2(s, k_0), \sqrt{2}/f \Lambda_{\min}(2s))$. Then on \mathcal{T}_a , the multi-step procedure returns $\widehat{\beta}$ such that for $B_3 = (1+a)(1+1/(16f^2\Lambda_{\min}^2(2s)))$,*

$$S \subseteq I := \text{supp}(\widehat{\beta}), \text{ where } |I \setminus S| < 1/(16f^2\Lambda_{\min}^2(2s)) \text{ and} \\ \|\widehat{\beta} - \beta\|_2^2 \leq \lambda_{\sigma, a, p}^2 |I| / \Lambda_{\min}^2(|I|) \leq B_3 (2 \log p/n) s \sigma^2 / (\Lambda_{\min}^2(|I|)).$$

In [Section 7](#), our simulation results using the Thresholded Lasso show that the exact recovery rate of the support is very high for a few types of random matrices once the number of samples passes a certain threshold. We note that the oracle inequality as in [\(1.12\)](#) is also achieved given that $\beta_{\min} \geq \sigma/\sqrt{n}$; hence $\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n) = s\sigma^2/n$. We next extend *model selection consistency* beyond the notion of exact recovery of the support set S as we introduced earlier, which has been considered in [Meinshausen and Bühlmann \(2006\)](#); [Wainwright \(2009b\)](#); [Zhao and Yu \(2006\)](#); Instead of having to make strong assumptions on either the signal strength, for example, on β_{\min} , or the incoherence conditions (or both), we focus on defining a meaningful criteria for *model selection consistency* when both are relatively weak.

1.4 Thresholding that achieves sparse oracle inequalities

The natural question upon obtaining [Theorem 1.1](#) is: is there a good thresholding rule that enables us to obtain a sufficiently *sparse* estimator $\widehat{\beta}$ which satisfies the *oracle inequality* as in [\(1.12\)](#), when some components of β_S (and hence β_{\min}) are well below σ/\sqrt{n} ? [Theorem 1.2](#) answers this question positively: under a uniform uncertainty principle (UUP), thresholding of an initial Dantzig selector β_{init} at the level of $C_1 \sqrt{2 \log p/n} \sigma$ for some constant C_1 , identifies a sparse model I of cardinality at most $2s_0$ such that its corresponding least-squares estimator $\widehat{\beta}$ based on the model I achieves the oracle inequality as in [\(1.12\)](#). This is accomplished without any knowledge of the significant coordinates or parameter values of β . [Theorem 1.3](#) shows that exactly the same type of sparse oracle inequalities hold for the Thresholded Lasso under the RE condition, which is both surprising but also mostly anticipated; this is also the key contribution of this paper. For simplicity, we always aim to bound $|I| < 2s_0$ while achieving the oracle inequality as in [\(1.12\)](#); One could aim to bound $|I| < cs_0$ for some other constant $c > 0$. We refer to estimators that satisfy both constraints as estimators that achieve the *sparse oracle inequalities*. Moreover, we note that thresholding of an initial estimator β_{init} which achieves ℓ_2 loss as in [\(1.12\)](#) at the level of $c_1 \sigma \sqrt{2 \log p/n}$ for some constant $c_1 > 0$, will always select nearly the best subset of variables in the spirit of [Theorem 1.2](#) and [1.3](#); Formal statements of such results are omitted.

Theorem 1.2. (Variable selection under UUP) Choose $\tau, a > 0$ and set $\lambda_n = \lambda_{p,\tau}\sigma$, where $\lambda_{p,\tau} := (\sqrt{1+a} + \tau^{-1})\sqrt{2\log p/n}$, in (1.3). Suppose β is s -sparse with $\delta_{2s} + \theta_{s,2s} < 1 - \tau$. Let threshold t_0 be chosen from the range $(C_1\lambda_{p,\tau}\sigma, C_4\lambda_{p,\tau}\sigma]$ for some constants C_1, C_4 to be defined. Then with probability at least $1 - (\sqrt{\pi\log pp^a})^{-1}$, the Gauss-Dantzig selector $\hat{\beta}$ selects a model $I := \text{supp}(\hat{\beta})$ such that we have

$$|I| \leq 2s_0 \text{ and } |I \setminus S| \leq s_0 \leq s \text{ and} \quad (1.16)$$

$$\|\hat{\beta} - \beta\|_2^2 \leq 2C_3^2 \log p \left(\sigma^2/n + \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n) \right) \quad (1.17)$$

where C_1 is defined in (4.2) and C_3 depends on $a, \tau, \delta_{2s}, \theta_{s,2s}$ and C_4 ; see (4.3).

Theorem 1.3. (Ideal model selection for the Thresholded Lasso) Suppose $RE(s_0, 6, X)$ holds with $K(s_0, 6)$, and conditions (1.4) and (1.5) hold. Let β_{init} be an optimal solution to (1.2) with $\lambda_n = d_0\sqrt{2\log p/n}\sigma \geq 2\lambda_{\sigma,a,p}$, where $a \geq 0$ and $d_0 \geq 2\sqrt{1+a}$. Suppose that we choose $t_0 = C_4\lambda\sigma$, for some constant $C_4 \geq D_1$, where $D_1 = \Lambda_{\max}(s - s_0) + 9K^2(s_0, 6)/2$; set $I = \{j \in \{1, \dots, p\} : \beta_{j,\text{init}} \geq t_0\}$. Then for $\mathcal{D} := \{1, \dots, p\} \setminus I$ and $\hat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$, we have on \mathcal{T}_a :

$$|I| \leq s_0(1 + D_1/C_4) < 2s_0, \quad |I \cup S| \leq s + s_0 \text{ and}$$

$$\|\hat{\beta} - \beta\|_2^2 \leq 2D_3^2 \log p (\sigma^2/n + \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n))$$

where D_3 depends on $a, K(s_0, 6), D_0$ and D_1 as in (5.2) and (5.3), $\Lambda_{\min}(|I|), \theta_{s,2s_0}$, and C_4 ; see (5.4).

Our analysis for Theorem 1.2 builds upon Candès and Tao (2007), which show that so long as β is sufficiently sparse the Dantzig selector as in (1.3) achieves the oracle inequality as in (1.12). Note that allowing t_0 to be chosen from a range (as wide as one would like, with the cost of increasing the constant C_3 in (1.17)), saves us from having to estimate C_1 , which indeed depends on δ_{2s} and $\theta_{s,2s}$. The same comment applies to Theorem 1.3 for D_3 . Assumption 1.2 implies that Assumption 1.1 holds for $k_0 = 1$ with $K(s, k_0) = \sqrt{\Lambda_{\min}(2s)}/(\Lambda_{\min}(2s) - \theta_{s,2s}) \leq \sqrt{\Lambda_{\min}(2s)}/(1 - \delta_{2s} - \theta_{s,2s})$ (see Bickel et al. (2009)). For a more comprehensive comparison between these conditions, we refer to van de Geer and Bühlmann (2009). We note that $RE(s_0, 6)$ is imposed on X with *sparsity* fixed at s_0 (rather than s) and $k_0 = 6$ in Theorem 5.1. Important consequences of this result is shown in Section 1.5. The term *sparsity oracle inequalities* has also been used in the literature, which is targeted at bounding prediction errors of the estimators with the best sparse approximation of the regression function known by an oracle; see Bickel et al. (2009) and more references therein. It would be interesting to explore such properties for the Thresholded Lasso under the RE conditions.

1.5 Connecting to the ℓ_0 penalized least squares estimators

Now why is the bound of $|I| \leq 2s_0$ interesting? We wish to point out that this would make the behavior of the Thresholded Lasso procedure somehow mimic that of the ℓ_0 penalized estimators, which is computational inefficient, as we introduce next. It is clear that for the least squares estimator based on I ,

$\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$, it holds that

$$\|X \widehat{\beta}_I - X \beta\|_2^2 = \|P_I(X \beta + \epsilon) - X \beta\|_2^2 = \|(P_I - \text{Id})X_{I^c} \beta_{I^c} + P_I \epsilon\|_2^2 \quad (1.18)$$

$$\text{and hence } \mathbb{E}\|X \widehat{\beta}_I - X \beta\|_2^2/n = \|(P_I - \text{Id})X_{I^c} \beta_{I^c}\|_2^2 + |I|\sigma^2, \quad (1.19)$$

which again shows the typical bias and variance tradeoff. Consider the best model I_0 upon which $\widehat{\beta}_{I_0} = (X_{I_0}^T X_{I_0})^{-1} X_{I_0}^T Y$ achieves the minimum in (1.19):

$$I_0 = \arg \min_{I \subset \{1, \dots, p\}} \|(P_I - \text{Id})X_{I^c} \beta_{I^c}\|_2 + |I|\sigma^2.$$

Now the question is: can one do nearly as well as $\widehat{\beta}_{I_0}$ in the sense of achieving mean square error within $\log p$ factor of $\mathbb{E}\|X \widehat{\beta}_{I_0} - X \beta\|_2^2$? It turns out that the answer is yes, if one solves the following ℓ_0 penalized least squares estimator with $\lambda_0 = \sqrt{\log p/n}$, as proposed in the RIC procedure (Foster and George, 1994):

$$\widehat{\beta} = \arg \min_{\beta} \|Y - X \beta\|_2^2/(2n) + \lambda_0^2 \sigma^2 \|\beta\|_0, \quad (1.20)$$

where $\|\beta\|_0$ is the number of nonzero components in β . This is shown in a series of papers in Barron et al. (1999); Birge and Massart (1997, 2001); Foster and George (1994). We refer to Barron et al. (1999); Foster and George (1994) for other procedures related to (1.20). Note that $\|Y - X \beta\|_2^2 \leq 2\|X \widehat{\beta} - X \beta\|_2^2 + 2\|\epsilon\|_2^2$; hence we only need to look at the tradeoff between $\|X \widehat{\beta} - X \beta\|_2^2$ and $\log p|I|$. Note that $\|X \widehat{\beta} - X \beta\|_2^2$ would be 0 if $\widehat{\beta} = \beta$, but $|I|$ would be large. Theorem 1.4 shows that (a) the thresholded estimators achieve a balance between the ‘‘complexity’’ measure $\log p|I|$ and $\|X \widehat{\beta} - X \beta\|_2^2$ which now have the same order of magnitude; (b) and in some sense, variables in model I are essential in predicting $X \beta$.

Theorem 1.4. *Let I be the model selected by thresholding an initial estimator β_{init} , under conditions as described in Theorem 1.2 or Theorem 1.3. Let $\mathcal{D} := \{1, \dots, p\} \setminus I$. Let s_0 be as defined in (1.14) and $\lambda = \sqrt{2 \log p/n}$. For $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$ and some constant C , we have on \mathcal{T}_a ,*

$$\frac{\|X \widehat{\beta}_I - X \beta\|_2}{\sqrt{n}} \leq \sqrt{\Lambda_{\max}(s)} \|\beta_{\mathcal{D}}\|_2 + \frac{\sqrt{|I| \Lambda_{\max}(|I|)} \lambda_{\sigma, a, p}}{\Lambda_{\min}(|I|)} \leq C \lambda \sigma \sqrt{s_0}.$$

Comparing (1.20) and (1.2), it is clear that for entries $\beta_{j, \text{init}} < \lambda_0 \sigma$ in a Lasso estimator, their contributions to the optimization function in (1.20) will be larger than that in (1.2) if $\lambda_n = \lambda_0 \sigma$; hence removing these entries from the initial estimator in some sense recovers the choices that would have been made by the complexity-based function as in (1.20). Put in another way, getting rid of variables $\{j : \beta_{j, \text{init}} < \lambda_0 \sigma\}$ from the solution to (1.2) with $\lambda_n \asymp \lambda_0 \sigma$ is in some way restoring the behavior of (1.20) in a brute-force manner. Proposition 1.5 (by setting $c' = 1$) shows that the number of variables in β at above and around $\sqrt{\log p/n} \sigma$ in magnitude is bounded by $2s_0$ (One could choose another target set: for example, $\{j : |\beta_j| \geq \sqrt{\log p/(c'n)} \sigma\}$, for some $c' > 1/2$.) Roughly speaking, we wish to include most of them by leaving $2s_0$ variables in the model I . Such connections will be made precise in our future work.

1.6 Controlling Type II errors

In Section 6 (cf. Theorem 6.3), we show that we can recover a subset S_L of variables accurately, where $S_L := \{j : |\beta_j| > \sqrt{2 \log p/n} \sigma\}$, under Assumption 1.1 when $\beta_{\min, S_L} := \min_{j \in S_L} |\beta_j|$ is large enough

(relative to the ℓ_2 loss of an initial estimator under the RE condition on the set S_L); in addition, a small number of extra variables from $\{1, \dots, p\} \setminus T_0 =: T_0^c$ are possibly also included in the model I , where T_0 denotes positions of the s_0 largest coefficients of β in absolute values. In this case, it is also possible to get rid of variables from T_0^c entirely by increasing the threshold t_0 while making the lower bound on β_{\min, S_L} a constant times stronger. We omit such details from the paper. Hence compared to Theorem 1.1, we have relaxed the restriction on β_{\min} : rather than requiring all non-zero entries to be large, we only require those in a subset S_L to be recovered to be large. In addition, we believe that our analysis can be extended to cases when β is not exactly sparse, but has entries decaying like a power law, for example, as studied by Candès and Tao (2007); We end with Proposition 1.5. For a set A , we use $|A|$ to denote its cardinality.

Proposition 1.5. *Let T_0 denote positions of the s_0 largest coefficients of β in absolute values. where s_0 is defined in (1.14). Let $a_0 = |S_L|$ (cf. (6.1)). Then $\forall c' > 1/2$, we have $\left| \{j \in T_0^c : |\beta_j| \geq \sqrt{\log p / (c'n)} \sigma \} \right| \leq (2c' - 1)(s_0 - a_0)$.*

1.7 Previous work

We briefly review related work in multi-step procedures and the role of sparsity for high-dimensional statistical inference. Before this work, hard thresholding idea has been shown in Candès and Tao (2007) (via Gauss-Dantzig selector) as a method to correct the bias of the initial Dantzig selector. The empirical success of the Gauss-Dantzig selector in terms of improving the statistical accuracy is strongly evident in their experimental results. Our theoretical analysis on the oracle inequalities, which hold for the Gauss-Dantzig selector under a uniform uncertainty principle, builds upon their theoretical analysis of the initial Dantzig selector under the same condition. For the Lasso, Meinshausen and Yu (2009) has also shown in theoretical analysis that thresholding is effective in obtaining a two-step estimator $\hat{\beta}$ that is consistent in its support with β when β_{\min} is sufficiently large; As pointed out by Bickel et al. (2009), a weakening of their condition is still sufficient for Assumption 1.1 to hold.

The sparse recovery problem under arbitrary noise is also well studied, see Candès et al. (2006); Needell and Tropp (2008); Needell and Vershynin (2009). Although as argued in Candès et al. (2006) and Needell and Tropp (2008), the best accuracy under arbitrary noise has essentially been achieved in both work, their bounds are worse than that in Candès and Tao (2007) (hence the present paper) under the stochastic noise as discussed in the present paper; Moreover, greedy algorithms in Needell and Tropp (2008); Needell and Vershynin (2009) require s to be part of the input, while algorithms in the present paper do not have such a requirement, and hence adapt to the unknown level of sparsity well. A more general framework on multi-step variable selection was studied by Wasserman and Roeder (2009). They control the probability of false positives at the price of false negatives, similar to what we aim for here; their analysis is constrained to the case when s is a constant. Recently, another two-stage procedure that is also relevant has been proposed in Zhang (2009), where in the second stage “selective penalization” is being applied to the set of *irrelevant features* which are defined as those below a certain threshold in the initial Lasso estimator; Incoherence conditions there are sufficiently different from the RE condition as we study in this paper for the Thresholded Lasso. Under conditions similar to Theorem 1.1, Zhou et al. (2009) requires $s = O(\sqrt{n / \log p})$ in order to achieve variable selection consistency using the adaptive Lasso (Zou, 2006) (see also Huang et al. (2008)), as the

second step procedure. Concurrent with the present work, the authors have revisited the adaptive Lasso and derived bounds in terms of prediction error [van de Geer et al. \(2010\)](#); there the number of false positives is also aimed at being in the same order as that of the set of *significant* variables which predicts $X\beta$ well; in addition, the adaptive Lasso method is compared with thresholding methods, under a stronger incoherence condition than the RE condition studied in the present paper. While the focus of the present paper is on variable selection and oracle inequalities for the ℓ_2 loss, prediction errors of the OLS estimators $\hat{\beta}$ are also explicitly derived; We also compare the performance in terms of variable selections between the adaptive and the thresholding methods in our simulation study, which is reported in Section 7.

Parts of this work was presented in a conference paper [Zhou \(2009b\)](#). The current version expands the original idea and elaborates upon the conceptual connections between the Thresholded Lasso and ℓ_0 penalized methods; in addition, we provide new results on the sparse oracle inequalities under the RE condition (cf. Theorem 1.3, Theorem 5.1 and Theorem 6.3).

1.8 Organization of the paper

Section 2 briefly discusses the relationship between linear sparsity and random design matrices, while highlighting the role thresholding plays in terms of recovering the best subset of variables, when s is a linear fraction of n , which in turn is a nonnegligible fraction of p . We prove Theorem 1.1 essentially in Section 3. A thresholding framework for the general setting is described in Section 4, which also sketches the proof of Theorem 1.2. The proof of Theorem 1.3 is shown in Section 5, where oracle inequalities for the original Lasso estimator is also shown. In Section 6, we show conditions under which one recovers a subset of strong signals. Section 7 includes simulation results showing that the Thresholded Lasso is consistent with our theoretical analysis on variable selection and on estimating β . Most of the technical proofs are included in the Appendix.

2 Linear sparsity and random matrices

A special case of design matrices that satisfy the Restricted Eigenvalue assumption are the random design matrices. This is shown in a large body of work, for example [Baraniuk et al. \(2008\)](#); [Candès et al. \(2006\)](#); [Candès and Tao \(2005, 2007\)](#); [Donoho \(2006b\)](#); [Mendelson et al. \(2008\)](#); [Szarek \(1991\)](#), which shows that the UUP holds for “generic” or random design matrices for very significant values of s . It is well known that for a random matrix the UUP holds for $s \asymp n/\log(p/n)$ with i.i.d. Gaussian random variables, subject to normalizations of columns, the Bernoulli, and in general the subgaussian random ensembles [Baraniuk et al. \(2008\)](#); [Mendelson et al. \(2008\)](#); [Adamczak et al. \(2009\)](#) show that UUP holds for $s \asymp n/\log^2(p/n)$ when X is a random matrix composed of columns that are independent isotropic vectors with log-concave densities. Hence this setup only requires Cs observations per nonzero value in β , where C is a small constant, when n is a nonnegligible fraction of p , in order to recover β ; we call this level of sparsity the linear sparsity. Our simulation results in Section 7 show that once $n \geq Cs \log(p/n)$, where C is a small constant, exact recovery rate of the sparsity pattern is very high for Gaussian (and Bernoulli) random ensembles, when

β_{\min} is sufficiently large; this shows a strong contrast with the ordinary Lasso, for which the probability of success in terms of exact recovery of the sparsity pattern tends to zero when $n < 2s \log(p-s)$ (Wainwright, 2009b).

A series of recent papers Raskutti et al. (2009); Zhou (2009a); Zhou et al. (2009) show that a broader class of subgaussian random matrices also satisfy the Restricted Eigenvalue condition; In particular, Zhou (2009a) shows that for subgaussian random matrices Ψ which are now well known to satisfy the UUP condition under linear sparsity, RE condition holds for $X := \Psi \Sigma^{1/2}$ with overwhelming probability with $n \asymp s \log(p/n)$ number of samples, where Σ is assumed to satisfy the follow condition: Suppose $\Sigma_{jj} = 1, \forall j = 1, \dots, p$, and for some integer $1 \leq s \leq p$ and a positive number k_0 , the following condition holds for all $v \neq 0$:

$$\frac{1}{K(s, k_0, \Sigma)} := \min_{\substack{J_0 \subseteq \{1, \dots, p\}, \\ |J_0| \leq s}} \min_{\|v_{J_0^c}\|_1 \leq k_0 \|v_{J_0}\|_1} \left\| \Sigma^{1/2} v \right\|_2 / \|v_{J_0}\|_2 > 0.$$

Thus the additional covariance structure Σ is explicitly introduced to the columns of Ψ in generating X . We believe similar results can be extended to other cases: for example, when X is the composition of a random Fourier ensemble, or randomly sampled rows of orthonormal matrices, see for example Candès and Tao (2006, 2007); Rudelson and Vershynin (2006), where the UUP holds for $s = O(n/\log^c p)$ for some constant $c > 0$.

3 Thresholding procedure when β_{\min} is large

In this section, we use a penalization parameter $\lambda_n \geq B\lambda_{\sigma, a, p}$ and assume $\beta_{\min} > C\lambda_n \sqrt{s}$ for some constants B, C ; we first specify the thresholding parameters in this case. We then show in Theorem 3.1 that our algorithm works under any condition so long as the rate of convergence of the initial estimator obeys the bounds in (3.2). Theorem 1.1 is a corollary of Theorem 3.1 under Assumption 1.1, given the rate of convergence bounds for β_{init} following derivations in (Bickel et al., 2009).

The Iterative Procedure. We obtain an initial estimator β_{init} using the Lasso or the Dantzig selector. Let $\widehat{S}_0 = \{j : \beta_{j, \text{init}} > 4\lambda_n\}$, and $\widehat{\beta}^{(0)} := \beta_{\text{init}}$; Iterate through the following steps twice, for $i = 0, 1$: (a) Set $t_i = 4\lambda_n \sqrt{|\widehat{S}_i|}$; (b) Threshold $\widehat{\beta}^{(i)}$ with t_i to obtain $I := \widehat{S}_{i+1}$, where

$$\widehat{S}_{i+1} = \left\{ j \in \widehat{S}_i : \widehat{\beta}_j^{(i)} \geq 4\lambda_n \sqrt{|\widehat{S}_i|} \right\} \quad (3.1)$$

and compute $\widehat{\beta}_I^{(i+1)} = (X_I^T X_I)^{-1} X_I^T Y$. Return the final set of variables in \widehat{S}_2 and output $\widehat{\beta}$ such that $\widehat{\beta}_{\widehat{S}_2} = \widehat{\beta}_{\widehat{S}_2}^{(2)}$ and $\widehat{\beta}_j = 0, \forall j \in \widehat{S}_2^c$.

Theorem 3.1. Let $\lambda_n \geq B\lambda_{\sigma, a, p}$, where $B \geq 1$ is a constant suitably chosen such that the initial estimator β_{init} satisfies on some event Q_b , for $v_{\text{init}} = \beta_{\text{init}} - \beta$,

$$\|v_{\text{init}, S}\|_2 \leq B_0 \lambda_n \sqrt{s} \text{ and } \|\beta_{\text{init}, S^c}\|_1 \leq B_1 \lambda_n s \quad (3.2)$$

where B_0, B_1 are some constants. Suppose for $B_2 = 1/(B\Lambda_{\min}(2s))$,

$$\beta_{\min} \geq \left(\max\left(\sqrt{B_1}, 2\right) 2\sqrt{2} + \max\left(B_0, \sqrt{2}B_2\right) \right) \lambda_n \sqrt{s}. \quad (3.3)$$

Then for $s \geq B_1^2/16$, it holds on $\mathcal{T}_a \cap Q_b$ that, (a): $\forall i = 1, 2$, $|\widehat{S}_i| \leq 2s$; and (b):

$$\|\widehat{\beta}^{(i)} - \beta\|_2 \leq \lambda_{\sigma, a, p} \sqrt{|\widehat{S}_i|/\Lambda_{\min}(|\widehat{S}_i|)} \leq \lambda_n B_2 \sqrt{2s} \quad (3.4)$$

where $\forall i = 1, 2$, $\widehat{\beta}^{(i)}$ are the OLS estimators based on \widehat{S}_i ; Moreover, the Iterative Procedure includes the set of relevant variables in \widehat{S}_2 such that $S \subseteq \widehat{S}_2 \subseteq \widehat{S}_1$ and

$$|\widehat{S}_2 \setminus S| := |\text{supp}(\widehat{\beta}) \setminus S| \leq 1/(16B^2 \Lambda_{\min}^2(|\widehat{S}_1|)) \leq B_2^2/16. \quad (3.5)$$

The proof of Theorem 3.1 appears in Section D. We now discuss its relationship to theorems in the subsequent sections. We first note that in order to obtain \widehat{S}_1 such that $|\widehat{S}_1| \leq 2s$ and $\widehat{S}_1 \supseteq S$ as above, we only need to threshold β_{init} at $t_0 = B_1 \lambda_n$; here instead of having to estimate the unknown B_1 , we can use $t_0 = c_0 \lambda_n \sqrt{s}$ for some constant c_0 to threshold β_{init} . In the general setting, we require that t_0 be chosen from the range $(C_1 \lambda_n, C_4 \lambda_n]$ for some constants C_1, C_4 to be specified; see Section 4 (Lemma 4.2) for example. We note that without the knowledge of σ , one could use $\widehat{\sigma} \geq \sigma$ in λ_n ; this will put a stronger requirement on β_{min} , but all conclusions of Theorem 3.1 hold. When β_{min} does not satisfy the constraint as in Theorem 3.1, we cannot really guarantee that all variables in S will be chosen. Hence (3.2) will be replaced by requirements on T_0 , which denotes locations of the s_0 largest coefficients of β in absolute values: ideally, we wish to have

$$\|(\beta_{\text{init}} - \beta)_{T_0}\|_2 \leq C_0 \lambda_n \sqrt{|T_0|} \quad \text{and} \quad \|\beta_{\text{init}, T_0^c}\|_1 \leq C_1 \lambda_n |T_0|; \quad (3.6)$$

for some constants C_0, C_1 , so that (1.16) and (1.17) hold under suitably chosen thresholding rules. This is the content of Theorem 5.1 and Theorem 6.3.

4 Nearly ideal model selections under the UUP

In this section, we wish to derive a meaningful criteria for consistency in variable selection, when β_{min} is well below the noise level. Suppose that we are given an initial estimator β_{init} that achieves the oracle inequality as in (1.12), which adapts nearly ideally not only to the uncertainty in the support set S but also the ‘‘significant’’ set. We show that although we cannot guarantee the presence of variables indexed by $S_R = \{j : |\beta_j| \leq \sigma \sqrt{2 \log p/n}\}$ to be included in the final set I (cf. (4.7)) due to their lack of strength, we wish to include in I most variables in $S_L = S \setminus S_R$ such that the OLS estimator based on I achieves (1.12) even though some non-zero variables are missing from I . Here we pay a price for the missing variables in order to obtain a sufficiently sparse model I . Toward this goal, we analyze the following algorithm.

The General Two-step Procedure: Assume $\delta_{2s} + \theta_{s, 2s} < 1 - \tau$, where $\tau > 0$;

1. First obtain an initial estimator β_{init} using the Dantzig selector in (1.3) with $\lambda_n = (\sqrt{1+a} + \tau^{-1}) \sqrt{2 \log p/n} \sigma$, where $a \geq 0$; then threshold β_{init} with t_0 , chosen from the range $(C_1 \lambda_{p, \tau} \sigma, C_4 \lambda_{p, \tau} \sigma]$, for C_1 as defined in (4.2), to obtain a set I of cardinality at most $2s_0$ (cf. Lemma 4.2):
set $I := \{j \in \{1, \dots, p\} : \beta_{j, \text{init}} \geq t_0\}$.

2. Given a set I as above, run the OLS regression to obtain $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$ and set $\widehat{\beta}_j = 0, \forall j \notin I$.

In Section 5, we analyze the Thresholded Lasso, where we obtain β_{init} via the Lasso under the RE condition and follow the same steps as above; see Theorem 5.1 and Lemma 5.2 for the new λ_n and t_0 to be specified. Under the UUP, Candès and Tao (2007) have shown that the Dantzig selector achieves nearly the ideal level of ℓ_2 loss. We then show in Lemma 4.2 that thresholding at the level of $C_1 \lambda \sigma$ at Step 1 selects a set I of at most $2s_0$ variables, among which at most s_0 are from S^c .

Proposition 4.1. (Candès and Tao, 2007) *Let $Y = X\beta + \epsilon$, for ϵ being i.i.d. $N(0, \sigma^2)$ and $\|X_j\|_2^2 = n$. Choose $\tau, a > 0$ and set $\lambda_n = (\sqrt{1+a} + \tau^{-1})\sigma\sqrt{2\log p/n}$ in (1.3). Then if β is s -sparse with $\delta_{2s} + \theta_{s,2s} < 1 - \tau$, the Dantzig selector obeys with probability at least $1 - (\sqrt{\pi \log pp^a})^{-1}$, $\|\widehat{\beta} - \beta\|_2^2 \leq 2C_2^2(\sqrt{1+a} + \tau^{-1})^2 \log p (\sigma^2/n + \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n))$.*

From this point on we let $\delta := \delta_{2s}$ and $\theta := \theta_{s,2s}$; Analysis in Candès and Tao (2007) (Theorem 2) and the current paper yields the following constants,

$$C_2 = 2C'_0 + \frac{1+\delta}{1-\delta-\theta} \quad \text{where } C'_0 = \frac{C_0}{1-\delta-\theta} + \frac{\theta(1+\delta)}{(1-\delta-\theta)^2}, \quad (4.1)$$

where $C_0 = 2\sqrt{2} \left(1 + \frac{1-\delta^2}{1-\delta-\theta}\right) + (1 + 1/\sqrt{2}) \frac{(1+\delta)^2}{1-\delta-\theta}$; We now define

$$C_1 = C'_0 + \frac{1+\delta}{1-\delta-\theta} \quad \text{and} \quad (4.2)$$

$$C_3^2 = 3(\sqrt{1+a} + \tau^{-1})^2((C'_0 + C_4)^2 + 1) + 4(1+a)/\Lambda_{\min}^2(2s_0) \quad (4.3)$$

where C_3 has not been optimized. Recall that s_0 is the smallest integer such that $\sum_{i=1}^p \min(\beta_i^2, \lambda^2 \sigma^2) \leq s_0 \lambda^2 \sigma^2$, where $\lambda = \sqrt{2\log p/n}$. We order the β_j 's in decreasing order of magnitude

$$|\beta_1| \geq |\beta_2| \dots \geq |\beta_p|. \quad (4.4)$$

Thus by definition of s_0 , the fact $0 \leq s_0 \leq s$, we have for $s < p$,

$$s_0 \lambda^2 \sigma^2 \leq \lambda^2 \sigma^2 + \sum_{i=1}^p \min(\beta_i^2, \lambda^2 \sigma^2) \leq 2 \log p \left(\frac{\sigma^2}{n} + \sum_{i=1}^p \min\left(\beta_i^2, \frac{\sigma^2}{n}\right) \right) \quad (4.5)$$

$$s_0 \lambda^2 \sigma^2 \geq \sum_{j=1}^{s_0+1} \min(\beta_j^2, \lambda^2 \sigma^2) \geq (s_0 + 1) \min(\beta_{s_0+1}^2, \lambda^2 \sigma^2) \quad (4.6)$$

which implies that (as shown in Candès and Tao (2007)) that $\min(\beta_{s_0+1}^2, \lambda^2 \sigma^2) < \lambda^2 \sigma^2$ and hence by (4.4), it holds that

$$|\beta_j| < \lambda \sigma \quad \text{for all } j > s_0. \quad (4.7)$$

Lemma 4.2. *Choose $\tau > 0$ such that $\delta_{2s} + \theta_{s,2s} < 1 - \tau$. Let β_{init} be the solution to (1.3) with $\lambda_n = \lambda_{p,\tau} \sigma := (\sqrt{1+a} + \tau^{-1})\sqrt{2\log p/n}\sigma$. Given some constant $C_4 \geq C_1$, for C_1 as in (4.2), choose a thresholding parameter t_0 such that $C_4 \lambda_{p,\tau} \sigma \geq t_0 > C_1 \lambda_{p,\tau} \sigma$ and set $I = \{j : |\beta_{j,\text{init}}| \geq t_0\}$. Then with probability at least $\mathbb{P}(\mathcal{I}_a)$, as detailed in Proposition 4.1, we have (1.16), and for C'_0 as in (4.1), $\|\beta_{\mathcal{D}}\|_2 \leq \sqrt{(C'_0 + C_4)^2 + 1} \lambda_{p,\tau} \sigma \sqrt{s_0}$, where $\mathcal{D} := \{1, \dots, p\} \setminus I$.*

It is clear by Lemma 4.2 that we cannot cut too many “significant” variables; in particular, for those that are $> \lambda\sigma\sqrt{s_0}$, we can cut at most a constant number of them. Next we show that even if we miss some columns of X in S , we can still hope to get the ℓ_2 loss as required in Theorem 1.2 so long as $\|\beta_{\mathcal{D}}\|_2$ is bounded, for example, as bounded in Lemma 4.2, and I is sufficiently sparse. Now Theorem 1.2 is an immediate corollary of Lemma 4.2 and 4.3 in view of (4.5). See Section E for its proof. We note that Lemma 4.3 yields a general result on the ℓ_2 loss for the OLS estimator, when a subset of relevant variables is missing from the chosen model I ; this is also an important technical contribution of this paper.

Lemma 4.3. (OLS estimator with missing variables) *Suppose that (1.4) and (1.5) hold. Let $\mathcal{D} := \{1, \dots, p\} \setminus I$ and $S_{\mathcal{D}} = \mathcal{D} \cap S$ such that $I \cap S_{\mathcal{D}} = \emptyset$. Suppose $|I \cup S_{\mathcal{D}}| \leq 2s$. Then, for $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$, it holds on \mathcal{T}_a that*

$$\left\| \widehat{\beta}_I - \beta \right\|_2^2 \leq \left(\theta_{|I|, |S_{\mathcal{D}}|} \|\beta_{\mathcal{D}}\|_2 + \lambda_{\sigma, a, p} \sqrt{|I|} \right)^2 / \Lambda_{\min}^2(|I|) + \|\beta_{\mathcal{D}}\|_2^2.$$

We note that Lemma 4.3 applies to X so long as conditions (1.4) and (1.5) hold, which guarantees that $\theta_{|I|, |S_{\mathcal{D}}|}$ is bounded within a reasonable constant, when $|I| + |S_{\mathcal{D}}| \leq 2s$ (cf. Lemma 5.4). It is clear from Lemma 4.3 and Theorem 1.4 that, except for the constants that appear before each term, namely, $\|\beta_{\mathcal{D}}\|_2$ and $\sqrt{|I|}\sqrt{2\log p}\sigma$, the bias and variance tradeoffs for the prediction error and the ℓ_2 loss follow roughly the same trend in their upper bounds. It will make sense to take a look at the bound on prediction error for the Gauss-Dantzig selector stated in Corollary 4.4, which follows immediately from Theorem 1.4 and Lemma 4.2.

Corollary 4.4. *Under conditions in Theorem 1.2, the Gauss-Dantzig selector chooses I , where $|I| \leq 2s_0$, such that for the OLS estimator $\widehat{\beta}$ based on I , we have $\left\| X\widehat{\beta} - X\beta \right\|_2 / \sqrt{n} \leq C_5 \sqrt{s_0} \lambda \sigma$, where $C_5 = \sqrt{\Lambda_{\max}(s)} (\sqrt{(C'_0 + C_4)^2 + 1} (\sqrt{1+a} + \tau^{-1})) + f(I)$, where $f(I) := \sqrt{2(1+a)} \Lambda_{\max}(|I|) / \Lambda_{\min}(|I|)$.*

5 On sparse oracle inequalities of the Lasso under the RE condition

In this section, in order to prove Theorem 1.3, we first show in Theorem 5.1 that under the RE condition, the Lasso estimator achieves essentially the same type of oracle properties as the Dantzig selector (under UUP). This result is new to the best of our knowledge; it improves upon a result in Bickel et al. (2009) (cf. Theorem 7.2) under slightly different RE conditions, and thus may be of independent interests. The sparse oracle properties of the Thresholded Lasso in terms of variable selection, ℓ_2 loss, and prediction error then all follow naturally from Theorem 5.1, Lemma 5.2 and Lemma 4.3 as derived in Section 4. The proof of Theorem 5.1 draws upon techniques from a concurrent work in van de Geer et al. (2010), where a stronger condition is required, while deriving bounds similar to the present paper.

Theorem 5.1. (Oracle inequalities of the Lasso) *Let $Y = X\beta + \epsilon$, for ϵ being i.i.d. $N(0, \sigma^2)$ and $\|X_j\|_2 = \sqrt{n}$. Let s_0 be as in (1.14) and T_0 denote locations of the s_0 largest coefficients of β in absolute values. Suppose that $RE(s_0, 6, X)$ holds with $K(s_0, 6)$, and (1.4) and (1.5) hold. Let β_{init} be an optimal solution to (1.2) with $\lambda_n = d_0 \lambda \sigma \geq 2\lambda_{\sigma, a, p}$, where $a \geq 0$ and $d_0 \geq 2\sqrt{1+a}$. Let $h = \beta_{\text{init}} - \beta_{T_0}$. Then on*

\mathcal{T}_a as in (1.15), we have for $\Lambda_{\max} := \Lambda_{\max}(s - s_0)$,

$$\begin{aligned} \|\beta_{\text{init}} - \beta\|_2^2 &\leq 2\lambda^2\sigma^2s_0(D_0^2 + D_1^2 + 1), \\ \|h_{T_0}\|_1 + \|\beta_{\text{init}, T_0^c}\|_1 &\leq \left(\frac{2\Lambda_{\max}}{d_0} + \max \left\{ 8K^2(s_0, 6)d_0, \frac{\Lambda_{\max}}{3d_0} \right\} \right) \lambda\sigma s_0, \\ \|X\beta_{\text{init}} - X\beta\|_2 / \sqrt{n} &\leq \lambda\sigma\sqrt{s_0}(\sqrt{\Lambda_{\max}} + 3d_0K(s_0, 6)) \end{aligned}$$

where D_0, D_1 are defined in (5.2) and (5.3). Moreover, for any subset $I_0 \subset S$, by assuming that $RE(|I_0|, 6, X)$ holds with $K(|I_0|, 6)$, we have

$$\|X\beta_{\text{init}} - X\beta\|_2^2 / n \leq 2\|X\beta - X\beta_{I_0}\|_2^2 / n + 9\lambda_n^2|I_0|K^2(|I_0|, 6). \quad (5.1)$$

Let T_1 denote the s_0 largest positions of h in absolute values outside of T_0 ; Let $T_{01} := T_0 \cup T_1$. The proof of Theorem 5.1 yields the following bounds: for $K := K(s_0, 6)$, $\|h_{T_{01}}\|_2 \leq D_0\lambda\sigma\sqrt{s_0}$ and $\|h_{T_0^c}\|_1 \leq D_1\lambda\sigma s_0$ where

$$D_0 = \max\{D, K\sqrt{2}(2\sqrt{\Lambda_{\max}(s - s_0)} + 3d_0K)\}, \quad (5.2)$$

$$\text{where } D = (\sqrt{2} + 1) \frac{\sqrt{\Lambda_{\max}(s - s_0)}}{\sqrt{\Lambda_{\min}(2s_0)}} + \frac{\theta_{s_0, 2s_0}\Lambda_{\max}(s - s_0)}{\Lambda_{\min}(2s_0)} \text{ and}$$

$$D_1 = 2\Lambda_{\max}(s - s_0)/d_0 + 9K^2d_0/2. \quad (5.3)$$

The proof of Lemma 5.2 follows exactly that of Lemma 4.2, and hence omitted. We then state the bound on prediction error for $\hat{\beta}$ for the Thresholded Lasso, which follows immediately from Theorem 1.4 and Lemma 5.2.

Lemma 5.2. *Suppose that X obeys $RE(s_0, 6, X)$, and conditions (1.4) and (1.5) hold. Let β_{init} be an optimal solution to (1.2) with $\lambda_n = d_0\lambda\sigma \geq 2\lambda_{\sigma, a, p}$, where $a \geq 0$, $d_0 \geq 2\sqrt{1+a}$, and $\lambda := \sqrt{2\log p/n}$ as in Theorem 5.1. Suppose that we choose $t_0 = C_4\lambda\sigma$ for some positive constant C_4 . Let $I = \{j : |\beta_{j, \text{init}}| \geq t_0\}$ and $\mathcal{D} := \{1, \dots, p\} \setminus I$. Then we have on \mathcal{T}_a ,*

$$|I| \leq s_0(1 + D_1/C_4) \text{ and } |I \cup S| \leq s + D_1s_0/C_4 \text{ and}$$

$$\|\beta_{\mathcal{D}}\|_2 \leq \sqrt{(D_0 + C_4)^2 + 1}\lambda\sigma\sqrt{s_0}, \text{ where } D_0, D_1 \text{ are as defined in (5.2) and (5.3).}$$

Corollary 5.3. *Under conditions in Theorem 1.3, the Thresholded Lasso chooses I , where $|I| \leq 2s_0$, such that for the OLS estimator $\hat{\beta}$ based on I , it holds that $\|X\hat{\beta}_I - X\beta\|_2 / \sqrt{n} \leq C_6\sqrt{s_0}\lambda\sigma$, where $C_6 = \sqrt{\Lambda_{\max}(s)}\sqrt{(D_0 + C_4)^2 + 1} + f(I)$, for $f(I)$ as defined in Corollary 4.4 and D_0 is defined in (5.2).*

We now state Lemma 5.4, which follows from Candès and Tao (2005) (Lemma 1.2); we then prove Theorem 1.3, where we give an explicit expression for D_3 .

Lemma 5.4. (Candès and Tao, 2005) *Suppose that (1.4) and (1.5) hold. Then for all disjoint sets $I, S_{\mathcal{D}} \subseteq \{1, \dots, p\}$ of cardinality $|S_{\mathcal{D}}| < s$ and $|I| + |S_{\mathcal{D}}| \leq 2s$,*

$$\theta_{|I|, |S_{\mathcal{D}}|} \leq (\Lambda_{\max}(2s) - \Lambda_{\min}(2s))/2;$$

In particular, if $\delta_{2s} < 1$, we have $\theta_{|I|, |S_{\mathcal{D}}|} \leq \delta_{|I|+|S_{\mathcal{D}}|} \leq \delta_{2s} < 1$.

Proof of Theorem 1.3. It holds by definition of $S_{\mathcal{D}}$ that $I \cap S_{\mathcal{D}} = \emptyset$. It is clear by Lemma 5.2 that for $C_4 \geq D_1$, $|I| \leq 2s_0$ and $|I \cup S_{\mathcal{D}}| \leq |I \cup S| \leq s + s_0 \leq 2s$, given that $|S_{\mathcal{D}}| < s$. We have by Lemma 4.3

$$\begin{aligned} \left\| \widehat{\beta}_I - \beta \right\|_2^2 &\leq \|\beta_{\mathcal{D}}\|_2^2 \left(1 + \frac{2\theta_{|I|, |S_{\mathcal{D}}|}^2}{\Lambda_{\min}^2(|I|)} \right) + \frac{2|I|}{\Lambda_{\min}^2(|I|)} \lambda_{\sigma, a, p}^2 \\ &\leq D_3^2 \lambda^2 \sigma^2 s_0 \leq 2D_3^2 \log p \left(\sigma^2/n + \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n) \right) \text{ where} \end{aligned}$$

$$D_3^2 = ((D_0 + C_4)^2 + 1) \left(1 + 2\theta_{|I|, |S_{\mathcal{D}}|}^2 / \Lambda_{\min}^2(|I|) \right) + 4(1 + a) / \Lambda_{\min}^2(|I|). \quad \square$$

It is clear by Lemma 5.4 that

$$D_3^2 \leq ((D_0 + C_4)^2 + 1) \left(1 + \frac{(\Lambda_{\max}(2s) - \Lambda_{\min}(2s))^2}{2\Lambda_{\min}^2(|I|)} \right) + \frac{4(1 + a)}{\Lambda_{\min}^2(|I|)}. \quad (5.4)$$

6 Controlling Type-II errors

In this section, we derive results that are parametrized based on the performance of an initial estimator, the smallest magnitude of variables in $\{j : |\beta_j| > \lambda\sigma\}$, where $\lambda := \sqrt{2 \log p/n}$, and the choice of the thresholding parameter t_0 . We emphasize that we do not necessarily require that $t_0 > \lambda\sigma$. We first introduce some more notation. Again order the β_j 's in decreasing order of magnitude: $|\beta_1| \geq |\beta_2| \dots \geq |\beta_p|$. Let $T_0 = \{1, \dots, s_0\}$. In view of (4.7), we decompose $T_0 = \{1, \dots, s_0\}$ into two sets: A_0 and $T_0 \setminus A_0$, where A_0 contains the set of coefficients of β strictly larger than $\lambda\sigma$, for which we define a constant:

$$A_0 = \{j : |\beta_j| > \lambda\sigma\} =: \{1, \dots, a_0\}; \quad \text{Let } \beta_{\min, A_0} := \min_{j \leq a_0} |\beta_j| > \lambda\sigma. \quad (6.1)$$

Our goal is to show when β_{\min, A_0} is sufficiently large, we have $A_0 \subset I$ while achieving the sparse oracle inequalities; This is shown in Theorem 6.3 under the RE condition, which is stated as a corollary of Lemma 6.2. First note that changing the coefficients of β_{A_0} will not change the values of s_0 or a_0 , so long as their absolute values stay strictly larger than $\lambda\sigma$. Thus one can increase t_0 as β_{\min, A_0} increases in order to reduce false positives while not increasing false negatives from the set A_0 . In Lemma 6.2, we impose a lower bound on β_{\min, A_0} (6.4) in order to recover the subset of variables in A_0 , while achieving the nearly ideal ℓ_2 loss with a sparse model I .

We now show In Lemma 6.1 that under no restriction on β_{\min} , we achieve an oracle bound on the ℓ_2 loss, which depends only on the ℓ_2 loss of the initial estimator on the set T_0 . Bounds in Lemma 4.2 and 5.2 are special cases (6.2) as we state now.

Lemma 6.1. *Let β_{init} be an initial estimator. Let $h = \beta_{\text{init}} - \beta_{T_0}$ and $\lambda := \sqrt{2 \log p/n}$. Suppose that we choose a thresholding parameter t_0 and set*

$$I = \{j : |\beta_j, \text{init}| \geq t_0\}.$$

Then for $\mathcal{D} := \{1, \dots, p\} \setminus I$, we have for $\mathcal{D}_{11} := \mathcal{D} \cap A_0$ and $a_0 = |A_0|$,

$$\|\beta_{\mathcal{D}}\|_2^2 \leq (s_0 - a_0)\lambda^2\sigma^2 + (t_0\sqrt{a_0} + \|h_{\mathcal{D}_{11}}\|_2)^2. \quad (6.2)$$

Suppose that $t_0 < \beta_{\min, A_0}$ as defined in (6.1). Then (6.2) can be replaced by

$$\|\beta_{\mathcal{D}}\|_2^2 \leq (s_0 - a_0)\lambda^2\sigma^2 + \|h_{\mathcal{D}_{11}}\|_2^2 (\beta_{\min, A_0} / (\beta_{\min, A_0} - t_0))^2. \quad (6.3)$$

Lemma 6.2. (Oracle Ideal MSE with ℓ_∞ bounds) Suppose that (1.4) and (1.5) hold. Let β_{init} be an initial estimator. Let $h = \beta_{\text{init}} - \beta_{T_0}$ and $\lambda := \sqrt{2 \log p/n}$. Suppose on some event Q_c , for β_{\min, A_0} as defined in (6.1), it holds that

$$\beta_{\min, A_0} \geq \|h_{A_0}\|_\infty + \min \left\{ (s_0)^{1/2} \|h_{T_0^c}\|_2, (s_0)^{-1} \|h_{T_0^c}\|_1 \right\}. \quad (6.4)$$

Now we choose a thresholding parameter t_0 such that on Q_c , for some $\check{s}_0 \geq s_0$,

$$\beta_{\min, A_0} - \|h_{A_0}\|_\infty \geq t_0 \geq \min \left\{ (\check{s}_0)^{-1/2} \|\beta_{\text{init}, T_0^c}\|_2, (\check{s}_0)^{-1} \|\beta_{\text{init}, T_0^c}\|_1 \right\} \quad (6.5)$$

holds and set $I = \{j : |\beta_{j, \text{init}}| \geq t_0\}$; Then we have on $\mathcal{T}_a \cap Q_c$,

$$A_0 \subset I \text{ and } |I \cap T_0^c| \leq \check{s}_0; \text{ and hence } |I| \leq s_0 + \check{s}_0; \quad (6.6)$$

$$\text{and } \|\beta_{\mathcal{D}}\|_2^2 \leq (s_0 - a_0)\lambda^2\sigma^2. \quad (6.7)$$

For $\widehat{\beta}_I$ being the OLS estimator based on (X_I, Y) and $\check{s}_0 \leq s$, we have on $\mathcal{T}_a \cap Q_c$,

$$\left\| \widehat{\beta}_I - \beta \right\|_2^2 \leq C_7 \check{s}_0 \lambda^2 \sigma^2 / \Lambda_{\min}^2(|I|) \quad (6.8)$$

where C_7 depends on $\theta_{|I|, |S_{\mathcal{D}}|}$ which is upper bounded by $(\Lambda_{\max}(2s) - \Lambda_{\min}(2s))/2$.

By introducing \check{s}_0 , the dependency of t_0 on the knowledge of s_0 is relaxed; in particular, it can be used to express a desirable level of sparsity for the model I that one wishes to select. We note that implicit in the statement of Lemma (6.2), we assume the knowledge of the bounds on various norms of $\beta_{\text{init}} - \beta$ (hence the name of ‘‘oracle’’). Theorem 6.3 is an immediate corollary of Lemma 6.2, with the difference being: we now let $\check{s}_0 = s_0$ everywhere and assume having an upper estimate \check{D}_1 of D_1 , so as not to depend on an ‘‘oracle’’ telling us an exact value.

Theorem 6.3. Suppose that $RE(s_0, 6, X)$ condition holds. Choose $\lambda_n \geq b\lambda_{\sigma, a, p}$, where $b \geq 2$. Let β_{init} be the Lasso estimator as in (1.2). Suppose that for some constants $\check{D}_1 \geq D_1$, and for D_0, D_1 as in (5.2) and (5.3), it holds that

$$\beta_{\min, A_0} \geq D_0 \lambda \sigma \sqrt{s_0} + \check{D}_1 \lambda \sigma, \text{ where } \lambda := \sqrt{2 \log p/n},$$

Choose a thresholding parameter t_0 and set

$$I = \{j : |\beta_{j, \text{init}}| \geq t_0\}, \text{ where } t_0 \geq \check{D}_1 \lambda \sigma.$$

Then on \mathcal{T}_a , (6.6), (6.7), and (6.8) all hold with $\check{s}_0 = s_0$ everywhere and $C_7 \leq \Lambda_{\min}^2(|I|) + (\Lambda_{\max}(2s) - \Lambda_{\min}(2s))^2/2 + 4(1 + a)$; Moreover, the OLS estimator $\widehat{\beta}$ based on I achieves on \mathcal{T}_a , for $f(I)$ as defined in Corollary 4.4, where $|I| \leq 2s_0$,

$$\left\| X \widehat{\beta}_I - X \beta \right\|_2 / \sqrt{n} \leq C_8 \sqrt{s_0} \lambda \sigma \text{ where } C_8 = \sqrt{\Lambda_{\max}(s)} + f(I).$$

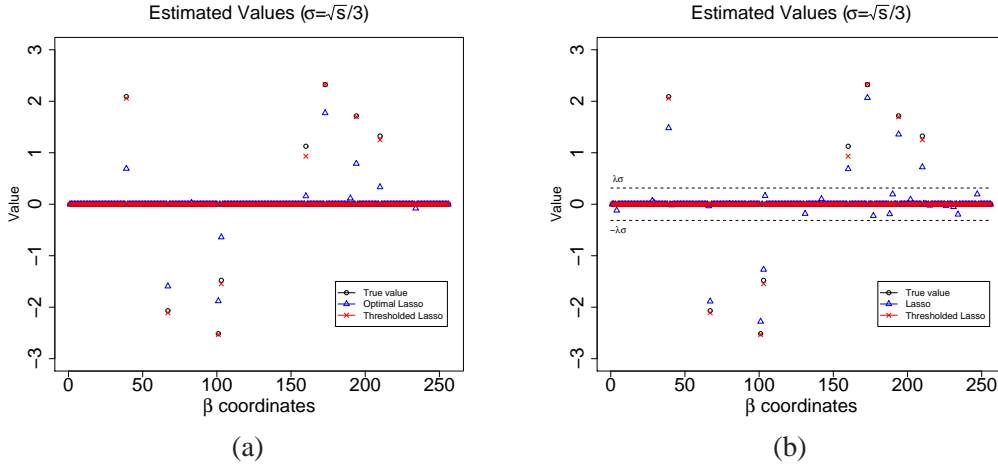


Figure 1: Illustrative example: i.i.d. Gaussian ensemble; $p = 256$, $n = 72$, $s = 8$, and $\sigma = \sqrt{s}/3$. (a) compare with the Lasso estimator $\tilde{\beta}$ which minimizes ℓ_2 loss. Here $\tilde{\beta}$ has only 3 FPs, but ρ^2 is large with a value of 64.73. (b) Compare with the β_{init} obtained using λ_n . The dotted lines show the thresholding level t_0 . The β_{init} has 15 FPs, all of which were cut after the 1st step; resulting $\rho^2 = 12.73$. After refitting with OLS in the 2nd step, for the $\hat{\beta}$, ρ^2 is further reduced to 0.51.

6.1 Discussions

Compared to Theorem 1.1, we now put a lower bound on β_{\min, A_0} rather than on the entire set S in Theorem 6.3, with the hope to recover A_0 . Choosing the set A_0 is rather arbitrary; one could for example, consider the set of variables that are strictly above $\lambda\sigma/2$ for instance. Bounds on $\|h_{A_0}\|_\infty$ are in general harder to obtain than $\|h_{A_0}\|_2$; Under stronger incoherence conditions, such bounds can be obtained; see for example Candès and Plan (2009); Lounici (2008); Wainwright (2009b). In general, we can still hope to bound $\|h_{A_0}\|_\infty$ by $\|h_{A_0}\|_2$. Having a tight bound on $\|h_{T_0}\|_2$ (or $\|h_{T_0}\|_\infty$) and $\|h_{T_0^c}\|_2$ naturally helps relaxing the requirement on β_{\min, A_0} for Lemma 6.2, while in Lemma 6.1, such tight upper bounds will help us to control both the size of I and $\|\beta_{\mathcal{D}}\|$ and therefore achieve a tight bound on the ℓ_2 loss in the expression of Lemma 4.3. In general, when the strong signals are close to each other in their strength, then a small β_{\min, A_0} implies that we are in a situation with low signal to noise ratio (low SNR); one needs to carefully tradeoff false positives with false negatives; this is shown in our experimental results in Section 7. We refer to Wainwright (2009a) and references therein for discussions on information theoretic limits on sparse recovery where the particular estimator is not specified.

7 Numerical experiments

In this section, we present results from numerical simulations designed to validate the theoretical analysis presented in previous sections. In our Thresholded Lasso implementation (we plan to release the imple-

mentation as an R package), we use a *Two-step* procedure as described in Section 1: we use the Lasso as the initial estimator, and OLS in the second step after thresholding. Specifically, we carry out the Lasso using procedure $\text{LARS}(Y, X)$ that implements the LARS algorithm Efron et al. (2004) to calculate the full regularization path. We then use λ_n , whose expression is fixed throughout the experiments as follows,

$$\lambda_n = 0.69\lambda\sigma, \quad \text{where } \lambda = \sqrt{2 \log p/n}, \quad \text{in (1.2)} \quad (7.1)$$

to select a β_{init} from this output path as our initial estimator. We then threshold the β_{init} using a value t_0 typically chosen between $0.5\lambda\sigma$ and $\lambda\sigma$. See each experiment for the actual value used. Given that columns of X being normalized to have ℓ_2 norm \sqrt{n} , for each input parameter β , we compute its SNR as follows:

$$\text{SNR} := \|\beta\|_2^2 / \sigma^2.$$

To evaluate $\widehat{\beta}$, we use metrics defined in Table 1; we also compute the ratio between squared ℓ_2 error and the ideal mean squared error, known as the ρ^2 ; see Section 7.3 for details.

7.1 Illustrative example

In the first example, we run the following experiment with a setup similar to what was used in Candès and Tao (2007) to conceptually compare the behavior of the Thresholded Lasso with the Gauss-Dantzig selector:

1. Generate an *i.i.d. Gaussian ensemble* $X_{n \times p}$, where $X_{ij} \sim N(0, 1)$ are independent, which is then normalized to have column ℓ_2 -norm \sqrt{n} .
2. Select a support set S of size $|S| = s$ uniformly at random, and sample a vector β with independent and identically distributed entries on S as follows, $\beta_i = \mu_i(1 + |g_i|)$, where $\mu_i = \pm 1$ with probability $1/2$ and $g_i \sim N(0, 1)$.
3. Compute $Y = X\beta + \epsilon$, where the noise $\epsilon \sim N(0, \sigma^2 I_n)$ is generated with I_n being the $n \times n$ identity matrix. Then feed Y and X to the Thresholded Lasso with thresholding parameter being t_0 to recover β using $\widehat{\beta}$.

In Figure 1, we set $p = 256$, $n = 72$, $s = 8$, $\sigma = \sqrt{s}/3$ and $t_0 = \lambda\sigma$. We compare the Thresholded Lasso estimator $\widehat{\beta}$ with the Lasso, where the full LARS regularization path is searched to find the *optimal* $\widetilde{\beta}$ that has the minimum ℓ_2 error.

7.2 Type I/II errors

We now evaluate the Thresholded Lasso estimator by comparing Type I/II errors under different values of t_0 and SNR. We consider Gaussian random matrices for the design X with both diagonal and Toeplitz covariance. We refer to the former as *i.i.d. Gaussian ensemble* and the latter as *Toeplitz ensemble*. In the Toeplitz case, the covariance is given by $T(\gamma)_{i,j} = \gamma^{|i-j|}$ where $0 < \gamma < 1$. We run under two noise levels: $\sigma = \sqrt{s}/3$ and $\sigma = \sqrt{s}$. For each σ , we vary the threshold t_0 from $0.01\lambda\sigma$ to $1.5\lambda\sigma$. For each σ and t_0

combination, we run the following experiment: First we generate X as in Step 1 above. After obtaining X , we keep it fixed and then repeat Steps 2 – 3 for 200 times with a new β and ϵ generated each time and we count the number of Type I and II errors in $\hat{\beta}$. We compute the average at the end of 200 runs, which will correspond to one data point on the curves in Figure 2 (a) and (b).

For both types of designs, similar behaviors are observed. For $\sigma = \sqrt{s}/3$, FNs increase slowly; hence there is a wide range of values from which t_0 can be chosen such that FNs and FPs are both zero. In contrast, when $\sigma = \sqrt{s}$, FNs increase rather quickly as t_0 increases due to the low SNR. It is clear that the low SNR and high correlation combination makes it the most challenging situation for variable selection, as predicted by our theoretical analysis and others. See discussions in Section 6. In (c) and (d), we run additional experiments for the low SNR case for Toeplitz ensembles. The performance is improved by increasing the sample size or lowering the correlation factor.

Table 1: Metrics for evaluating $\hat{\beta}$

Metric	Definition
Type I errors or False Positives (FPs)	# of incorrectly selected non-zeros in $\hat{\beta}$
Type II errors or False Negatives (FNs)	# of non-zeros in β that are not selected in $\hat{\beta}$
True positives (TPs)	# of correctly selected non-zeros
True Negatives (TNs)	# of zeros in $\hat{\beta}$ that are also zero in β
False Positive Rate (FPR)	$FPR = FP/(FP + TN) = FP/(p - s)$
True Positive Rate (TPR)	$TPR = TP/(TP + FN) = TP/s$

7.3 ℓ_2 loss

We now compare the performance of the Thresholded Lasso with the ordinary Lasso by examining the metric ρ^2 defined as follows:

$$\rho^2 = \frac{\sum_{i=1}^p (\hat{\beta}_i - \beta_i)^2}{\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)}.$$

We first run the above experiment using i.i.d. Gaussian ensemble under the following thresholds: $t_0 = \lambda\sigma$ for $\sigma = \sqrt{s}/3$, and $t_0 = 0.36\lambda\sigma$ for $\sigma = \sqrt{s}$. These are chosen based on the desire to have low errors of both types (as shown in Figure 2 (a)). Naturally, for low SNR cases, small t_0 will reduce Type II errors. In practice, we suggest using cross-validations to choose the exact constants in front of $\lambda\sigma$. We plot the histograms of ρ^2 in Figure 2 (e) and (f). In (e), the mean and median are 1.45 and 1.01 for the Thresholded Lasso, and 46.97 and 41.12 for the Lasso. In (f), the corresponding values are 7.26 and 6.60 for the Thresholded Lasso and 10.50 and 10.01 for the Lasso. With high SNR, the Thresholded Lasso performs extremely well; with low SNR, the improvement of the Thresholded Lasso over the ordinary Lasso is less prominent; this is in close correspondence with the Gauss-Dantzig selector’s behavior as shown by Candès and Tao (2007).

Next we run the above experiment under different sparsity values of s . We again use i.i.d. Gaussian ensemble

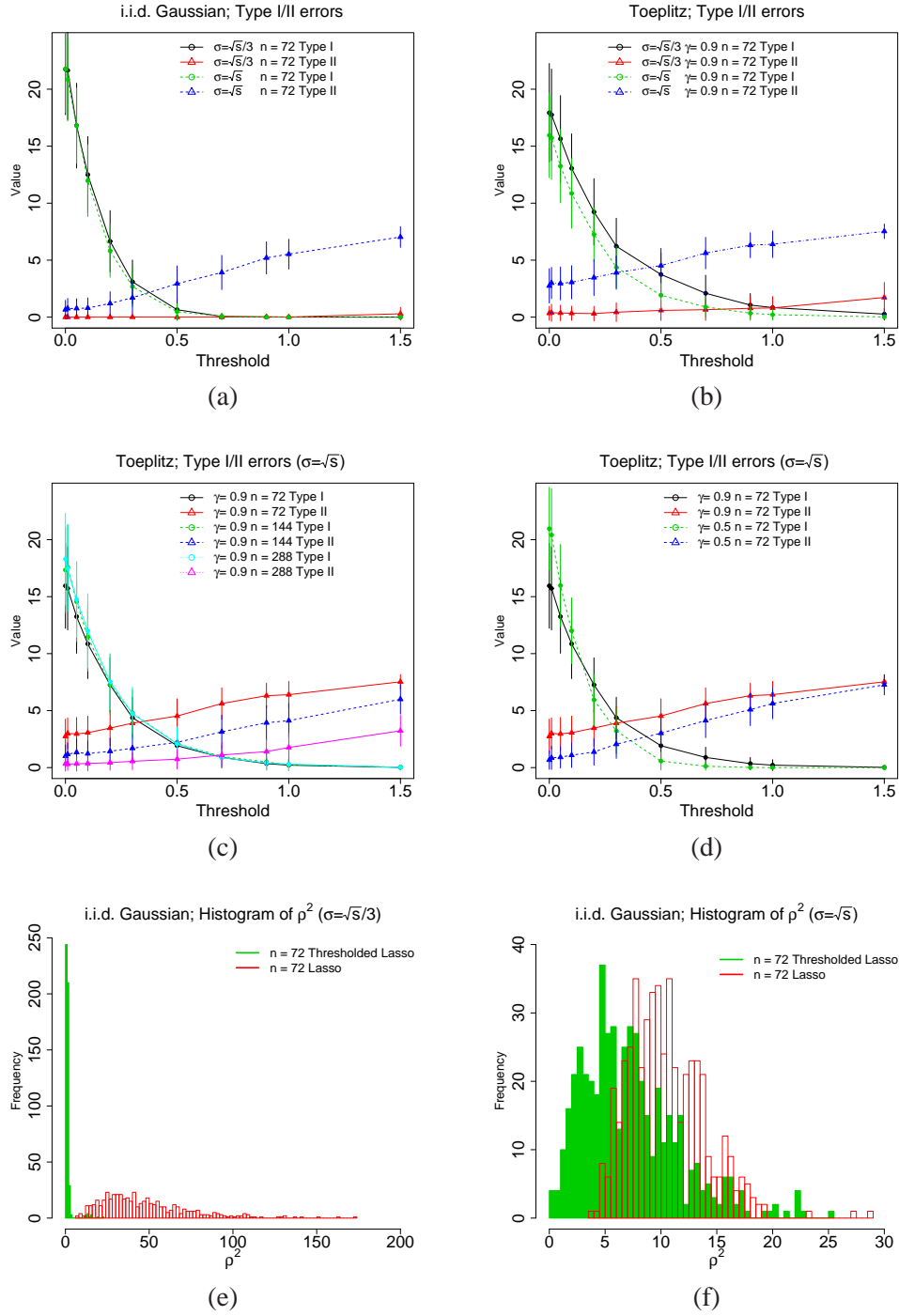


Figure 2: $p = 256$ $s = 8$. (a) (b) Type I/II errors for i.i.d. Gaussian and Toeplitz ensembles. Each vertical bar represents ± 1 std. The unit of x -axis is in $\lambda\sigma$. For both types of design matrices, FPs decrease and FNs increase as the threshold increases. For Toeplitz ensembles, in (c) with fixed correlation γ , FNs decrease with more samples, and in (d) with fixed sample size, FNs decrease as the correlation γ decreases. (e) (f) Histograms of ρ^2 under i.i.d Gaussian ensembles from 500 runs.

with $p = 2000$, $n = 400$, and $\sigma = \sqrt{s}/3$. The threshold is set at $t_0 = \lambda\sigma$. The SNR for different s is fixed at around 32.36. Table 2 shows the mean of the ρ^2 for the Lasso and the Thresholded Lasso estimators. The Thresholded Lasso performs consistently better than the ordinary Lasso until about $s = 80$, after which both break down. For the Lasso, we always choose from the full regularization path the *optimal* $\tilde{\beta}$ that has the minimum ℓ_2 loss.

Table 2: ρ^2 under different sparsity and fixed SNR. Average over 100 runs for each s .

s	5	18	20	40	60	80	100
SNR	34.66	32.99	32.29	32.08	32.28	32.56	32.54
Lasso	17.42	22.01	44.89	52.68	31.88	29.40	47.63
Thresholded Lasso	1.02	0.96	1.11	1.54	10.32	29.38	53.81

7.4 Linear Sparsity

We next present results demonstrating that the Thresholded Lasso recovers a sparse model using a small number of samples per non-zero component in β when X is a subgaussian ensemble. We run under three cases of $p = 256, 512, 1024$; for each p , we increase the sparsity s by roughly equal steps from $s = 0.2p/\log(0.2p)$ to $p/4$. For each p and s , we run with different sample size n . For each tuple (n, p, s) , we run an experiment similar to the one described in Section 7.2 with an i.i.d. Gaussian ensemble X being fixed while repeating Steps 2 – 3 100 times. In Step 2, each randomly selected non-zero coordinate of β is assigned a value of ± 0.9 with probability $1/2$. After each run, we compare $\hat{\beta}$ with the true β ; if all components match in signs, we count this experiment as a success. At the end of the 100 runs, we compute the percentage of successful runs as the probability of success. We compare with the ordinary Lasso, for which we search over the full regularization path of LARS and choose the $\tilde{\beta}$ that best matches β in terms of support.

We experiment with $\sigma = 1$ and $\sigma = \sqrt{s}/3$. For $\sigma = 1$, we set $t_0 = f_t \sqrt{|\hat{S}_0|} \lambda \sigma$, where $\hat{S}_0 = \{j : \beta_{j,\text{init}} \geq 0.5\lambda_n = 0.35\lambda\sigma\}$ for λ_n as in (7.1), and f_t is chosen from the range of $[0.12, 0.24]$ (cf. Section 3). For $\sigma = \sqrt{s}/3$, we set $t_0 = 0.7\lambda\sigma$ with SNR being fixed. The results are shown in Figure 3. We observe that under both noise levels, the Thresholded Lasso estimator requires much fewer samples than the ordinary lasso in order to conduct exact recovery of the sparsity pattern of the true linear model when all non-zero components are sufficiently large. When σ is fixed as s increases, the SNR is increasing; the experimental results illustrate the behavior of sparse recovery when it is close to the noiseless setting. Given the same sparsity, more samples are required for the low SNR case to reach the same level of success rate. Similar behavior was also observed for Toeplitz and Bernoulli ensembles with i.i.d. ± 1 entries.

7.5 ROC comparison

We now compare the performance of the Thresholded Lasso estimator with the Lasso and the Adaptive Lasso by examining their ROC curves. Our parameters are $p = 512$, $n = 330$, $s = 64$ and we run under

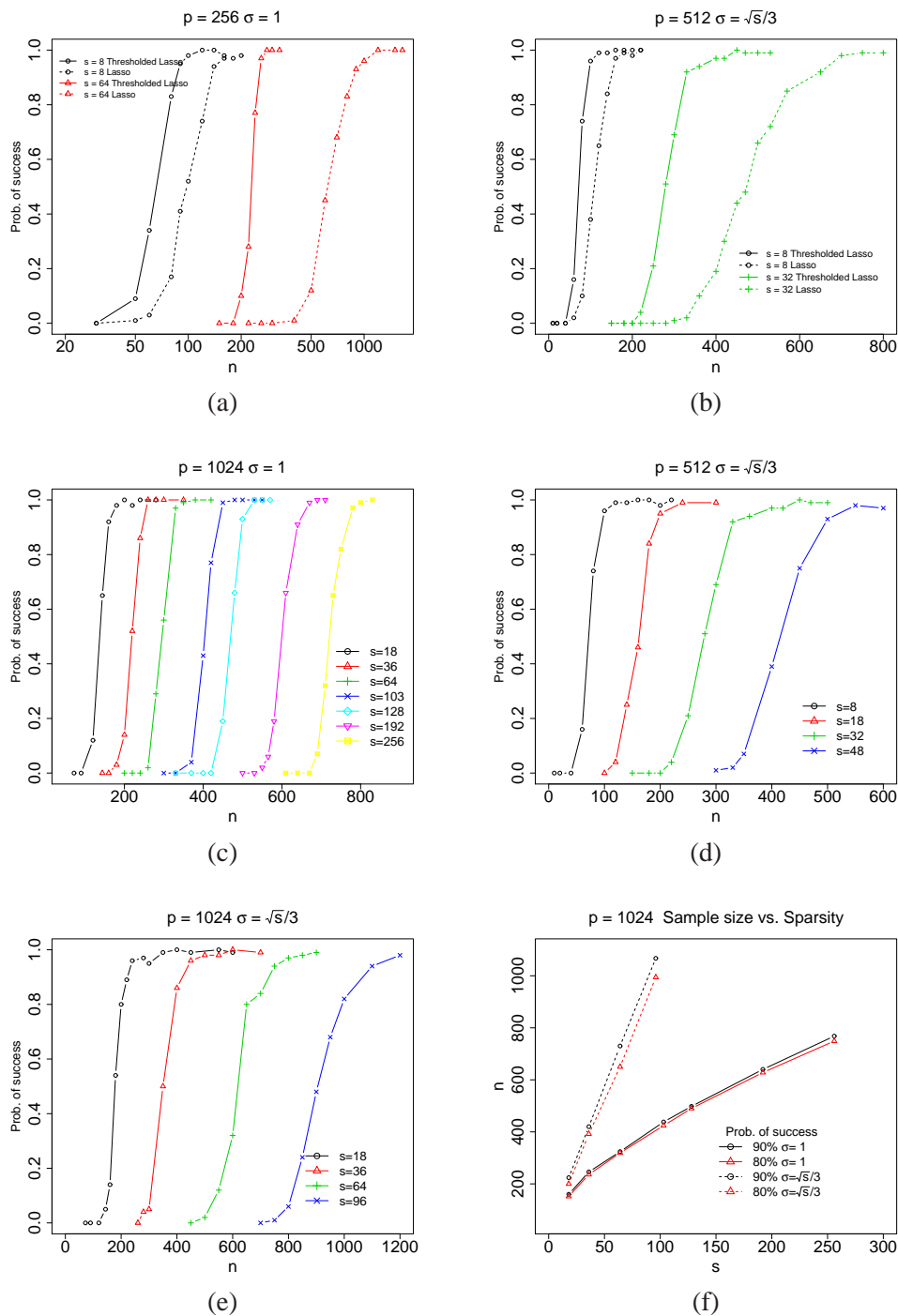


Figure 3: (a) (b) Compare the probability of success for $p = 256$ and $p = 512$ under two noise levels. The Thresholded Lasso estimator requires much fewer samples than the ordinary Lasso. (c) (d) (e) show the probability of success of the Thresholded Lasso under different levels of sparsity and noise levels when n increases for $p = 512$ and 1024. (f) The number of samples n increases almost linearly with s for $p = 1024$. More samples are required to achieve the same level of success when $\sigma = \sqrt{s}/3$ due to the relatively low SNR.

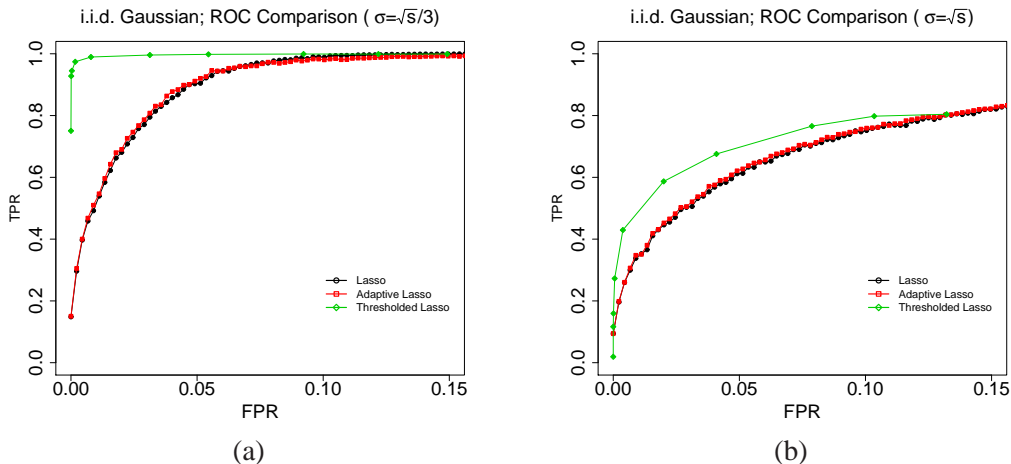


Figure 4: $p = 512$ $n = 330$ $s = 64$. ROC for the Thresholded Lasso, ordinary Lasso and Adaptive Lasso. The Thresholded Lasso clearly outperforms the ordinary Lasso and the Adaptive Lasso for both high and low SNRs.

two cases: $\sigma = \sqrt{s}/3$ and $\sigma = \sqrt{s}$. In the Thresholded Lasso, we vary the threshold level from $0.01\lambda\sigma$ to $1.5\lambda\sigma$. For each threshold, we run the experiment described in Section 7.2 with an i.i.d. Gaussian ensemble X being fixed while repeating Steps 2 – 3 100 times. After each run, we compute the FPR and TPR of the $\hat{\beta}$, and compute their averages after 100 runs as the FPR and TPR for this threshold. For the Lasso, we compute the FPR and TPR for each output vector along its entire regularization path. For the Adaptive Lasso, we use the *optimal* output $\tilde{\beta}$ in terms of ℓ_2 loss from the initial Lasso penalization path as the input to its second step, that is, we set $\beta_{\text{init}} := \tilde{\beta}$ and use $w_j = 1/\beta_{\text{init},j}$ to compute the weights for penalizing those non-zero components in β_{init} in the second step, while all zero components of β_{init} are now removed. We then compute the FPR and TPR for each vector that we obtain from the second step's LARS output. We implement the algorithms as given in Zou (2006), the details of which are omitted here as its implementation has become standard. The ROC curves are plotted in Figure 4. The Thresholded Lasso performs better than both the ordinary Lasso and the Adaptive Lasso; its advantage is more apparent when the SNR is high.

8 Conclusion

In this paper, we show that the thresholding method is effective in variable selection and accurate in statistical estimation. It improves the ordinary Lasso in significant ways. For example, we allow very significant number of non-zero elements in the true parameter, for which the ordinary Lasso would have failed. On the theoretical side, we show that if X obeys the RE condition and if the true parameter is sufficiently sparse, the Thresholded Lasso achieves the ℓ_2 loss within a logarithmic factor of the *ideal mean square error* one would achieve with an oracle, while selecting a sufficiently sparse model I . This is accomplished when threshold level is at about $\sqrt{2 \log p/n}\sigma$, assuming that columns of X have ℓ_2 norm \sqrt{n} . We also report a similar result on the Gauss-Dantzig selector under the UUP, built upon results from Candès and Tao (2007).

When the SNR is high, almost exact recovery of the non-zeros in β is possible as shown in our theory; exact recovery of the support of β is shown in our simulation study when n is only linear in s for several Gaussian and Bernoulli random ensembles. When the SNR is relatively low, the inference task is difficult for any estimator. In this case, we show that Thresholded Lasso tradeoffs Type I and II errors nicely: we recommend choosing the thresholding parameter conservatively. Algorithmic issues such as how to get an estimate on σ and parameters related to the incoherence conditions is left as future work. While the current focus is on ℓ_2 loss, we are also interested in exploring the *sparsity oracle inequalities* for the Thresholded Lasso under the RE condition as studied in [Bickel et al. \(2009\)](#) in our future work.

A Proof of Theorem 1.1

Proving Theorem 1.1 involves showing that the Lasso and the Dantzig selector satisfy (3.2). These have been proved in [Bickel et al. \(2009\)](#). Theorem 1.1 is then an immediate corollary of Theorem 3.1 under assumptions therein. We note that on \mathcal{T}_a , it holds that $\|v_{\text{init}, S^c}\|_1 \leq k_0 \|v_{\text{init}, S}\|_1$, where $k_0 = 1$ for the Dantzig selector when $\lambda_n \geq \lambda_{\sigma, a, p}$ and $k_0 = 3$ for the Lasso, when $\lambda_n \geq 2\lambda_{\sigma, a, p}$ for the Lasso. Then on \mathcal{T}_a as in (1.15), (3.2) holds with $B_0 = 4K^2(s, 3)$ and $B_1 = 3K^2(s, 3)$ for Lasso under $RE(s, 3, X)$ and (3.2) holds with $B_0 = B_1 = 4K^2(s, 1)$ for the Dantzig selector under $RE(s, 1, X)$; See [Zhou \(2009a\)](#) for deriving the exact constants here. \square

B Proof of Theorem 1.4

Proof of Theorem 1.4. It is clear by construction that under \mathcal{T}_a , $X\widehat{\beta}_I = P_I Y$ and $|I| \leq 2s_0$. Hence

$$\begin{aligned} \left\| X\widehat{\beta}_I - X\beta \right\|_2 / \sqrt{n} &= \left\| (P_I - \text{Id})X\beta + P_I\epsilon \right\|_2 / \sqrt{n} \\ &\leq \|X_{I^c}\beta_{\mathcal{D}}\|_2 / \sqrt{n} + \|P_I\epsilon\|_2 / \sqrt{n} \\ &\leq \sqrt{\Lambda_{\max}(s)} \|\beta_{\mathcal{D}}\|_2 + \frac{\sqrt{|I|(1+a)\Lambda_{\max}(|I|)}\lambda\sigma}{\Lambda_{\min}(|I|)} \end{aligned}$$

where we have on \mathcal{T}_a , for $\lambda_{\sigma, a, p} = \sqrt{1+a}\lambda\sigma$, where $\lambda = \sqrt{2\log p/n}$,

$$\begin{aligned} \left\| X_I(X_I^T X_I)^{-1} X_I^T \epsilon \right\|_2 / \sqrt{n} &\leq \left\| X_I(X_I^T X_I/n)^{-1} / \sqrt{n} \right\|_2 \left\| X_I^T \epsilon / n \right\|_2 \\ &\leq \frac{\sqrt{\Lambda_{\max}(|I|)}\sqrt{|I|}\lambda_{\sigma, a, p}}{\Lambda_{\min}(|I|)} \leq \frac{\sqrt{|I|(1+a)\Lambda_{\max}(|I|)}\lambda\sigma}{\Lambda_{\min}(|I|)} \end{aligned}$$

Now by Lemma 4.2 and 5.2, we have $\|\beta_{\mathcal{D}}\|_2 \leq C\sqrt{s_0}\lambda\sigma$ for some constant C . \square

C Proof of Proposition 1.5

Recall that $|\beta_j| \leq \lambda\sigma$ for all $j > a_0$ as defined in (6.1); hence for $\lambda = \sqrt{2\log p/n}$, we have by (G.1), $\sum_{i>a_0}^p \min(\beta_i^2, \lambda^2\sigma^2) = \sum_{i>a_0}^s \beta_i^2 \leq (s_0 - a_0)\lambda^2\sigma^2$; hence

$$\left| \{j \in A_0^c : |\beta_j| \geq \sqrt{\log p/(c'n)}\sigma \} \right| \leq 2c'(s_0 - a_0) \text{ where } |T_0 \setminus A_0| = s_0 - a_0.$$

Now given that $\beta_i \geq \beta_j$ for all $i \in T_0, j \in T_0^c$, the proposition holds. \square

D Proof of Theorem 3.1

We first state two lemmas. Define $v_{\text{init}} = \beta_{\text{init}} - \beta$ and $v^{(i)} = \hat{\beta}^{(i)} - \beta$.

Lemma D.1. *Under assumptions in Theorem 3.1, suppose on $\mathcal{T}_a \cap Q_b$,*

$$\beta_{\min} \geq \Xi + \Gamma \text{ where } \Xi := \max_{i=0,1} \left\| v_S^{(i)} \right\|_{\infty} \text{ and } \Gamma := \max_{i=0,1} t_i. \quad (\text{D.1})$$

Then $S \subseteq \hat{S}_2 \subseteq \hat{S}_1$.

Proof. We have $\forall j \in S \beta_{\text{init},j} \geq \beta_{\min} - \|v_{\text{init},S}\|_{\infty} \geq \beta_{\min} - \Xi \geq \Gamma = t_0$ and

$\hat{\beta}_j^{(1)} \geq \beta_{\min} - \|v_S^{(1)}\|_{\infty} \geq \beta_{\min} - \Xi \geq \Gamma \geq t_1$. Thus the lemma holds by definition of \hat{S}_i , for $i = 0, 1, 2$. \square

The following lemma follows from Lemma 4.3, by plugging in $\|\beta_{\mathcal{D}}\|_2 = 0$.

Lemma D.2. (ℓ_2 -loss for the OLS estimators) *Suppose that $I \supseteq S$ and $|I| \leq 2s$, then the OLS estimator $\hat{\beta}_I := (X_I^T X_I)^{-1} X_I^T Y$ satisfies on \mathcal{T}_a , $\|\hat{\beta}_I - \beta\|_2 \leq \lambda_{\sigma,a,p} \sqrt{|I|}/\Lambda_{\min}(|I|)$ which satisfies (3.4) with $B_2 = 1/(B\Lambda_{\min}(2s))$.*

Proof of Theorem 3.1. It is clear by construction that

$$\hat{S}_2 \subseteq \hat{S}_1 \subseteq \hat{S}_0. \quad (\text{D.2})$$

Recall that \hat{S}_0 is obtained by thresholding β_{init} with $4\lambda_n$, hence by (3.2), we have

$$|\hat{S}_0 \setminus S| \leq \frac{\|v_{\text{init},S^c}\|_1}{4\lambda_n} \leq \frac{B_1\lambda_n s}{4\lambda_n} \leq \frac{B_1 s}{4}.$$

1. If $B_1 \leq 4$, we have that $|\hat{S}_0| \leq 2s$;
2. Otherwise, we have $|\hat{S}_0| \leq s + B_1 s/4 \leq B_1 s/2$.

Hence for $t_i = 4\lambda_n \sqrt{|\hat{S}_i|}$, $\forall i = 0, 1$ and Γ as in (D.1), it holds by (D.2) that

$$\Gamma = t_0 = 4\lambda_n \sqrt{|\hat{S}_0|} \leq \lambda_n \sqrt{s} \max\left(2\sqrt{2B_1}, 4\sqrt{2}\right). \quad (\text{D.3})$$

Now given (3.3) and (3.2), we have $\forall j \in S$,

$$\beta_{\text{init},j} \geq \beta_{\min} - \|v_{\text{init},S}\|_{\infty} \geq \beta_{\min} - \|v_{\text{init},S}\|_2 \geq \Gamma = t_0,$$

and hence it holds that $S \subseteq \widehat{S}_1 \subseteq \widehat{S}_0$ by construction of \widehat{S}_1 , and hence $t_0 \geq 4\lambda_n\sqrt{s}$. Now by (3.2), we have for $s \geq B_1^2/16$,

$$|\widehat{S}_1 \setminus S| < \frac{\|v_{\text{init},S^c}\|_1}{t_0} \leq \frac{B_1\lambda_n s}{4\lambda_n\sqrt{s}} < \frac{B_1\sqrt{s}}{4} < s; \quad \text{and} \quad |\widehat{S}_1| < 2s. \quad (\text{D.4})$$

For the OLS estimator $\widehat{\beta}^{(1)}$ with $I = \widehat{S}_1$, by Lemma D.2, we have on \mathcal{T}_a

$$\left\| \widehat{\beta}^{(1)} - \beta \right\|_2 \leq \frac{\lambda_{\sigma,a,p}\sqrt{s_1}}{\Lambda_{\min}(s_1)} \leq \frac{\lambda_n\sqrt{s_1}}{B\Lambda_{\min}(2s)} \leq B_2\lambda_n\sqrt{2s}, \quad \text{where} \quad s_1 := |\widehat{S}_1|$$

where $\lambda_n \geq B\lambda_{\sigma,a,p}$, for $\lambda_{\sigma,a,p}$ as in (1.15), and $B_2 = 1/(B\Lambda_{\min}(2s))$. Clearly we have by definition of Ξ in (D.1),

$$\Xi \leq \max_{i=0,1} \left\| v_S^{(i)} \right\|_2 \leq \max\{B_0, \sqrt{2}B_2\}\lambda_n\sqrt{s}$$

and thus $\beta_{\min} \geq \Xi + \Gamma$ holds given (3.3) and (D.3). By Lemma D.1, we have $\widehat{S}_i \supseteq S, \forall i = 0, 1, 2$. It remains to show (3.5) and (3.4); Upon thresholding $\widehat{\beta}^{(1)}$ with t_1 , we have for $s_1 := |\widehat{S}_1|$ and $\lambda_n \geq B\lambda_{\sigma,a,p}$,

$$|\widehat{S}_2 \setminus S| \leq \left\| v^{(1)} \right\|_2^2 / t_1^2 \leq \left(\frac{\lambda_{\sigma,a,p}\sqrt{s_1}}{\Lambda_{\min}(s_1)} \cdot \frac{1}{4\lambda_n\sqrt{s_1}} \right)^2 \leq \frac{1}{16B^2\Lambda_{\min}^2(s_1)}.$$

Now for the final estimator in (3.1), we have on $\mathcal{T}_a \cap Q_b$ by Lemma D.2,

$$\left\| \widehat{\beta}^{(2)} - \beta \right\|_2 = \left\| \widehat{\beta} - \beta \right\|_2 = \lambda_{\sigma,a,p} \sqrt{|\widehat{S}_2|} / \Lambda_{\min}(|\widehat{S}_2|) \leq \lambda_n B_2 \sqrt{2s}. \quad \square$$

E Proofs for the Gauss-Dantzig selector

Recall β_{init} is the solution to the Dantzig selector. We write $\beta = \beta^{(1)} + \beta^{(2)}$ where

$$\beta_j^{(1)} = \beta_j \cdot 1_{1 \leq j \leq s_0} \quad \text{and} \quad \beta_j^{(2)} = \beta_j \cdot 1_{j > s_0}.$$

Let $h = \beta_{\text{init}} - \beta^{(1)}$, where $\beta^{(1)}$ is hard-thresholded version of β , localized to $T_0 = \{1, \dots, s_0\}$. Let T_1 be the s_0 largest positions of h outside of T_0 ; Let $T_{01} = T_0 \cup T_1$. The proof of Proposition 4.1 (cf. Candès and Tao (2007)) yields the following:

$$\|h_{T_{01}}\|_2 \leq C'_0 \lambda_{p,\tau} \sigma \sqrt{s_0}, \quad \text{for } C'_0 \text{ as in (4.1)} \quad (\text{E.1})$$

$$\|h_{T_0^c}\|_1 \leq C_1 \lambda_{p,\tau} \sigma s_0, \quad \text{where } C_1 = \left(C'_0 + \frac{1 + \delta}{1 - \delta - \theta} \right), \quad \text{and} \quad (\text{E.2})$$

$$\|h_{T_{01}^c}\|_2 \leq \|h_{T_0^c}\|_1 / \sqrt{s_0} \leq C_1 \lambda_{p,\tau} \sigma \sqrt{s_0}, \quad (\text{cf. Lemma F.2}). \quad (\text{E.3})$$

Proof of Lemma 4.2. Consider the set $I \cap T_0^c := \{j \in T_0^c : |\beta_{j,\text{init}}| > t_0\}$. It is clear by definition of $h = \beta_{\text{init}} - \beta^{(1)}$ and (E.2) that

$$|I \cap T_0^c| \leq \|\beta_{T_0^c, \text{init}}\|_1 / t_0 = \|h_{T_0^c}\|_1 / t_0 < s_0, \quad (\text{E.4})$$

where $t_0 \geq C_1 \lambda_{p,\tau} \sigma$. Thus $|I| = |I \cap T_0| + |I \cap T_0^c| \leq 2s_0$; Now (1.16) holds given (E.4) and $|I \cup S| = |S| + |I \cap S^c| \leq s + |I \cap T_0^c| < s + s_0$. We now bound $\|\beta_{\mathcal{D}}\|_2^2$. By (E.1) and (6.2), where $\mathcal{D}_{11} \subset T_0$, we have for $t_0 < C_4 \lambda_{p,\tau} \sigma \sqrt{s_0}$,

$$\|\beta_{\mathcal{D}}\|_2^2 \leq (s_0 - a_0) \lambda^2 \sigma^2 + (t_0 \sqrt{s_0} + \|h_{T_0}\|_2)^2 \leq ((C_4 + C'_0)^2 + 1) \lambda_{p,\tau}^2 \sigma^2. \quad \square$$

Proof of Lemma 4.3. Note that $X_{I^c} \beta_{I^c} = X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}$. We have

$$\begin{aligned} \widehat{\beta}_I &= (X_I^T X_I)^{-1} X_I^T Y = (X_I^T X_I)^{-1} X_I^T (X_I \beta_I + X_{I^c} \beta_{I^c} + \epsilon) \\ &= \beta_I + (X_I^T X_I)^{-1} X_I^T X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}} + (X_I^T X_I)^{-1} X_I^T \epsilon; \\ \text{Hence } \|\widehat{\beta}_I - \beta_I\|_2 &= \|(X_I^T X_I)^{-1} X_I^T X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}} + (X_I^T X_I)^{-1} X_I^T \epsilon\|_2 \\ &\leq \|(X_I^T X_I)^{-1} X_I^T X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}\|_2 + \|(X_I^T X_I)^{-1} X_I^T \epsilon\|_2, \end{aligned} \quad (\text{E.5})$$

where the second term is bounded as Lemma D.2: we have on \mathcal{T}_a ,

$$\|(X_I^T X_I)^{-1} X_I^T \epsilon\|_2 \leq \left\| \left(\frac{X_I^T X_I}{n} \right)^{-1} \right\|_2 \left\| \frac{X_I^T \epsilon}{n} \right\|_2 \leq \frac{\sqrt{|I|}}{\Lambda_{\min}(|I|)} \lambda_{\sigma,a,p} \quad (\text{E.6})$$

by (1.8), where $\lambda_{\sigma,a,p} = \sqrt{1+a} \lambda \sigma$ for $\lambda = \sqrt{\log p/n}$. We now focus on bounding the first term in (E.5). Let P_I denote the orthogonal projection onto I . Let

$$c = (X_I^T X_I)^{-1} X_I^T X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}, \text{ hence } X_{I^c} = P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}.$$

By the disjointness of I and $S_{\mathcal{D}}$, we have for $P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}} := X_{I^c}$,

$$\begin{aligned} \|P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}\|_2^2 &= \langle P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}, X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}} \rangle = \langle X_{I^c}, X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}} \rangle \\ &\leq n \theta_{|I|,|S_{\mathcal{D}}|} \|c\|_2 \|\beta_{S_{\mathcal{D}}}\|_2 \text{ where} \\ \|c\|_2 &\leq \frac{\|X_{I^c}\|_2}{\sqrt{n \Lambda_{\min}(|I|)}} \leq \frac{\|P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}\|_2}{\sqrt{n \Lambda_{\min}(|I|)}}; \text{ Hence} \end{aligned} \quad (\text{E.7})$$

$$\|P_I X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}\|_2 \leq \frac{\sqrt{n} \theta_{|I|,|S_{\mathcal{D}}|}}{\sqrt{\Lambda_{\min}(|I|)}} \|\beta_{S_{\mathcal{D}}}\|_2 \text{ where } \|\beta_{S_{\mathcal{D}}}\|_2 = \|\beta_{\mathcal{D}}\|_2 \quad (\text{E.8})$$

and $\|c\|_2 \leq \theta_{|I|,|S_{\mathcal{D}}|} \|\beta_{\mathcal{D}}\|_2 / \Lambda_{\min}(|I|)$. Now we have on \mathcal{T}_a , by (E.6),

$$\begin{aligned} \|\widehat{\beta}_I - \beta_I\|_2 &\leq \|(X_I^T X_I)^{-1} X_I^T X_{S_{\mathcal{D}}} \beta_{S_{\mathcal{D}}}\|_2 + \|(X_I^T X_I)^{-1} X_I^T \epsilon\|_2 \\ &\leq \frac{\theta_{|I|,|S_{\mathcal{D}}|}}{\Lambda_{\min}(|I|)} \|\beta_{\mathcal{D}}\|_2 + \frac{\sqrt{|I|}}{\Lambda_{\min}(|I|)} \lambda_{\sigma,a,p}. \end{aligned}$$

Now the lemma holds given $\|\widehat{\beta}_I - \beta\|_2^2 = \|\widehat{\beta}_I - \beta_I\|_2^2 + \|\beta_I - \beta\|_2^2$. \square

Proof of Theorem 1.2. It holds by definition of $S_{\mathcal{D}}$ that $I \cap S_{\mathcal{D}} = \emptyset$. It is clear by Lemma 4.2 that $|S_{\mathcal{D}}| < s$ and $|I| \leq 2s_0$ and $|I \cup S_{\mathcal{D}}| \leq |I \cup S| \leq s + s_0 \leq 2s$; Thus for $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$, we have for $\lambda = \sqrt{2 \log p/n}$, and by (4.5)

$$\begin{aligned} \|\widehat{\beta}_I - \beta\|_2^2 &\leq \|\beta_{\mathcal{D}}\|_2^2 \left(1 + \frac{2\theta_{|I|, |S_{\mathcal{D}}|}^2}{\Lambda_{\min}^2(|I|)}\right) + \frac{2|I|}{\Lambda_{\min}^2(|I|)} \lambda_{\sigma, a, p}^2 \\ &\leq \lambda^2 \sigma^2 s_0 \left((\sqrt{1+a} + \tau^{-1})^2 ((C'_0 + C_4)^2 + 1) \left(1 + \frac{2\theta_{s, 2s_0}^2}{\Lambda_{\min}^2(2s_0)}\right) + \frac{4(1+a)}{\Lambda_{\min}^2(2s_0)} \right). \end{aligned}$$

Thus the theorem holds for C_3 as in (4.3) by (4.5), where it holds for $\tau > 0$ that

$$\frac{\theta_{s, 2s_0}}{\Lambda_{\min}(2s_0)} \leq \frac{\theta_{s, 2s}}{\Lambda_{\min}(2s_0)} \leq \frac{1 - \delta_{2s} - \tau}{\Lambda_{\min}(2s)} < 1$$

given that $\theta_{s, 2s} < 1 - \tau - \delta_{2s} < \Lambda_{\min}(2s)$ for $\tau > 0$. \square

F Oracle properties of the Lasso

We first show Lemma F.1, which gives us the prediction error using β_{T_0} .

Lemma F.1. *Suppose that (1.5) holds. We have for $\lambda = \sqrt{(2 \log p)/n}$.*

$$\|X\beta - X\beta_{T_0}\|_2 / \sqrt{n} \leq \sqrt{\Lambda_{\max}(s - s_0)} \lambda \sigma \sqrt{s_0}. \quad (\text{F.1})$$

Proof. The lemma holds given that $\|\beta_{T_0^c}\|_2 \leq \lambda \sigma \sqrt{s_0}$, and $\|X\beta - X\beta_{T_0}\|_2 / \sqrt{n} = \|X\beta_{T_0^c}\|_2 / \sqrt{n} \leq \sqrt{\Lambda_{\max}(s - s_0)} \|\beta_{T_0^c}\|_2$. \square

We then state Lemma F.2, followed by the proof of Theorem 5.1, where we do not focus on obtaining the best constants. Lemma F.2 is the same (up to normalization) as Lemma 3.1 in Candès and Tao (2007). We note that in their original statement, the UUP condition is assumed; a careful examination of their proof shows that it is a sufficient but not necessary condition; indeed we only need to assume that $\Lambda_{\min}(2s_0) > 0$ and $\theta_{s_0, 2s_0} < \infty$, as we show below. The proof is included by the end of this section for the purpose of a self-complete presentation.

Lemma F.2. *Suppose $\Lambda_{\min}(2s_0) > 0$ and $\theta_{s_0, 2s_0} < \infty$. Then*

$$\begin{aligned} \|h_{T_{01}}\|_2 &\leq \frac{1}{\sqrt{\Lambda_{\min}(2s_0)} \sqrt{n}} \|Xh\|_2 + \frac{\theta_{s_0, 2s_0}}{\sqrt{s_0} \Lambda_{\min}(2s_0)} \|h_{T_0^c}\|_1 \\ \|h_{T_{01}^c}\|_2^2 &\leq \|h_{T_0^c}\|_1^2 \sum_{k \geq s_0+1} 1/k^2 \leq \|h_{T_0^c}\|_1^2 / s_0 \quad \text{and thus} \\ \|h\|_2 &\leq \|h_{T_{01}}\|_2^2 + s_0^{-1} \|h_{T_0^c}\|_1^2 \end{aligned}$$

Proof of Theorem 5.1. Throughout this proof, we assume that \mathcal{T}_a holds. We use $\widehat{\beta} := \beta_{\text{init}}$ to represent the solution to the Lasso estimator in (1.2); By the optimality of $\widehat{\beta}$, we have

$$\begin{aligned} \frac{1}{2n} \left\| Y - X\widehat{\beta} \right\|_2^2 - \frac{1}{2n} \left\| Y - X\beta_{T_0} \right\|_2^2 &\leq \lambda_n \|\beta_{T_0}\|_1 - \lambda_n \left\| \widehat{\beta} \right\|_1, \quad \text{where} \\ \left\| Y - X\widehat{\beta} \right\|_2^2 &= \left\| X\beta - X\widehat{\beta} + \epsilon \right\|_2^2 = \left\| X\widehat{\beta} - X\beta \right\|_2^2 + 2(\beta - \widehat{\beta})^T X^T \epsilon + \|\epsilon\|_2^2 \end{aligned} \quad (\text{F.2})$$

and similarly, we have for $\beta_0 = \beta_{T_0}$,

$$\left\| Y - X\beta_0 \right\|_2^2 = \left\| X\beta - X\beta_0 + \epsilon \right\|_2^2 = \left\| X\beta - X\beta_0 \right\|_2^2 + 2(\beta - \beta_0)^T X^T \epsilon + \|\epsilon\|_2^2;$$

Let $h = \widehat{\beta} - \beta_0$. Thus by (F.2) and the triangle inequality, we have on \mathcal{T}_a

$$\begin{aligned} \frac{\left\| X\widehat{\beta} - X\beta \right\|_2^2}{n} &\leq \frac{\left\| X\beta - X\beta_0 \right\|_2^2}{n} + \frac{2h^T X^T \epsilon}{n} + 2\lambda_n (\|\beta_0\|_1 - \|h + \beta_0\|_1) \\ &\leq \frac{\left\| X\beta - X\beta_0 \right\|_2^2}{n} + 2\|h\|_1 \left\| \frac{X^T \epsilon}{n} \right\|_\infty + 2\lambda_n (\|h_{T_0}\|_1 - \|h_{T_0^c}\|_1) \\ &\leq \frac{\left\| X\beta - X\beta_0 \right\|_2^2}{n} + 3\lambda_n \|h_{T_0}\|_1 - \lambda_n \|h_{T_0^c}\|_1, \end{aligned}$$

where we have used the fact that $\lambda_n \geq 2\lambda_{\sigma,a,p}$ for $a \geq 0$; Thus we have on \mathcal{T}_a ,

$$\left\| X\widehat{\beta} - X\beta \right\|_2^2 / n + \lambda_n \|h_{T_0^c}\|_1 \leq \left\| X\beta - X\beta_0 \right\|_2^2 / n + 3\lambda_n \|h_{T_0}\|_1, \quad (\text{F.3})$$

which is also the starting point of our analysis on the oracle inequalities of the Lasso estimator. Now we differentiate between two cases.

1. Suppose that on \mathcal{T}_a , $\left\| X\beta - X\beta_0 \right\|_2^2 / n \geq 3\lambda_n \|h_{T_0}\|_1$. We then have that

$$\left\| X\widehat{\beta} - X\beta \right\|_2^2 / n + \lambda_n \|h_{T_0^c}\|_1 \leq 2 \left\| X\beta - X\beta_0 \right\|_2^2 / n \quad (\text{F.4})$$

and hence for $\lambda_n = d_0 \lambda \sigma$, where $d_0 \geq 2$, we have by Lemma F.1,

$$\|h_{T_0^c}\|_1 \leq 2\Lambda_{\max}(s - s_0) \lambda \sigma s_0 / d_0 \leq \Lambda_{\max}(s - s_0) \lambda \sigma s_0.$$

Now by (F.3), we have

$$\begin{aligned} \|h\|_1 &\leq \left\| X\beta - X\beta_0 \right\|_2^2 / (n\lambda_n) + 4\|h_{T_0}\|_1 \\ &\leq 7 \left\| X\beta - X\beta_0 \right\|_2^2 / (3n\lambda_n) \leq 7\Lambda_{\max}(s - s_0) \lambda \sigma s_0 / (3d_0) \text{ and clearly} \\ \|Xh\|_2 &\leq \left\| X\widehat{\beta} - X\beta \right\|_2 + \left\| X\beta - X\beta_0 \right\|_2 \leq (\sqrt{2} + 1) \left\| X\beta - X\beta_0 \right\|_2. \end{aligned}$$

By Lemma F.2, we have on \mathcal{T}_a ,

$$\begin{aligned} \|h_{T_0}\|_2 &\leq \frac{1}{\sqrt{n} \sqrt{\Lambda_{\min}(2s_0)}} \|Xh\|_2 + \frac{\theta_{s_0, 2s_0}}{\Lambda_{\min}(2s_0) \sqrt{s_0}} \|h_{T_0^c}\|_1 \\ &\leq \lambda \sigma \sqrt{s_0} \frac{\sqrt{\Lambda_{\max}(s - s_0)}}{\sqrt{\Lambda_{\min}(2s_0)}} \left((\sqrt{2} + 1) + \frac{\theta_{s_0, 2s_0} \sqrt{\Lambda_{\max}(s - s_0)}}{\sqrt{\Lambda_{\min}(2s_0)}} \right) \\ &= D \lambda \sigma \sqrt{s_0}, \text{ for } D = (\sqrt{2} + 1) \frac{\sqrt{\Lambda_{\max}(s - s_0)}}{\sqrt{\Lambda_{\min}(2s_0)}} + \frac{\theta_{s_0, 2s_0} \Lambda_{\max}(s - s_0)}{\Lambda_{\min}(2s_0)}. \end{aligned}$$

2. Otherwise, suppose on \mathcal{T}_a , we have $\|X\beta - X\beta_0\|_2^2/n \leq 3\lambda_n \|h_{T_0}\|_1$; thus

$$\left\|X\widehat{\beta} - X\beta\right\|_2^2/(n\lambda_n) + \|h_{T_0^c}\|_1 \leq 6 \|h_{T_0}\|_1$$

and $\|h_{T_0^c}\|_1 \leq 6 \|h_{T_0}\|_1$, which under the $RE(s_0, 6, X)$ condition immediately implies that

$$\|h_{T_0}\|_2 \leq K(s_0, 6) \|Xh\|_2/\sqrt{n}. \quad (\text{F.5})$$

The rest of the proof is devoted to this second case.

We use $K := K(s_0, 6)$ as a shorthand below. By (F.3), we have on \mathcal{T}_a ,

$$\begin{aligned} & \left\|X\widehat{\beta} - X\beta\right\|_2^2/n + \lambda_n \|h_{T_0^c}\|_1 - \|X\beta - X\beta_0\|_2^2/n \\ & \leq 3\lambda_n \|h_{T_0}\|_1 \leq 3\lambda_n \sqrt{s_0} \|h_{T_0}\|_2 \leq \frac{3K\lambda_n \sqrt{s_0}}{\sqrt{n}} \|Xh\|_2 \quad (\text{by (F.5)}) \\ & 3K\lambda_n \sqrt{s_0} \left\|X\widehat{\beta} - X\beta\right\|_2 + 3K\lambda_n \sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} \\ & \leq 3K\lambda_n \sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} + \left\|X\widehat{\beta} - X\beta\right\|_2^2/n + (3K\lambda_n \sqrt{s_0})^2, \end{aligned} \quad (\text{F.6})$$

from which the following immediately follows: for $\lambda_n = d_0\lambda\sigma \geq 2\lambda_{\sigma,a,p}$, we have

$$\begin{aligned} \|h_{T_0^c}\|_1 & \leq \|X\beta - X\beta_0\|_2^2/(n\lambda_n) + 3K\sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} + (3K/2)^2 \lambda_n s_0 \\ & = \left(\|X\beta - X\beta_0\|_2/\sqrt{n\lambda_n} + (3K/2)\sqrt{\lambda_n s_0} \right)^2 := D'_1 \lambda \sigma s_0 \end{aligned}$$

where $D'_1 = (\sqrt{\Lambda_{\max}(s-s_0)/d_0} + 3K(s_0, 6)\sqrt{d_0}/2)^2$. Similarly, we can derive a bound on $\|h\|_1$ from (F.3); we have on \mathcal{T}_a ,

$$\begin{aligned} & \left\|X\widehat{\beta} - X\beta\right\|_2^2/n + \lambda_n \|h_{T_0^c}\|_1 + \lambda_n \|h_{T_0}\|_1 - \|X\beta - X\beta_0\|_2^2/n \leq 4\lambda_n \|h_{T_0}\|_1 \\ & \leq 4 \lambda_n \sqrt{s_0} \|h_{T_0}\|_2 \leq 4K\lambda_n \sqrt{s_0} \|Xh\|_2/\sqrt{n} \quad (\text{by (F.5)}) \\ & \leq 4K\lambda_n \sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} + \left\|X\widehat{\beta} - X\beta\right\|_2^2/n + (2K\lambda_n \sqrt{s_0})^2 \end{aligned}$$

Hence it is clear that for $\lambda_n = d_0\lambda\sigma \geq 2\lambda_{\sigma,a,p}$, we have by Lemma F.2, on \mathcal{T}_a ,

$$\begin{aligned} \|h\|_1 & \leq \|X\beta - X\beta_0\|_2^2/(n\lambda_n) + 4K\sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} + 4K^2 \lambda_n s_0 \\ & = \left(\|X\beta - X\beta_0\|_2/\sqrt{n\lambda_n} + 2K\sqrt{\lambda_n s_0} \right)^2 = D_2 \lambda \sigma s_0 \end{aligned}$$

where $D_2 = (\sqrt{\Lambda_{\max}(s-s_0)/d_0} + 2K(s_0, 6)\sqrt{d_0})^2$. Now we derive a bound for $\left\|X\widehat{\beta} - X\beta\right\|_2^2/n$; our starting point is (F.6), from which by shifting items around and adding $(3K\lambda_n \sqrt{s_0})^2$ to both sides, we obtain

$$\begin{aligned} & \left\|X\widehat{\beta} - X\beta\right\|_2^2/n - 3K\lambda_n \sqrt{s_0} \left\|X\widehat{\beta} - X\beta\right\|_2/\sqrt{n} + (3K\lambda_n \sqrt{s_0})^2 + \lambda_n \|h_{T_0^c}\|_1 \\ & \leq \|X\beta - X\beta_0\|_2^2/n + 3K\lambda_n \sqrt{s_0} \|X\beta - X\beta_0\|_2/\sqrt{n} + (3K\lambda_n \sqrt{s_0}/2)^2 \end{aligned}$$

Thus we have for $\lambda_n = d_0 \lambda \sigma \geq 2\lambda_{\sigma, a, p}$,

$$\left(\frac{1}{\sqrt{n}} \|X\hat{\beta} - X\beta\|_2 - \frac{3K\lambda_n\sqrt{s_0}}{2} \right)^2 + \lambda_n \|h_{T_0^c}\|_1 \leq (\|X\beta - X\beta_0\|_2/\sqrt{n} + 3K\lambda_n\sqrt{s_0}/2)^2$$

and hence

$$\begin{aligned} \|X\hat{\beta} - X\beta\|_2/\sqrt{n} &\leq \|X\beta - X\beta_0\|_2/\sqrt{n} + 3K\lambda_n\sqrt{s_0} \\ &\leq \lambda\sigma\sqrt{s_0} \left(\sqrt{\Lambda_{\max}(s - s_0)} + 3d_0K(s_0, 6) \right). \end{aligned} \quad (\text{F.7})$$

by Lemma F.2. Under $RE(s_0, 6, X)$ condition, we have by (F.7)

$$\begin{aligned} \|h_{T_0}\|_2 &\leq K(s_0, 6) \|Xh\|_2/\sqrt{n} \leq \frac{K(s_0, 6)}{\sqrt{n}} \left(\|X\hat{\beta} - X\beta\|_2 + \|X\beta - X\beta_0\|_2 \right) \\ &\leq K(s_0, 6) (2\|X\beta - X\beta_0\|_2/\sqrt{n} + 3K(s_0, 6)\lambda_n\sqrt{s_0}) \\ &\leq \lambda\sigma\sqrt{s_0}K(s_0, 6)(2\sqrt{\Lambda_{\max}(s - s_0)} + 3d_0K(s_0, 6)). \end{aligned}$$

Let T_1 be the s_0 largest positions of h outside of T_0 ; Now by a property as derived in Zhou (2009a) (Proposition A.1), we also have for $K := K(s_0, 6)$.

$$\|h_{T_{01}}\|_2 \leq \sqrt{2/n}K(s_0, 6) \|Xh\|_2 \leq \lambda\sigma\sqrt{2s_0}K(2\sqrt{\Lambda_{\max}(s - s_0)} + 3d_0K) \leq D_0\lambda\sigma\sqrt{s_0}$$

Moreover, we have by Lemma F.2,

$$\begin{aligned} \|\hat{\beta} - \beta\|_2^2 &\leq 2\|\hat{\beta} - \beta_{T_0}\|_2^2 + 2\|\beta - \beta_{T_0}\|_2^2 \leq 2\|h\|_2^2 + 2\lambda^2\sigma^2s_0 \\ &\leq 2(\|h_{T_{01}}\|_2^2 + \|h_{T_0^c}\|_1^2/s_0) + 2\lambda^2\sigma^2s_0 \leq 2\lambda^2\sigma^2s_0(D_0^2 + D_1^2 + 1) \end{aligned}$$

We note that (5.1) holds given (F.4) and (F.7). \square

Remark F.3. We could have bounded $\|h_{T_{01}}\|_2$ for the second case also by Lemma F.2; we take the form here for simplicity.

Proof of Lemma F.2. Decompose $h_{T_{01}^c}$ into h_{T_2}, \dots, h_{T_K} such that T_2 corresponds to locations of the s_0 largest coefficients of $h_{T_{01}^c}$ in absolute values, and T_3 corresponds to locations of the next s_0 largest coefficients of $h_{T_{01}^c}$ in absolute values, and so on. Let V be the span of columns of X_j , where $j \in T_{01}$, and P_V be the orthogonal projection onto V . Decompose $P_V Xh$:

$$P_V Xh = P_V Xh_{T_{01}} + \sum_{j \geq 2} P_V Xh_{T_j} = Xh_{T_{01}} + \sum_{j \geq 2} P_V Xh_{T_j}, \text{ where}$$

$$\|P_V Xh_{T_j}\|_2 \leq \frac{\sqrt{n}\theta_{s_0, 2s_0}}{\Lambda_{\min}(2s_0)} \|h_{T_j}\|_2 \text{ and } \sum_{j \geq 2} \|h_{T_j}\|_2 \leq \|h_{T_0^c}\|_1/\sqrt{s_0}$$

see Candès and Tao (2007)) for details; Thus we have

$$\begin{aligned} \|Xh_{T_{01}}\|_2 &= \left\| P_V Xh - \sum_{j \geq 2} P_V Xh_{T_j} \right\|_2 \leq \|P_V Xh\|_2 + \left\| \sum_{j \geq 2} P_V Xh_{T_j} \right\|_2 \\ &\leq \|Xh\|_2 + \sum_{j \geq 2} \|P_V Xh_{T_j}\|_2 \leq \|Xh\|_2 + \frac{\sqrt{n}\theta_{s_0, 2s_0}}{\sqrt{\Lambda_{\min}(2s_0)}\sqrt{s_0}} \|h_{T_0^c}\|_1, \end{aligned}$$

where we used the fact that $\|P_V\|_2 \leq 1$. Hence the lemma follows given $\|h_{T_{01}}\|_2 \leq \frac{1}{\sqrt{\Lambda_{\min}(2s_0)}\sqrt{n}} \|Xh_{T_{01}}\|_2$. For other bounds, the fact that the k th largest value of $h_{T_0^c}$ obeys $|h_{T_0^c}|_{(k)} \leq \|h_{T_0^c}\|_1/k$ has been used; see [Candès and Tao \(2007\)](#). \square

G Proofs for Lemmas in Section 6

Let $\lambda = \sqrt{2 \log p/n}$. By definition of s_0 as in (1.14), we have $\sum_{i=1}^p \min(\beta_i^2, \lambda^2 \sigma^2) \leq s_0 \lambda^2 \sigma^2$. We write $\beta = \beta^{(11)} + \beta^{(12)} + \beta^{(2)}$ where

$$\beta_j^{(11)} = \beta_j \cdot 1_{1 \leq j \leq a_0}, \quad \beta_j^{(12)} = \beta_j \cdot 1_{a_0 < j \leq s_0}, \quad \text{and} \quad \beta_j^{(2)} = \beta_j \cdot 1_{j > s_0}.$$

Now it is clear that $\sum_{j \leq a_0} \min(\beta_j^2, \lambda^2 \sigma^2) = a_0 \lambda^2 \sigma^2$ and hence

$$\sum_{j > a_0} \min(\beta_j^2, \lambda^2 \sigma^2) = \left\| \beta^{(12)} + \beta^{(2)} \right\|_2^2 \leq (s_0 - a_0) \lambda^2 \sigma^2. \quad (\text{G.1})$$

Proof of Lemma 6.1. It is clear for $\mathcal{D}_{11} = \mathcal{D} \cap A_0$, we have $\mathcal{D}_{11} \subset A_0 \subset T_0 \subset S$. Let $\beta_{\mathcal{D}}^{(11)} := (\beta_j)_{j \in A_0 \cap \mathcal{D}}$ consist of coefficients of β that are above $\lambda \sigma$ in their absolute values but are dropped as $\beta_{j, \text{init}} < t_0$. Now by (G.1), we have

$$\|\beta_{\mathcal{D}}\|_2^2 \leq \left\| \beta_{\mathcal{D}}^{(11)} \right\|_2^2 + \left\| \beta^{(12)} + \beta^{(2)} \right\|_2^2 \leq \left\| \beta_{\mathcal{D}}^{(11)} \right\|_2^2 + (s_0 - a_0) \lambda^2 \sigma^2,$$

where $|\mathcal{D}_{11}| \leq a_0$ and thus we have by the triangle inequality,

$$\begin{aligned} \left\| \beta_{\mathcal{D}}^{(11)} \right\|_2 &\leq \|\beta_{\mathcal{D}_{11}, \text{init}}\|_2 + \left\| \beta_{\mathcal{D}_{11}, \text{init}} - \beta_{\mathcal{D}}^{(11)} \right\|_2 \leq t_0 \sqrt{|\mathcal{D}_{11}|} + \|h_{\mathcal{D}_{11}}\|_2 \\ &\leq t_0 \sqrt{a_0} + \|h_{\mathcal{D}_{11}}\|_2; \end{aligned} \quad (\text{G.2})$$

Thus (6.2) holds. Now we replace the bound of $|\mathcal{D}_{11}| \leq a_0$ with $|\mathcal{D}_{11}| \leq \frac{\|h_{\mathcal{D}_{11}}\|_2^2}{|\beta_{\min, A_0} - t_0|^2}$ in (G.2) to obtain

$$\left\| \beta_{\mathcal{D}}^{(11)} \right\|_2 \leq t_0 \frac{\|h_{\mathcal{D}_{11}}\|_2}{\beta_{\min, A_0} - t_0} + \|h_{\mathcal{D}_{11}}\|_2 = \|h_{\mathcal{D}_{11}}\|_2 \frac{\beta_{\min, A_0}}{\beta_{\min, A_0} - t_0}$$

which proves (6.3). \square

Proof of Lemma 6.2. Suppose $\mathcal{T}_a \cap Q_c$ holds. It is clear by the choice of t_0 in (6.5) and by (6.4) that $\min_{i \in A_0} \widehat{\beta}_i \geq \beta_{\min, A_0} - \|h_{A_0}\|_\infty \geq t_0$ and $\mathcal{D}_{11} = \emptyset$. Thus by (6.5), we can bound $|I \cap T_0^c|$, depending on which one is applicable, by $|I \cap T_0^c| \leq \|\beta_{T_0^c, \text{init}}\|_1/t_0 \leq \check{s}_0$ or by $|I \cap T_0^c| \leq \|\beta_{T_0^c, \text{init}}\|_2^2/t_0^2 \leq \check{s}_0$. Moreover, the bounds on $\left\| \widehat{\beta}_I - \beta \right\|_2^2$ follows immediately from Lemma 4.3, on event \mathcal{T}_a , where $\theta_{|I|, |S_{\mathcal{D}}|}^2$ is bounded in Lemma 5.4 given $|I| + |S_{\mathcal{D}}| \leq s + |I \cap T_0^c| \leq s + \check{s}_0 \leq 2s$. \square

Acknowledgment. The author thanks Larry Wasserman and Peter Bühlmann for helpful discussions at the early stage of this work, and Sara van de Geer for her positive feedback when I presented this work in the Workshop of the DFG-SNF Research Group at University of Bern, in September 2009. The author would like to thank the NIPS 2009 reviewers for their constructive comments, and John Lafferty and Jun Gao for their encouragements and support.

References

- ADAMCZAK, R., LITVAK, A. E., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2009). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. 0904.4723v1.
- BARANIUK, R. G., DAVENPORT, M., DEVORE, R. A. and WAKIN, M. B. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* **28** 253–263.
- BARRON, A., BIRGE, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301–413.
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732.
- BIRGE, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*.
- BIRGE, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268.
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007a). Aggregation for gaussian regression. *Annals of Statistics* **35** 1674–1697.
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007b). Sparse density estimation with ℓ_1 penalties. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT'07)*.
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007c). Sparsity oracle inequalities for the Lasso. *The Electronic Journal of Statistics* **1** 169–194.
- CAI, T., WANG, L. and XU, G. (2009). Stable recovery of sparse signals and an oracle inequality. Tech. rep., Department of Statistics, The Wharton School, University of Pennsylvania.
- CANDÈS, E. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics* **37** 2145–2177.
- CANDÈS, E., ROMBERG, J. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics* **59** 1207–1223.
- CANDÈS, E. and TAO, T. (2005). Decoding by Linear Programming. *IEEE Trans. Info. Theory* **51** 4203–4215.
- CANDÈS, E. and TAO, T. (2006). Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory* **52** 5406–5425.
- CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics* **35** 2313–2351.

- CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing* **20** 33–61.
- DONOHO, D. (2006a). Compressed sensing. *IEEE Trans. Info. Theory* **52** 1289–1306.
- DONOHO, D. (2006b). For most large underdetermined systems of equations, the minimal ℓ_1 -norm solution is also the sparsest solution. *Communications in Pure and Applied Mathematics* **59** 797–829.
- DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–499.
- FOSTER, D. and GEORGE, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.
- GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.
- HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618.
- KOLTCHINSKII, V. (2009a). Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828.
- KOLTCHINSKII, V. (2009b). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7–57.
- LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2** 90–102.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.
- MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246–270.
- MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2008). Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation* **28** 277–289.
- NEEDEL, D. and TROPP, J. A. (2008). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis* **26** 301–321.
- NEEDEL, D. and VERSHYNIN, R. (2009). Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, to appear .

- RASKUTTI, G., WAINWRIGHT, M. and YU, B. (2009). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. In *Allerton Conference on Control, Communication and Computer*. Longer version in arXiv:0910.2042v1.pdf.
- RAVIKUMAR, P., WAINWRIGHT, M. and LAFFERTY, J. (2008). High-dimensional ising model selection using ℓ_1 -regularized logistic regression. *Annals of Statistics* To appear; Also in Technical Report 750, Department of Statistics, U.C. Berkeley.
- RUDELSON, M. and VERSHYNIN, R. (2006). Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *40th Annual Conference on Information Sciences and Systems (CISS 2006)*.
- SZAREK, S. (1991). Condition numbers of random matrices. *J. Complexity* **7** 131–148.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
- VAN DE GEER, S. and BUHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3** 1360–1392.
- VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2010). Prediction and variable selection with the adaptive lasso. ArXiv:1001.5176v1.pdf.
- VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36** 614–645.
- WAINWRIGHT, M. (2009a). Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741.
- WAINWRIGHT, M. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming. *IEEE Trans. Inform. Theory* **55** 2183–2202.
- WASSERMAN, L. and ROEDER, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37** 2178–2201.
- ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567–1594.
- ZHANG, T. (2009). Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics* **37** 2109–2144.
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2567.
- ZHOU, S. (2009a). Restricted eigenvalue conditions on subgaussian random matrices. ArXiv:0904.4723v2.
- ZHOU, S. (2009b). Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems 22*. MIT Press.

ZHOU, S., VAN DE GEER, S. and BÜHLMANN, P. (2009). Adaptive Lasso for high dimensional regression and gaussian graphical modeling. *ArXiv:0903.2515*.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.