PIVOTAL ESTIMATION OF NONPARAMETRIC FUNCTIONS VIA
SQUARE-ROOT LASSO

ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV, AND LIE WANG

ABSTRACT. In a nonparametric linear regression model we study a variant of LASSO, called \( \sqrt{\text{LASSO}} \), which does not require the knowledge of the scaling parameter \( \sigma \) of the noise or bounds for it. This work derives new finite sample upper bounds for prediction norm rate of convergence, \( \ell_1 \)-rate of convergence, \( \ell_\infty \)-rate of convergence, and sparsity of the \( \sqrt{\text{LASSO}} \) estimator. A lower bound for the prediction norm rate of convergence is also established.

In many non-Gaussian noise cases, we rely on moderate deviation theory for self-normalized sums and on new data-dependent empirical process inequalities to achieve Gaussian-like results provided \( \log p = o(n^{1/3}) \) improving upon results derived in the parametric case that required \( \log p \lesssim \log n \).

In addition, we derive finite sample bounds on the performance of ordinary least square (OLS) applied to the model selected by \( \sqrt{\text{LASSO}} \) accounting for possible misspecification of the selected model. In particular, we provide mild conditions under which the rate of convergence of OLS post \( \sqrt{\text{LASSO}} \) is not worse than \( \sqrt{\text{LASSO}} \).

We also study two extreme cases: parametric noiseless and nonparametric unbounded variance. \( \sqrt{\text{LASSO}} \) does have interesting theoretical guarantees for these two extreme cases. For the parametric noiseless case, differently than LASSO, \( \sqrt{\text{LASSO}} \) is capable of exact recovery. In the unbounded variance case it can still be consistent since its penalty choice does not depend on \( \sigma \).

Finally, we conduct Monte Carlo experiments which show that the empirical performance of \( \sqrt{\text{LASSO}} \) is very similar to the performance of LASSO when \( \sigma \) is known. We also emphasize that \( \sqrt{\text{LASSO}} \) can be formulated as a convex programming problem and its computation burden is similar to LASSO. We provide theoretical and empirical evidence of that.
1. Introduction

We consider the problem of recovering a nonparametric regression function, where the underlying function of interest has unknown function form of basic covariates. To be more specific, we consider a nonparametric regression model:

\[ y_i = f(z_i) + \sigma \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1} \]

where \( y_i \) are the outcomes, \( z_i \) are vectors of fixed basic covariates, \( \varepsilon_i \) are i.i.d. noises, \( f \) is the regression function, and \( \sigma \) is a scaling parameter. Our goal is to recover the regression function \( f \). To achieve this goal, we use linear combinations of regressors \( x_i = P(z_i) \) to approximate \( f \), where \( P(z_i) \) is a \( p \)-vector of transformations of \( z_i \). We are interested in the high dimension low sample size case, in which we potentially have \( p \geq n \), to attain a flexible functional form. In particular, we are interested in a sparse model over the regressors \( x_i \) to describe the regression function.

Now the model is written as

\[ y_i = x_i^T \beta_0 + u_i, \quad u_i = r_i + \sigma \varepsilon_i. \]

In the above expression, \( r_i := f_i - x_i^T \beta_0 \) is the approximation error, where \( f_i = f(z_i) \).

Assume that the cardinality of the support of coefficient \( \beta_0 \) is

\[ s := |T| = \| \beta_0 \|_0, \]

where \( T = \text{supp}(\beta_0) \). It is well known that ordinary least square is generally inconsistent when \( p > n \). However, the sparsity assumption makes it possible to estimate these models effectively by searching for approximately the right set of the regressors. In particular, \( \ell_1 \)-based penalization methods have been playing a central role in this question. Many papers have studied the estimation of high dimensional mean regression models with the \( \ell_1 \)-norm acting as a penalty function \([6, 11, 15, 19, 34, 39, 38]\). In these references, under appropriate choice of penalty level, it was demonstrated that the \( \ell_1 \)-penalized least squares estimators achieve the rate \( \sigma \sqrt{s/n} \sqrt{\log p} \), which is very close to the oracle rate \( \sigma \sqrt{s/n} \) achievable when the true model is known. We refer to \([4, 6, 8, 9, 11, 13, 17, 24, 34]\) for many other developments and a more detailed review of the existing literature.

An important \( \ell_1 \)-based estimator proposed in \([30]\) is the LASSO estimator defined as

\[ \hat{\beta}^L \in \arg \min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n} \| \beta \|_1, \tag{1.2} \]
where, for observations of a response variable $y_i$ and regressors $x_i$, \( \hat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2] = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2, \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \) and $\lambda$ is the penalty level. We note that the LASSO estimator minimizes a convex function. Therefore, from a computational complexity perspective, (1.2) is a computationally efficient (polynomial time) alternative to exhaustive combinatorial search over all possible submodels.

The performance of LASSO relies heavily on the penalization parameter $\lambda$ which should majorate the non-negligible spurious correlation between the noise terms and the large number of additional regressors. The typical choice of $\lambda$ is proportional to the unknown scaling parameter $\sigma$ of the noise (typically the standard deviation). Simple upper bounds for $\sigma$ can be derived based on the empirical variance of the response variable. However, upper bounds on $\sigma$ can lead to unnecessary over regularization which translates into larger bias and slower rates of convergence. Moreover, such over regularization can lead to the exclusion of relevant regressors from the selected model harming post model selection estimators.

In this paper, we have three sets of main results. The first contribution is to study a variant of (1.2), called $\sqrt{\text{LASSO}}$, which does not require the knowledge of $\sigma$ or bounds of it but it is still computationally attractive. The $\sqrt{\text{LASSO}}$ estimator is defined as

\[
\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \sqrt{\hat{Q}(\beta)} + \frac{\lambda}{n} \|\beta\|_1.
\] (1.3)

In the parametric model studied in [5], the choice of the penalty parameter becomes pivotal given the covariates and the distribution of the error term. In contrast, in the nonparametric setting we need to account for the impact of the approximation error to derive a practical and theoretical justified choice of penalty level. We rely on moderate deviation theory for self-normalized sums and on data-dependent empirical process inequalities to achieve Gaussian-like results in many non-Gaussian cases provided $\log p = o(n^{1/3})$ improving upon results derived in the parametric case that required $\log p \lesssim \log n$, see [5]. We perform a thorough non-asymptotic theoretical analysis of the choice of the penalty parameter.

The second set of contributions is to derive upper bounds for prediction norm rate of convergence, $\ell_1$-rate of convergence, $\ell_\infty$-rate of convergence, and sparsity of the $\sqrt{\text{LASSO}}$ estimator. A lower bound on the rate of convergence for the prediction norm is also established. Furthermore, we also study two extreme cases: (i) parametric noiseless and (ii)
nonparametric unbounded variance. $\sqrt{\text{LASSO}}$ does have interesting theoretical guarantees for these two extreme cases. For the parametric noiseless case, for a wide range of the penalty level, $\sqrt{\text{LASSO}}$ achieves exact recovery in sharp contrast to LASSO. In the nonparametric unbounded variance case, $\sqrt{\text{LASSO}}$ estimator can still be consistent since its penalty choice does not depend on the standard deviation of the noise. We develop the necessary modifications on the oracle definition and penalty level, and derive finite sample bounds for the case the noise has a Student’s $t$-distribution with 2 degrees of freedom.

The third contribution aims to remove the potentially significant bias towards zero introduced by the $\ell_1$-norm regularization employed in (1.3). We consider the post model selection estimator that applies ordinary least squares (OLS) regression to the model $\hat{T}$ selected by $\sqrt{\text{LASSO}}$. Formally, set

$$\hat{T} = \text{supp}(\hat{\beta}) = \{ j \in \{1, \ldots, p\} : |\hat{\beta}_j| > 0 \},$$

and define the OLS post $\sqrt{\text{LASSO}}$ estimator $\tilde{\beta}$ as

$$\tilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sqrt{\hat{Q}(\beta)} : \beta_j = 0 \text{ if } j \notin \hat{T}^c. \quad (1.4)$$

It follows that if the model selection works perfectly (i.e., $\hat{T} = T$) then the OLS post $\sqrt{\text{LASSO}}$ estimator is simply the oracle estimator whose properties are well known. Unfortunately, perfect model selection might be unlikely for many designs of interest. This is usually the case in a nonparametric setting. Thus, we are also interest on the properties of OLS post $\sqrt{\text{LASSO}}$ when $\hat{T} \neq T$, including cases where $T \not\subseteq \hat{T}$.

Finally, we emphasize that $\sqrt{\text{LASSO}}$ can be formulated as a convex programming problem which allows us to rely on many efficient algorithmic implementations to compute the estimator (interior point methods [31, 32], first order methods [1, 2], and coordinatewise methods). Importantly, the computation cost of $\sqrt{\text{LASSO}}$ is similar to LASSO. We conduct Monte Carlo experiments which show that the empirical performance of $\sqrt{\text{LASSO}}$ is very similar to the performance of LASSO when $\sigma$ is known.

2. Nonparametric Regression Model

Recall the nonparametric regression model:

$$y_i = f(z_i) + \sigma \epsilon_i, \quad \epsilon_i \sim F_0, \quad E[\epsilon_i] = 0, \quad E[\epsilon_i^2] = 1, \quad i = 1, \ldots, n. \quad (2.5)$$
The model can also be written as
\[ y_i = x'_i \beta_0 + u_i, \quad u_i = r_i + \sigma \epsilon_i. \] (2.6)

In many applications of interest there is no exact sparse model or, due to noise, it might be inefficient to rely on an exact model. However, there might be a sparse model that yields a good approximation to the true regression function \( f \) in equation (2.5). In this case, the target linear combination is given by any vector \( \beta_0 \) that solves the following “oracle” risk minimization problem:
\[
\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(f_i - x'_i \beta)^2] + \sigma^2 \| \beta_0 \| / n,
\] (2.7)
where \( \mathbb{E}_n[z_i] = (1/n) \sum_{i=1}^n z_i, f_i = f(z_i), \) and the corresponding cardinality of its support \( T = \text{supp}(\beta_0) \) is
\[ s := |T| = \| \beta_0 \|_0. \]

The oracle balances the approximation error \( \mathbb{E}_n[(f_i - x'_i \beta)^2] \) with the variance term \( \sigma^2 \| \beta_0 \| / n \), where the latter is determined by the complexity of the model – the number of non-zero coefficients of \( \beta \). The average square error from approximating \( f_i \) by \( x'_i \beta_0 \) is denote by
\[ c^2_s := \mathbb{E}_n[r_i^2] = \mathbb{E}_n[(f_i - x'_i \beta_0)^2], \]
so that \( c^2_s + \sigma^2 s / n \) is the optimal value of (2.7). In general the support of the best sparse approximation \( T = \text{supp}(\beta_0) \) is unknown since we do not observe \( f_i \).

We consider the case of fixed design, namely we treat the covariate values \( x_1, \ldots, x_n \) as fixed. Without loss of generality, we normalize the covariates so that
\[ \mathbb{E}_n[x^2_{ij}] = 1 \text{ for } j = 1, \ldots, p. \] (2.8)

We summarize the previous setting in the following condition.

**Condition 1 (ASM).** We have data \( \{(y_i, z_i) : i = 1, \ldots, n\} \) that for each \( n \) obey the regression model (2.5), which admits the approximately sparse form (2.6) with \( \beta_0 \) defined by (2.7). The regressors \( x_i = P(z_i) \) are normalized as in (2.8).

The main focus of the literature is on deriving rate of convergence results in the prediction norm, which measures the accuracy of predicting the true regression function over the
design points $x_1, \ldots, x_n$, $\|\delta\|_{2,n} = \sqrt{E_n[(x_i^T\delta)^2]}$. It follows that the least square criterion $\hat{Q}(\beta) = E_n[(y_i - x_i^T\beta)^2]$ satisfies,

$$\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) = \|\hat{\beta} - \beta_0\|_{2,n}^2 - 2E_n[(\sigma \epsilon_i + r_i)x_i^T(\hat{\beta} - \beta_0)].$$

(2.9)

Thus, $\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0)$ provides noisy information about $\|\hat{\beta} - \beta_0\|_{2,n}$, and the discrepancy is controlled by the noise and approximation term

$$S = 2E_n[\sigma \epsilon_i x_i] \quad \text{and} \quad 2E_n[r_i x_i].$$

Moreover, the estimation of $\beta_0$ in the prediction norm allows to bound the empirical risk by the triangular inequality:

$$\sqrt{E_n[(x_i^T\hat{\beta} - f_i)^2]} \leq \|\hat{\beta} - \beta_0\|_{2,n} + c_s.$$  

(2.10)

2.1. Conditions on the Gram Matrix. It is known that the Gram matrix $E_n[x_i x_i^T]$ plays an important role in the analysis of estimators in this setup. In our case, the smallest eigenvalue of the Gram matrix has zero value if $p > n$ which creates potential identification problems. Thus, to restore identification, one needs to restrict the type of deviation vectors $\delta$ from $\beta_0$ that we will consider. It will be important to consider vectors $\delta$ that belong to the restricted set $\Delta_{\bar{c}}$ defined as

$$\Delta_{\bar{c}} = \{ \delta \in \mathbb{R}^p : \|\delta\|_{1} \leq \bar{c} \|\delta\|_{1}, \delta \neq 0 \}, \quad \text{for} \quad \bar{c} \geq 1.$$  

We will state the bounds in terms of the following restricted eigenvalue of the Gram matrix $E_n[x_i x_i^T]$:

$$\kappa_{\bar{c}} := \min_{\delta \in \Delta_{\bar{c}}} \frac{\sqrt{\mathbb{E}[\|\delta\|_{2,n}^2]}}{\|\delta\|_{1}}.$$  

(2.11)

The restricted eigenvalue can depend on $n$ and $T$, but we suppress the dependence in our notations. The restricted eigenvalue [2.11] is a variant of the restricted eigenvalue introduced in Bickel, Ritov and Tsybakov [6].

Next consider the minimal and maximal $m$-sparse eigenvalues of the Gram matrix,

$$\phi_{\min}(m) := \min_{\|\delta\|_{0} \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_{2}^2}, \quad \text{and} \quad \phi_{\max}(m) := \max_{\|\delta\|_{0} \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_{2}^2}.$$  

(2.12)

They also play an important role in the analysis. Moreover, sparse eigenvalues provide a simple sufficient condition to bound restricted eigenvalues. Indeed, following [6], we can
bound $\kappa_c$ from below by

$$\kappa_c \geq \max_{m \geq 0} \phi_{\min}(m) \left( 1 - \frac{\phi_{\max}(m)}{\phi_{\min}(m)} c \sqrt{s/m} \right).$$

Thus, if $m$-sparse eigenvalues are bounded away from zero and from above

$$0 < k \leq \phi_{\min}(m) \leq \phi_{\max}(m) \leq k' < \infty, \quad \text{for all} \quad m \leq 4(k'/k)^2 c^2 s,$$

then $\kappa_c \geq \phi_{\min}(4(k'/k)^2 c^2 s)/2$. We note that (2.13) only requires the eigenvalues of certain “small” $m \times m$ submatrices of the large $p \times p$ Gram matrix to be bounded from above and below. Many sufficient conditions for (2.13) are provided by [6], [39], and [19]. Bickel, Ritov, and Tsybakov [6] and others also provide different sets of sufficient primitive conditions for $\kappa_c$ to be bounded away from zero.

3. Pivotal Penalty Level for $\sqrt{\text{LASSO}}$

Here we propose a pivotal, data-driven choice for the regularization parameter value $\lambda$ in the case that the distribution of the disturbances is known up to a scaling factor (in the traditional case we would have unknown variance but normality of $\epsilon_i$). Since the objective function in the optimization problem (1.3) is not pivotal in either small or large samples, finding a pivotal $\lambda$ appears to be difficult a priori. However, the insight is to consider the gradient at the true parameter that summarizes the noise in the estimation problem. We note that the principle of setting $\lambda$ to dominate the score of the criterion function is motivated by [6]’s choice of penalty level for LASSO. In fact, this is a general principle that carries over to other convex problems and that leads to the near-oracle performance of $\ell_1$-penalized estimators.

The key quantity determining the choice of the penalty level for $\sqrt{\text{LASSO}}$ is the score

$$\tilde{S} := \frac{\mathbb{E}_n[x_i(\sigma \epsilon_i + r_i)]}{\sqrt{\mathbb{E}_n[(\sigma \epsilon_i + r_i)^2]}} \quad \text{if} \quad \mathbb{E}_n[(\sigma \epsilon_i + r_i)^2] > 0, \quad \text{and} \quad \tilde{S} := 0 \quad \text{otherwise.}$$

The score $\tilde{S}$ summarizes the estimation noise and approximation error in our problem, and we may set the penalty level $\lambda/n$ to dominate this term. For efficiency reasons, we set $\lambda/n$ at a smallest level that dominates the estimation noise, namely we choose the smallest $\lambda$ such that

$$\lambda \geq c\bar{\Lambda}, \quad \text{for} \quad \bar{\Lambda} := n \|\tilde{S}\|_{\infty}, \quad \text{(3.14)}$$
with a high probability, say $1 - \alpha$, where $\tilde{\Lambda}$ is the maximal score scaled by $n$, and $c > 1$ is a theoretical constant of [6] to be stated later. The event (3.14) implies that

$$\hat{\beta} - \beta_0 \in \Delta_{\tilde{c}}, \quad \text{where } \tilde{c} = \frac{c + 1}{c - 1}$$

so that rates of convergence can be attained using the restricted eigenvalue $\kappa_{\tilde{c}}$.

It is worth pointing out that in the parametric case, $r_i = 0$, $i = 1, \ldots, n$, the score $\tilde{S}$ does not depend on the unknown scaling parameter $\sigma$ or the unknown true parameter value $\beta_0$. Therefore, the score is pivotal with respect to these parameters. Moreover, under the additional classical normality assumption, namely $F_0 = \Phi$, the score is in fact completely pivotal, conditional on $X$. This means that in principle we know the distribution of $\tilde{S}$ in this case, or at least we can compute it by simulation, see [5].

However, in the nonparametric case the design $X$ also determines the unknown approximation errors $r_i$. To achieve an implementable choice of penalty level, we will consider the following random variable

$$\Lambda := \frac{n\|E_n[\sigma \epsilon_i x_i]\|_\infty + \sigma \sqrt{n}}{\sqrt{E_n[\sigma^2 \epsilon_i^2]}} = \frac{n\|E_n[\epsilon_i x_i]\|_\infty + \sqrt{n}}{\sqrt{E_n[\epsilon_i^2]}}. \quad (3.15)$$

$\Lambda$ also does not depend on the unknown scaling parameter $\sigma$ or the unknown true parameter value $\beta_0$, and therefore, it is pivotal with respect to these parameters. We will show that

if $\sqrt{E_n[\sigma^2 \epsilon_i^2]} \leq (1 + u_n)\sqrt{E_n[(\sigma \epsilon_i + r_i)^2]}$ we have $\tilde{\Lambda} \leq (1 + u_n)\Lambda. \quad (3.16)$

It will follow that the condition above is satisfied with high probability even for $u_n = o(1)$.

We propose two choices for the penalty level $\lambda$. For $0 < \alpha < 1$, $c > 1$, and $u_n \geq 0$ define:

- **exact** $\lambda = (1 + u_n) \cdot c \cdot \Lambda(1 - \alpha |X)$,
- **asymptotic** $\lambda = (1 + u_n) \cdot c \cdot \sqrt{n} \Phi^{-1}(1 - \alpha/2p)$,

where $\Phi(\cdot)$ is the cumulative distribution function for a standard Gaussian random variable, and $\Lambda(1 - \alpha |X) := (1 - \alpha)$-quantile of $\Lambda |X, \epsilon_i \sim F_0$. We can accurately approximate the latter by simulation and the former by numerical integration.

The parameter $1 - \alpha$ is a confidence level which guarantees near-oracle performance with probability at least $1 - \alpha$; we recommend $1 - \alpha = 95\%$. The constant $c > 1$ is the slack parameter in (3.14) used in [6]; we recommend $c = 1.1$. The parameter $u_n$ is intended to account for the approximation errors to achieve (3.16); we recommend $u_n = 0.05$. These
recommendations are valid either in finite or large samples under the conditions stated below. They are also supported by the finite-sample experiments reported in Section 4. The exact option is applicable when we know that the distribution of the errors $F_0$, for example in the classical Gaussian errors case. As stated previously, we can approximate the quantiles $\Lambda(1 - \alpha |X)$ by simulation. Therefore, we can implement the exact option. The asymptotic option is applicable when $F_0$ and design $X$ satisfy some moment conditions stated below. We recommend using the exact, when applicable, since it is better tailored to the given design $X$ and sample size $n$. However, the asymptotic option is trivial to compute and it often provides a very similar penalty level.

3.1. Analysis of the Penalty Choices. In this section we formally analyze the penalty choices described in (3.17). In particular we are interested on establishing bounds on the probability of (3.14) occurring.

In order to achieve Gaussian-like behavior for non-Gaussian disturbances we have to rely on some moment conditions for the noise, on some restrictions on the growth of $p$ relative to $n$, and also consider $\alpha$ that is either bounded away from zero or approaches zero not too rapidly. In this section we focus on the following set of conditions.

**CONDITION M’.** There exist a finite constant $q > 8$ such that the disturbance obeys

$$\sup_{n \geq 1} E_{F_0}[|\epsilon|^q] < \infty,$$

and the design $X$ obeys

$$\sup_{n \geq 1} \max_{1 \leq j \leq p} E_n[|x_{ij}|^q] < \infty.$$

**CONDITION R’.** Let $r_n = \left(\alpha^{-1} \log n C_q E[|\epsilon|^{q/4}] \right)^{1/q} / n^{1/4} < 1/2$, and define the constants $c_1 = 1 + 4 \sqrt{\frac{\log(\log n)}{n \alpha / \log n}} \left(\frac{3E[\epsilon]^q}{\alpha / \log n}\right)^{4/q} \max_{1 \leq j \leq p} (E_n[|x_{ij}|^q])^{1/4}$ and $c_2 = 1 / (1 - 3^{1/q} r_n)$. Also, assume that $n^{1/6} \geq (\Phi^{-1}(1 - \alpha/2p) + 1) \max_{1 \leq j \leq p} (E_n[|x_{ij}|^3] E[|\epsilon|])^{1/3}$. Let $u_n \geq 0$ be such that $u_n \Phi^{-1}(1 - \alpha/2p) \geq 4 / (1 - r_n)$, $u_n \geq 4(\sqrt{c_1 c_2} - 1)$.

The growth condition depends on the number of bounded moments $q$ of regressors and of the noise term. Under condition M’ and $\alpha$ fixed, condition R’ is satisfied if $\log p = o(n^{1/3})$. This is asymptotically less restrictive that the condition $\log p \leq (q - 2) \log n$ required in [5]. However, condition M’ is more stringent than some conditions in [5] thus neither set of condition dominates the other. (Due to similarities, we defer the result under the conditions considered in [5] to the appendix.)

The main insight of the new analysis is the use of the theory of moderate deviation for self normalized sums. Theorems 1 and 2 show that the options (3.17) implement the
regularization event \( \lambda > c\Lambda \) in the non-Gaussian case with asymptotic probability at least \( 1 - \alpha \). Theorem 3 bounds the magnitude of the penalty level \( \lambda \) for the exact option.

**Theorem 1 (Coverage of the Exact Penalty Level).** Suppose that conditions ASM, M' and R' hold. Then, the exact option in (3.17) implements \( \lambda > c\Lambda \) with probability at least

\[
1 - \alpha \left( 1 + \frac{1}{\log n} \right) - \frac{4(1 + u_n)}{nu_n}.
\]

**Theorem 2 (Coverage of the Asymptotic Penalty Level).** Suppose that conditions ASM, M' and R' hold. Then, the asymptotic option in (3.17) implements \( \lambda > c\Lambda \) with probability at least

\[
1 - \alpha \left( 1 + \frac{A}{\ell_n^3} + \frac{2}{\log n} \right) - \frac{20}{n(u_n \wedge 1)}
\]

where \( \ell_n = n^{1/6}/[(\Phi^{-1}(1-\alpha/2p) + 1)\max_{1 \leq j \leq p}(\mathbb{E}[|x_{ij}|^3]\mathbb{E}[|\epsilon_i|^3])^{1/3}] \), and \( A \) is an universal constant.

**Theorem 3 (Bound on the Exact Penalty Level).** Suppose that conditions ASM, M' and R' hold. Then, the approximate score \( \Lambda(1-\alpha|X) \) in (3.17) satisfies

\[
\Lambda(1-\alpha|X) \leq 2\sqrt{n} + \sqrt{n}\Phi^{-1}(1-\alpha/2p)\sqrt{c_1c_2} \left( 1 + \frac{2\log(3 + 3A/\ell_n^3)/\Phi^{-1}(1-\alpha/2p)}{\mathbb{E}[|x_{ij}|^3]\mathbb{E}[|\epsilon_i|^3])^{1/3}] \right)
\]

where \( \ell_n = n^{1/6}/[(\Phi^{-1}(1-\alpha/2p) + 1)\max_{1 \leq j \leq p}(\mathbb{E}[|x_{ij}|^3]\mathbb{E}[|\epsilon_i|^3])^{1/3}] \), and \( \Phi^{-1}(1-\alpha/2p) \leq \sqrt{2\log(p/\alpha)} \).

Under conditions on the growth of \( p \) relative to \( n \), these theorems establish that many nice properties of the penalty level in the Gaussian case continue to hold in many non-Gaussian cases. The following corollary summarizes the asymptotic behavior of the penalty choices (3.17).

**Corollary 1.** Suppose that conditions ASM, M' and R' hold, and \( \lambda \) is chosen either following the exact or asymptotic rule in (3.17). If \( \alpha/p = o(1) \), \( \log(p/\alpha) = o(n^{1/3}) \), \( 1/\alpha = o(n/\log p) \) then there exists \( u_n = 1 + o(1) \) such that \( \lambda \) satisfies

\[
P(\lambda \geq c\Lambda|X) \geq 1 - \alpha(1 + o(1)) \quad \text{and} \quad \lambda \leq (1 + o(1))c\sqrt{2n\log(p/\alpha)}.
\]

4. **Finite-Sample Analysis of \( \sqrt{\text{LASSO}} \)**

Next we establish several finite sample results regarding the \( \sqrt{\text{LASSO}} \) estimator. We highlight several differences between the \( \sqrt{\text{LASSO}} \) analysis conducted here and traditional
analysis of LASSO. First, we do not assume Gaussian or sub-Gaussian noise. Second, the
value of the scaling parameter $\sigma$ is unknown, and the penalty level $\lambda$ does not depend on
the scaling parameter $\sigma$. Third, the necessity of a mild side condition to hold which ensures
that the penalty does not overrule the identification. Fourth, most of the analysis of this
section is conditional not only on the covariates $x_1, \ldots, x_n$, but also on the noise $\epsilon_1, \ldots, \epsilon_n$,
through the event $\lambda \geq c\tilde{A}$. Therefore, by choosing $\lambda$ as in (3.17) the event $\lambda \geq c\tilde{A}$ occurs
with high probability and the stated results hold.

4.1. Finite-Sample Bounds on Different Norms. We begin with a finite sample bound
for the prediction norm which is similar to the bound obtained by the LASSO estimator
that knows $\sigma$.

**Theorem 4 (Finite Sample Bounds on Estimation Error).** Under condition ASM, let $c > 1$,
$\bar{c} = (c + 1)/(c - 1)$, and suppose that $\lambda$ obeys the growth restriction $\lambda\sqrt{s} < n\kappa\epsilon$. If $\lambda \geq c\tilde{A}$,
then

$$||\hat{\beta} - \beta_0||_{2,n} \leq \frac{2(1 + 1/c)}{1 - \frac{(\lambda\sqrt{s})}{n\kappa\epsilon}} \sqrt{Q(\beta_0)} \frac{\lambda\sqrt{s}}{n\kappa\epsilon}.$$ 

We recall that the choice of $\lambda$ does not depend on the scaling parameter $\sigma$. The impact
of $\sigma$ in the bound above comes through the factor

$$\sqrt{Q(\beta_0)} \leq \sigma \sqrt{\mathbb{E}_n[\epsilon^2_1]} + c_s.$$ 

Thus, this result leads to the same rate of convergence as in the case of the LASSO estimator
that knows $\sigma$ since $\mathbb{E}_n[\epsilon^2_1]$ concentrates around one under (2.5) and the law of large numbers.

As mentioned before, the analysis of $\sqrt{\text{LASSO}}$ raises several different issues from that of
LASSO, and so the proof of Theorem 4 is different. In particular, we need to invoke the
additional growth restriction, $\lambda\sqrt{s} < n\kappa\epsilon$, which is not present in the LASSO analysis that
treats $\sigma$ as known. This is required because the introduction of the square-root removes
the quadratic growth which would eventually dominates the $\ell_1$ penalty for large enough
deviations from $\beta_0$. This condition ensures that the penalty is not too large so identification
of $\beta_0$ is still possible. However, when this side condition fails and $\sigma$ is bounded away from
zero, LASSO is not guaranteed to be consistent since its rate of convergence is typically
given by $\sigma\lambda\sqrt{s}/[n\kappa\epsilon]$. 

Also, the event \( \lambda \geq c \Lambda \) accounts for the approximation errors \( r_1, \ldots, r_n \). That has two implications. First, the impact of \( c_s \) on the estimation of \( \beta_0 \) is diminished by a factor of \( \lambda \sqrt{s}/[n \kappa \tilde{c}] \). Second, despite of the approximation errors, we have \( \tilde{\beta} - \beta_0 \in \Delta_{\tilde{c}} \). This is in contrast to the analysis that relied on \( \lambda \geq cn \| E_n [\epsilon_i x_i] \|_\infty \) instead, see [6, 3]. We build on the latter to establish \( \ell_1 \)-rate and \( \ell_\infty \)-rate of convergence.

**Theorem 5** (\( \ell_1 \)-rate of convergence). Under condition ASM, if \( \lambda \geq c \Lambda \), for \( c > 1 \) and \( \tilde{c} := (c + 1)/(c - 1) \), then

\[
\| \tilde{\beta} - \beta_0 \|_1 \leq (1 + \tilde{c}) \sqrt{n} \| \tilde{\beta} - \beta_0 \|_{2, n} .
\]

**Theorem 6** (\( \ell_\infty \)-rate of convergence). Under condition ASM, if \( \lambda \geq c \Lambda \), for \( c > 1 \) and \( \tilde{c} := (c + 1)/(c - 1) \), then

\[
\| \tilde{\beta} - \beta_0 \|_\infty \leq \left( 1 + \frac{1}{c} \right) \frac{\lambda \sqrt{\hat{Q}(\beta_0)}}{n} + \left( \frac{\lambda^2}{n^2} + \| E_n [x_i x_i'] - I \|_\infty \right) \| \tilde{\beta} - \beta_0 \|_1 .
\]

Regarding the \( \ell_\infty \)-rate, since we have \( \| \cdot \|_\infty \leq \| \cdot \|_1 \), the result is meaningful for nearly orthogonal designs so that \( \| E_n [x_i x_i'] - I \|_\infty \) is small. In fact, near orthogonality also allows to bound the restricted eigenvalues \( \kappa \tilde{c} \) from below. [6] and [16] have established that if for some \( u \geq 1 \) we have \( \| E_n [x_i x_i'] - I \|_\infty \leq 1/(u(1 + \tilde{c})s) \) then \( \kappa \tilde{c} \geq \sqrt{1 - 1/u} \).

We close this subsection establishing relative finite sample bound on the estimation of \( \hat{Q}(\beta_0) \) based on \( \hat{Q}(\tilde{\beta}) \) under the assumptions of Theorem 4.

**Theorem 7** (Relative Estimation of \( \hat{Q}(\beta_0) \)). Under condition ASM, if \( \lambda \geq c \Lambda \) and \( \lambda \sqrt{s} < n \kappa \tilde{c} \), for \( c > 1 \) and \( \tilde{c} := (c + 1)/(c - 1) \), we have

\[
- \frac{4 \tilde{c}}{c} \left( \frac{\lambda \sqrt{s}}{n \kappa \tilde{c}} \right)^2 \sqrt{\hat{Q}(\beta_0)} \leq \sqrt{\hat{Q}(\beta_0)} - \sqrt{\hat{Q}(\hat{\beta})} \leq 2 \left( 1 + \frac{1}{c} \right) \frac{\left( \frac{\lambda \sqrt{s}}{n \kappa \tilde{c}} \right)^2}{1 - \left( \frac{\lambda \sqrt{s}}{n \kappa \tilde{c}} \right)^2} \sqrt{\hat{Q}(\beta_0)} .
\]

Thus, if in addition \( \lambda \sqrt{s} = o(n \kappa \tilde{c}) \) holds, Theorem 7 establishes that

\[
\sqrt{\hat{Q}(\beta_0)} = (1 + o(1)) \sqrt{\hat{Q}(\beta_0)} .
\]

The quantity \( \sqrt{\hat{Q}(\tilde{\beta})} \) is particularly relevant for \( \sqrt{\text{LASSO}} \) since it appears in the first order condition which is the key to establish sparsity properties.
4.2. Finite-Sample Bounds Relating Sparsity and Prediction Norm. In this section we investigate sparsity properties and lower bounds on the rate of convergence in the prediction norm of the $\sqrt{\text{LASSO}}$ estimator. It turns out these results are connected via the first order condition. We start with a technical lemma.

**Lemma 1** (Relating Sparsity and Prediction Norm). Under condition ASM, let $\hat{T} = \text{supp}(\hat{\beta})$ and $\hat{m} = |\hat{T} \setminus T|$. For any $\lambda > 0$ we have

$$
\frac{\lambda}{n} \sqrt{\hat{Q}(\beta)} \sqrt{|\hat{T}|} \leq \sqrt{|\hat{T}|} \|E_n[x_i(\sigma \epsilon_i + r_i)]\|_\infty + \sqrt{\phi_{\max}(\hat{m})} \|\hat{\beta} - \beta_0\|_{2,n}.
$$

The proof of the lemma above rely on the optimality conditions which implies that the selected support has binding dual constraints. Intuitively, for any selected component, there is a shrinkage bias which introduces a bound on how close the estimated coefficient can be from the true coefficient. Based on the inequality above and Theorem 7, we establish the following result.

**Theorem 8** (Lower Bound on Prediction Norm). Under condition ASM, if $\lambda \geq c \tilde{\Lambda}$, $\lambda \sqrt{s} < nk_{\epsilon}$, where $c > 1$ and $\tilde{c} := (c + 1)/(c - 1)$, and letting $\hat{T} = \text{supp}(\hat{\beta})$ and $\hat{m} = |\hat{T} \setminus T|$, we have

$$
|| \hat{\beta} - \beta_0 ||_{2,n} \geq \frac{\lambda \sqrt{|\hat{T}|}}{n \sqrt{\phi_{\max}(\hat{m})}} \sqrt{\hat{Q}(\beta_0)} \left( 1 - \frac{1}{c} - \frac{4 \tilde{c}}{c} \frac{\lambda \sqrt{s}}{nk_{\epsilon}} \right)^2.
$$

In the case of LASSO, as derived in [17], the lower bound does not have the term $\sqrt{\hat{Q}(\beta_0)}$ since the impact of the scaling parameter $\sigma$ is accounted in the penalty level $\lambda$. Thus, under condition ASM, the lower bounds for LASSO and $\sqrt{\text{LASSO}}$ are very close.

Next we proceed to bound the size of the selected support $\hat{T} = \text{supp}(\hat{\beta})$ for the $\sqrt{\text{LASSO}}$ estimator relative to the size $s$ of the support of the oracle estimator $\beta_0$.

**Theorem 9** (Sparsity bound for $\sqrt{\text{LASSO}}$). Under condition ASM, let $\hat{\beta}$ denote the $\sqrt{\text{LASSO}}$ estimator, $\hat{T} = \text{supp}(\hat{\beta})$, and let $\hat{m} := |\hat{T} \setminus T|$. If $\lambda \geq c \tilde{\Lambda}$ and $\lambda \sqrt{s} \leq nk_{\epsilon} \rho$, where $2 \rho^2 \leq (c - 1)/(c - 1 + 4 \tilde{c})$, for $c > 1$ and $\tilde{c} = (c + 1)/(c - 1)$, we have that

$$
\hat{m} \leq s \left( \min_{m \in \mathcal{M}} \phi_{\max}(m) \right) \cdot \left(4 \tilde{c} / \kappa_{\epsilon}\right)^2
$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > s \phi_{\max}(m) \cdot 2(4 \tilde{c} / \kappa_{\epsilon})^2\}$. 
The slightly more stringent side condition ensures that the right hand side of the bound in Theorem 8 is positive. Asymptotically, under mild conditions on the design matrix, for example
\[
\frac{\phi_{\max}(s \log n)}{\phi_{\min}(s \log n)} \lesssim 1,
\]
the event (3.14) and the side condition \( s \log(p/\alpha) = o(n) \), imply that for \( n \) large enough, the size of the selected model is of the same order of magnitude as the oracle model, namely
\[
\hat{m} \lesssim s.
\]

4.3. Finite Sample Bounds on the Estimation Error of OLS post \( \sqrt{\text{LASSO}} \).

Based on the model selected by \( \sqrt{\text{LASSO}} \) estimator, \( \hat{T} := \text{supp}(\hat{\beta}) \), we consider the OLS estimator restricted to these data-driven selected components. If model selection works perfectly (as it will under some rather stringent conditions), then this estimator is simply the oracle estimator and its properties are well known. However, of more interest is the case when model selection does not work perfectly, as occurs for many designs of interest in applications.

The following theorem establishes bounds on the prediction error of the OLS post \( \sqrt{\text{LASSO}} \) estimator. The analysis accounts for the data-driven choice of components and for the possibly having a mispecified model (i.e. \( T \not\subseteq \hat{T} \)). The analysis build upon sparsity and rate bounds of the \( \sqrt{\text{LASSO}} \) estimator, and on a data-dependent empirical process inequality.

**Theorem 10** (Performance of OLS post \( \sqrt{\text{LASSO}} \)). Suppose condition ASM holds, let \( c > 1 \), \( \bar{c} = (c + 1)/(c - 1) \) and \( \hat{m} = |\hat{T} \setminus T| \). If \( \lambda \geq c\Lambda \) occurs with probability at least \( 1 - \alpha \), and \( \lambda \sqrt{s} \leq n\kappa_1 \rho_1 \), for some \( \rho_1 < 1 \), then for \( C \geq 1 \), with probability at least \( 1 - \alpha - 1/C^2 - 1/[9C^2 \log p] \), we have
\[
\|\hat{\beta} - \beta_0\|_{2,n} \leq \frac{C\sigma}{\sqrt{\phi_{\min}(\hat{m})}} \sqrt{\frac{s}{n}} + 2c_s + \sqrt{\frac{\hat{m} \log p}{n}} \left( \frac{24C\sigma}{\sqrt{\phi_{\min}(\hat{m})}} + \sqrt{\frac{1}{n} \max_{j=1,...,p} \mathbb{E}_n \left[ x_{ij}^2 \epsilon_i^2 \right]} \right) + \lambda \sqrt{s} \frac{n\kappa_1}{\rho_1} \\mathbb{Q}(\beta_0) \frac{4(1 + 1/c)}{1 - \rho_1^2} \left( 1 + \frac{(1 + 1/c)\rho_1^2}{1 - \rho_1^2} \right).
\]

We note that the random term in the bound above can be controlled in a variety of ways. For example, under conditions M’ and R’, if \( \log p = o(n^{1/3}) \) Lemma 10 establishes that
\[
\max_{j=1,...,p} \mathbb{E}_n \left[ x_{ij}^2 \epsilon_i^2 \right] = 1 + o_P(1).
\]
\textbf{Corollary 2} (Asymptotic Performance of OLS post $\sqrt{\text{LASSO}}$). Suppose conditions ASM, $M'$ and $R'$ hold, let $c > 1$, $ar{c} = (c + 1)/(c - 1)$ and $\hat{m} = |\hat{T} \setminus T|$. If $\lambda$ is set as the exact or asymptotic choice in (3.17), $\alpha = o(1)$, $\phi_{\text{max}}(s \log n)/\phi_{\text{min}}(s \log n) \lesssim 1$, $s \log (p/\alpha) = o(n)$, $\log p = o(n^{1/3})$, we have that

$$\|\hat{\beta} - \beta_0\|_{2,n} \lesssim_P c_s + \sigma \sqrt{\frac{s \log p}{n}}.$$ 

Moreover, if $\hat{m} = o(s)$ and $T \subseteq \hat{T}$ with probability going to 1,

$$\|\hat{\beta} - \beta_0\|_{2,n} \lesssim_P c_s + \sigma \sqrt{\frac{o(s) \log p}{n}} + \sigma \sqrt{\frac{s}{n}},$$

and if $T = \hat{T}$ with probability going to 1, we have

$$\|\hat{\beta} - \beta_0\|_{2,n} \lesssim_P c_s + \sigma \sqrt{\frac{s}{n}}.$$

Under the conditions of the corollary above, the upper bounds on the rates of convergence of $\sqrt{\text{LASSO}}$ and OLS post $\sqrt{\text{LASSO}}$ coincide. This occurs despite the fact that $\sqrt{\text{LASSO}}$ may in general fail to correctly select the oracle model $T$ as a subset, that is $T \not\subseteq \hat{T}$. Nonetheless, there is a class of well-behaved models in which OLS post $\sqrt{\text{LASSO}}$ rate improves upon the rate achieved by $\sqrt{\text{LASSO}}$. More specifically, this occurs if $\hat{m} = o_P(s)$ and $T \subseteq \hat{T}$ with probability going to 1 or in the case of perfect model selection, when $T = \hat{T}$ with probability going to 1. Results on LASSO’s model selection performance derived on Wainright [38], Candès and Plan [10], and Belloni and Chernozhukov [3] can be extended to the $\sqrt{\text{LASSO}}$ based on Theorem 6 and 7. Moreover, under mild conditions, we know from Theorem 8 that the upper bounds for the rates found for $\sqrt{\text{LASSO}}$ are sharp, i.e. the rate of convergence cannot be faster than $\sigma \sqrt{\log p \log s/n}$. Thus the use of the post model selection estimator leads to a strict improvement in the rate of convergence on these well-behaved models.

4.4. Extreme Cases: Parametric Noiseless and Nonparametric Unbounded Variance. In this section we show that the robustness advantage of $\sqrt{\text{LASSO}}$ with respect to LASSO extends to two extreme cases as well: $\sigma + c_s = 0$ and $E[\epsilon_i^2] = \infty$. Since the the traditional choice of the penalty level $\lambda$ for LASSO depends on $\sigma \sqrt{E[\epsilon_i^2]}$, setting
\[
\lambda = \sigma \sqrt{\mathbb{E}[\epsilon_i^2]} 2c \sqrt{n \Phi^{-1}(1 - \alpha/2p)}
\] cannot be directly applied in either case. In contrast, the penalty of LASSO is indirectly self normalized by the factor \(\sqrt{\hat{Q}(\hat{\beta})}\) which is well defined in both cases.

4.4.1. Parametric Noiseless Case. The analysis developed in the previous section immediately covers the case \(\sigma = 0\) if \(c_s > 0\). The case that \(c_s = 0\) is also zero, thus \(\hat{Q}(\beta_0) = 0\), allows for exact recovery under less stringent restrictions.

**Theorem 11** (Exact Recovery under Parametric Noiseless Case). *Under condition ASM, let \(\sigma = 0\) and \(c_s = 0\). Suppose that \(\lambda > 0\) obeys the growth restriction \(\lambda \sqrt{s} < nk_{1}\). Then we have that \(\hat{\beta} = \beta_0\).*

It is worth mentioning that for any \(\lambda > 0\), unless \(\beta_0 = 0\), LASSO cannot achieve exact recovery. Moreover, it is not obvious how to properly set the penalty level for LASSO even if we knew a priori that it is a parametric noiseless model. In contrast, square-root lasso intrinsically adjusts the penalty \(\lambda\) by a factor of \(\sqrt{\hat{Q}(\hat{\beta})}\). Under mild conditions Theorem 7 ensures that \(\sqrt{\hat{Q}(\hat{\beta})} = \sqrt{\hat{Q}(\beta_0)} = 0\) which allows for the perfect recovery. Also note that the lower bound derived in Theorem 8 becomes trivially zero.

4.4.2. Nonparametric Unbounded Variance. Next we turn to the unbounded variance case. Although the analysis of the prediction norm of the estimator goes through with no change, one needs to redefine the model and oracle properly. Importantly, the exact choice of \(\lambda\) is still valid but the asymptotic penalty choice of \(\lambda\) no longer applies. In order to account for unbounded variance, the main insight is to interpret \(\sigma\) as a scaling factor (and not the standard deviation), but the effective standard deviation \(\hat{\sigma}\) to be

\[
\hat{\sigma} := \sigma \sqrt{\mathbb{E}_n[\epsilon_i^2]} < \infty.
\]

That allows to properly define the oracle in finite sample. However, it follows that \(\hat{\sigma} \to \infty\) with probability one since the noise has infinite variance. Therefore, meaningful estimation of \(\beta_0\) is still possible provided that \(\hat{\sigma}\) does not diverge too quickly. In this case, the oracle estimator needs to be redefined as any any solution to

\[
\min_{0 \leq k \leq p \wedge n} \min_{\|\beta\|_0 = k} \mathbb{E}_n[(f_i - x_i'\beta)^2] + \hat{\sigma}^2 \frac{k}{n}.
\] (4.18)
Regarding the choice of the penalty parameter $\lambda$, we stress that the asymptotic option no longer applies and we cannot achieve Gaussian-like rates. However, the exact option for $\lambda$ is still valid. In fact, the scaling parameter $\sigma$ still cancels out. By inspection of the proof, Lemma 7 can be adjusted to yield $|r_i| \leq \hat{\sigma}/\sqrt{n}$. That motivates the definition of $\Lambda$ as

$$\hat{\Lambda} = n \left\| \frac{\mathbb{E}_n[\epsilon_i x_i]}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} \right\|_\infty + \sqrt{n} \geq \left\| \frac{\mathbb{E}_n[(\sigma \epsilon_i + r_i)x_i]}{\sqrt{\mathbb{E}_n[\sigma^2]} \right\|_\infty.$$ 

Thus, the rate of convergence will be affected by how fast $\mathbb{E}_n[\epsilon_i^2]$ diverges and $\lambda/n$ goes to zero. That is, the final rates will depend on the particular tail properties of the distribution of the noise. The next lemma establishes finite sample bounds in the case of $\epsilon_i \sim t(2), i = 1, \ldots, n$.

**Theorem 12** ($\sqrt{\text{LASSO}}$ prediction norm for $\epsilon_i \sim t(2)$). Consider a nonparametric regression model with data $\{(y_i, x_i) : i = 1, \ldots, n\}$, such that $\mathbb{E}_n[x_{ij}^2] = 1$ for every $j = 1, \ldots, p$, $\epsilon_i \sim t(2)$ are i.i.d., and $\beta_0$ defined as any solution to (4.18). Letting $\bar{x} = \max_{1 \leq j \leq p, 1 \leq i \leq n} |x_{ij}|$ and $\lambda = c(1 + u_n)\Lambda(1 - \alpha - \frac{64}{n^2} \log(4\sqrt{n}/\alpha))$, and $\lambda \sqrt{s} < n\kappa_c$. Then, for any $\tau \in (0, 1)$, with probability at least $1 - \alpha - \tau - \frac{48(1+u_n)\log(4n/\tau)}{nu_n \log n}$ we have

$$\|\hat{\beta} - \beta_0\|_{2,n} \leq \frac{2(1 + 1/c)}{1 - (\lambda \sqrt{s}/[n\kappa_c])^2} \left( c_s + \sigma \sqrt{\log(4n/\tau)} + 2\sqrt{2}/\tau \right) \frac{\lambda \sqrt{s}}{n\kappa_c}$$

where

$$\lambda \leq c(1 + u_n) \left( \frac{4\bar{x} \sqrt{2n \log(4/p/\alpha)} \sqrt{\log(4n/\alpha)} + 2\sqrt{2}/\alpha}{\sqrt{\log n} / \sqrt{24}} + \sqrt{n} \right).$$

Asymptotically, if $1/\alpha = o(\log n)$ and $s \log(p/\alpha) = o(n\kappa_c)$, the result above yields that with probability $1 - \alpha(1 + o(1))$

$$\|\hat{\beta} - \beta_0\|_{2,n} \leq \bar{x}(c_s + \sigma \sqrt{\log n}) \sqrt{\frac{s \log p}{n}}$$

where the scaling factor $\sigma < \infty$ is fixed. Thus, despite of the infinite variance of the noise in the $t(2)$ case, for bounded designs, $\sqrt{\text{LASSO}}$ rate of convergence differs from the Gaussian case only by a $\sqrt{\log n}$ factor.

5. **Empirical Performance of $\sqrt{\text{LASSO}}$ Relative to LASSO**

Next we proceed to evaluate the empirical performance of $\sqrt{\text{LASSO}}$ relative to LASSO. In particular we discuss their computational burden and their estimation performance.
5.1. Computational Performance of $\sqrt{\text{LASSO}}$ Relative to LASSO. Since model selection is particularly relevant in high-dimensional problems, the computational tractability of the optimization problem associated with $\sqrt{\text{LASSO}}$ is an important issue. It will follow that the optimization problem associated with $\sqrt{\text{LASSO}}$ can be cast as a tractable conic programming problem. Conic programming consists of the following optimization problem

$$\min_{x} \; c(x)$$

$$A(x) = b$$

$$x \in K$$

where $K$ is a cone, $c$ is a linear functional, $A$ is a linear operator, and $b$ is an element in the counter domain of $A$. We are particularly interested on the case that $K$ is also convex. Convex conic programming problems have greatly extended the scope of applications of linear programming problems in several fields including optimal control, learning theory, eigenvalue optimization, combinatorial optimization and many others. Under mild regularities conditions, the duality theory for conic programs have been fully developed and allows for characterization of optimal conditions via dual variables, much like linear programming problems.

In the past two decades, the study of the computational complexity and the developments of efficient computational algorithms for conic programming have played a central role in the optimization community. In particular, for the case of self-dual cones, which encompasses the non-negative orthant, second-order cones, and the cone of semi-definite positive matrices, interior point methods have been highly specialize. A sound theoretical foundation, establishing polynomial computational complexity $^{[22, 23]}$, and efficient software implementations $^{[33]}$ made large instances of these problems computational tractable. More recently, first order methods have also been propose to approximately solve even even larger instances of structured conic problem $^{[20, 21, 18]}$.

It follows that (1.3) can be written as a conic programming problem whose relevant cone is self-dual. Letting $Q^{n+1} := \{(t, v) \in \mathbb{R} \times \mathbb{R}^n : t \geq \|v\|\}$ denote the second order cone in

$^{1}$The relevant cone in linear programs is the non-negative orthant, $\min_{w} \{c'w : Aw = b, w \in \mathbb{R}^k\}$.
In $\mathbb{R}^{n+1}$, we can recast (1.3) as the following conic program:

$$
\min_{t,v,\beta^+,\beta^-} \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \sum_{i=1}^{p} \left( \beta^+_j + \beta^-_j \right) \\
v_i = y_i - x_i'\beta^+ + x_i'\beta^-, \; i = 1, \ldots, n \\
(t,v) \in Q^{n+1}, \; \beta^+ \geq 0, \; \beta^- \geq 0.
$$

(5.19)

Conic duality immediately yields the following dual problem

$$
\max_a \mathbb{E}_n [y_i a_i] \\
|\mathbb{E}_n [x_{ij} a_i]| \leq \lambda/n, \; j = 1, \ldots, p \\
\|a\| \leq \sqrt{n}.
$$

(5.20)

From a statistical perspective, the dual variables represent the normalized residuals. Thus the dual problem maximizes the correlation of the dual variable $a$ subject to the constraint that $a$ are approximately uncorrelated with the regressors. It follows that these dual variables play a role in deriving necessary conditions for a component $\hat{\beta}_j$ to be non-zero and therefore on sparsity bounds.

The fact that $\sqrt{\text{LASSO}}$ can be formulated as a convex conic programming problem allows the use of several computational methods tailored for conic problems to be used to compute the $\sqrt{\text{LASSO}}$ estimator. In this section we compare three different methods to compute $\sqrt{\text{LASSO}}$ with their counterparts to compute LASSO. We note that these methods have different initialization and stopping criterion that could impact the running times significantly. Therefore we do not aim to compare different methods but instead we focus on the comparison of the performance of each method to LASSO and $\sqrt{\text{LASSO}}$ since the same initialization and stopping criterion are used.

Table 5.1 illustrates that the average computational time to solve LASSO and $\sqrt{\text{LASSO}}$ optimization problems are comparable. Table 5.1 also reinforces typical behavior of these methods. As the size increases, the running time for interior point methods grows faster than other first order methods. Simple componentwise methods are particular effective when the solution is highly sparse. This is the case of the parametric design used here. We emphasize the performance of each method depends on the particular design and choice of $\lambda$.

5.2. Estimation Performance of $\sqrt{\text{LASSO}}$ Relative to LASSO. In this section we use Monte-Carlo experiments to assess the finite sample performance of the following estimators:
### Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Componentwise</th>
<th>First Order</th>
<th>Interior Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100, p = 500$</td>
<td>LASSO</td>
<td>0.2173</td>
<td>10.99</td>
</tr>
<tr>
<td></td>
<td>\sqrt{\text{LASSO}}</td>
<td>0.3268</td>
<td>7.345</td>
</tr>
<tr>
<td>$n = 200, p = 1000$</td>
<td>LASSO</td>
<td>0.6115</td>
<td>19.84</td>
</tr>
<tr>
<td></td>
<td>\sqrt{\text{LASSO}}</td>
<td>0.6448</td>
<td>19.96</td>
</tr>
<tr>
<td>$n = 400, p = 2000$</td>
<td>LASSO</td>
<td>2.625</td>
<td>84.12</td>
</tr>
<tr>
<td></td>
<td>\sqrt{\text{LASSO}}</td>
<td>2.687</td>
<td>77.65</td>
</tr>
</tbody>
</table>

In these instances we had $s = 5$, $\sigma = 1$, and each value was computed by averaging 100 simulations.

- the (infeasible) LASSO, which knows $\sigma$ (which is unknown outside the experiments),
- OLS post LASSO, which applies OLS to the model selected by (infeasible) LASSO,
- \(\sqrt{\text{LASSO}}\), which does not know $\sigma$, and
- OLS post \(\sqrt{\text{LASSO}}\), which applies OLS to the model selected by \(\sqrt{\text{LASSO}}\).

We set the penalty level for LASSO as the standard choice in the literature, 
\[ \lambda = c^2 \sigma \sqrt{n} \Phi^{-1}(1-\alpha/2p), \]
and \(\sqrt{\text{LASSO}}\) according to the asymptotic option, 
\[ \lambda = c \sqrt{n} \Phi^{-1}(1-\alpha/2p), \]
both with $1 - \alpha = .95$ and $c = 1.1$ to both estimators. (The results obtained using the exact option are similar to the case with the penalty level set according to the asymptotic option, so we only report the results for the latter.)

We use the linear regression model stated in the introduction as a data-generating process, with either standard normal or $t(4)$ errors:

(a) $\epsilon_i \sim N(0, 1)$ or (b) $\epsilon_i \sim t(4)/\sqrt{2}$,

so that $E[\epsilon_i^2] = 1$ in either case. We set the regression function as

\[ f(x_i) = x_i' \beta_0^*, \quad \text{where} \quad \beta_0^* = 1/j^{3/2}, \quad j = 1, \ldots, p. \]  

(5.21)

The scaling parameter $\sigma$ vary between 0.25 and 5. For a fixed design, as the scaling parameter $\sigma$ increases, the number of non-zero components in the oracle vector $s$ decreases. The number of regressors is $p = 500$, the sample size is $n = 100$, and we used 100 simulations for
each design. We generate regressors as $x_i \sim N(0, \Sigma)$ with the Toeplitz correlation matrix
$\Sigma_{jk} = (1/2)^{|j-k|}$.

![Empirical Risk](image1)

**Figure 1.** The average empirical risk of the estimators as a function of the scaling parameter $\sigma$.

![Bias](image2)

**Figure 2.** The norm of the bias of the estimators as a function of the scaling parameter $\sigma$. 
\[ \epsilon_i \sim N(0,1) \quad \epsilon_i \sim t(4) \]

**Figure 3.** The average number of regressors selected as a function of the scaling parameter \( \sigma \).

We present the results of computational experiments for designs a) and b) in Figures 1, 2, 3. The left plot of each figure reports the results for the normal errors, and the right plot of each figure reports the results for \( t(4) \) errors. For each model, the figures show the following quantities as a function of the signal strength \( C \) for each estimator \( \tilde{\beta} \):

- Figure 1 – the average empirical risk, \( E[\mathbb{E}_n[x_i'(\tilde{\beta} - \beta_0)]^2] \),
- Figure 2 – the norm of the bias, \( \|E[\tilde{\beta} - \beta_0]\| \), and
- Figure 3 – the average number of regressors selected, \( E|\text{support}(\tilde{\beta})| \).

Figure 1, left panel, shows the empirical risk for the Gaussian case. We see that, for a wide range of the scaling parameter \( \sigma \), LASSO and \( \sqrt{\text{LASSO}} \) perform similarly in terms of empirical risk, although standard LASSO outperforms somewhat \( \sqrt{\text{LASSO}} \). At the same time, OLS post LASSO outperforms slightly OLS post \( \sqrt{\text{LASSO}} \) for larger signal strengths. This is expected since \( \sqrt{\text{LASSO}} \) over regularize to simultaneously estimate \( \sigma \) when compared to LASSO (since it essentially uses \( \sqrt{\hat{Q}(\tilde{\beta})} \) as an estimate of \( \sigma \)). In the nonparametric model considered here, the coefficients are not well separated from zero. These two issues combined leads to a smaller selected support.
Overall, the empirical performance of √LASSO and OLS post √LASSO is achieves its goal. Despite not knowing σ, √LASSO performs comparably to the standard LASSO that knows σ. These results are in close agreement with our theoretical results, which state that the upper bounds on empirical risk for √LASSO asymptotically approach the analogous bounds for standard LASSO.

Figures 2 and 3 provide additional insight into the performance of the estimators. On the one hand, Figure 2 shows that the finite-sample differences in empirical risk for LASSO and √LASSO arise primarily due to √LASSO having a larger bias than standard LASSO. This bias arises because √LASSO uses an effectively heavier penalty. Figure 3 shows that such heavier penalty translates into √LASSO achieving a smaller support than LASSO on average.

Finally, Figure 1, right panel, shows the empirical risk for the t(4) case. We see that the results for the Gaussian case carry over to the t(4) case. In fact, the performance of LASSO and √LASSO under t(4) errors nearly coincides with their performance under Gaussian errors. This is exactly what is predicted by our theoretical results.

**Appendix A. Probability Inequalities**

**Lemma 2 (Rosenthal Inequality).** Let $X_1, \ldots, X_n$ be independent zero-mean random variables, then for $r \geq 2$

$$E \left[ \left| \sum_{i=1}^{n} X_i \right|^r \right] \leq C(r) \max \left\{ \sum_{t=1}^{n} E[|X_i|^r], \left( \sum_{t=1}^{n} E[X_i^2] \right)^{r/2} \right\}.$$

**Corollary 3 (Rosenthal LLN).** Let $r \geq 2$, and consider the case of independent and identically distributed zero-mean variables $X_i$ with $E[X_i^2] = 1$ and $E[|X_i|^r]$ bounded by $C$. Then for any $\ell_n > 0$

$$Pr \left( \frac{1}{n} \sum_{i=1}^{n} X_i > \ell_n n^{-1/2} \right) \leq \frac{2C(r)C}{\ell_n},$$

where $C(r)$ is a constant depend only on $r$. 
Remark. To verify the corollary, note that by Rosenthal’s inequality $E \left| \sum_{i=1}^{n} X_i \right| \leq Cn^{r/2}$. By Markov inequality,

$$P \left( \frac{\sum_{i=1}^{n} X_i}{n} > c \right) \leq \frac{C(r)Cn^{r/2}}{c^r n^r} \leq \frac{C(r)C}{c^r n^r},$$

so the corollary follows. We refer [25] for complete proofs.

**Lemma 3** *(Vonbahr-Esseen inequality)*. Let $X_1, \ldots, X_n$ be independent zero-mean random variables. Then for $1 \leq r \leq 2$

$$E \left[ \left| \sum_{i=1}^{n} X_i \right|^r \right] \leq (2 - n^{-1}) \cdot \sum_{k=1}^{n} E[|X_k|^r].$$

We refer to [36] for proofs.

**Corollary 4** *(Vonbahr-Esseen’s LLN)*. Let $r \in [1, 2]$, and consider the case of identically distributed zero-mean variables $X_i$ with $E|X_i|^r$ bounded by $C$. Then for any $\ell_n > 0$

$$P \left( \frac{\sum_{i=1}^{n} X_i}{n} > \ell_n n^{-(1-1/r)} \right) \leq \frac{2C}{\ell_n^r}.$$  

Remark. By Markov and Vonbahr-Esseen’s inequalities,

$$P \left( \frac{\sum_{i=1}^{n} X_i}{n} > c \right) \leq \frac{E\left[ \left( \sum_{i=1}^{n} X_i \right)^r \right]}{c^r n^r} \leq \frac{(2n - 1)E[|X_i|^r]}{c^r n^r} \leq \frac{2C}{c^r n^r-1},$$

which implies the corollary.


Next we consider Slătquistov-Rubin-Sethuraman Moderate Deviation Theorem.

Let $X_{ni}, i = 1, \ldots, k_n; n \geq 1$ be a double sequence of row-wise independent random variables with $E[X_{ni}] = 0$, $E[X_{ni}^2] < \infty$, $i = 1, \ldots, k_n; n \geq 1$, and $B_n^2 = \sum_{i=1}^{k_n} E[X_{ni}^2] \to \infty$ as $n \to \infty$. Let

$$F_n(x) = P \left( \frac{\sum_{i=1}^{k_n} X_{ni}}{B_n} < x \right).$$

**Lemma 4** *(Slătquistov, Theorem 1.1)*. If for sufficiently large $n$ and some positive constant $c$,

$$\sum_{i=1}^{k_n} E[|X_{ni}|^{2+c}] \rho(|X_{ni}|) \log^{-1+c^2/2} (3 + |X_{ni}|) \leq g(B_n)B_n^2,$$
where $\rho(t)$ is slowly varying function monotonically growing to infinity and $g(t) = o(\rho(t))$ as $t \to \infty$, then

$$1 - F_n(x) \sim 1 - \Phi(x), F_n(-x) \sim \Phi(-x), \quad n \to \infty,$$

uniformly in the region $0 \leq x \leq c\sqrt{\log B_n^2}$

**Corollary 5** (Slastnikov, Rubin-Sethuraman). If $q > c^2 + 2$ and

$$\sum_{i=1}^{\varphi_n} \mathbb{E}[|X_{ni}|^q] \leq KB_n^2,$$

then there is a sequence $\gamma_n \to 1$, such that

$$\left| \frac{1 - F_n(x) + F_n(-x)}{2\Phi(x)} - 1 \right| \leq \gamma_n - 1 \to 0, \quad n \to \infty,$$

uniformly in the region $0 \leq x \leq c\sqrt{\log B_n^2}$

Remark. Rubin-Sethuraman derived the corollary for $x = t\sqrt{\log B_n^2}$ for fixed $t$. Slastnikov’s result adds uniformity and relaxes the moment assumption.

We refer to [28] for proofs.

### A.3. Moderate Deviations for Self-Normalized Sums

We shall be using the following result – Theorem 7.4 in [12].

Let $X_1, \ldots, X_n$ be independent, mean-zero variables, and

$$S_n = \sum_{i=1}^{n} X_i, \quad V_n^2 = \sum_{i=1}^{n} X_i^2.$$

For $0 < \delta \leq 1$ set

$$B_n^2 = \sum_{i=1}^{n} \mathbb{E}X_i^2, \quad L_n, \delta = \sum_{i=1}^{n} \mathbb{E}|X_i|^{2+\delta}, \quad d_n, \delta = B_n/L_n^{1/(2+\delta)}.$$

Then for uniformly in $0 \leq x \leq d_n, \delta$,

$$\frac{\Pr(S_n/V_n \geq x)}{\Phi(x)} = 1 + O(1) \left( \frac{1 + x}{d_n, \delta} \right)^{2+\delta},$$

$$\frac{\Pr(S_n/V_n \leq -x)}{\Phi(-x)} = 1 + O(1) \left( \frac{1 + x}{d_n, \delta} \right)^{2+\delta},$$

where the terms $O(1)$ are bounded in absolute value by a universal constant $A$, and $\Phi := 1 - \Phi.$
Application of this result gives the following lemma:

**Lemma 5** (Moderate Deviations for Self-Normalized Sums). Let $X_{1,n},...,X_{n,n}$ be the triangular array of i.i.d, zero-mean random variables. Suppose that

$$M_n = \frac{(EX_{1,n}^2)^{1/2}}{(E|X_{1,n}|^3)^{1/3}} > 0$$

and that for some $\ell_n \to \infty$

$$n^{1/6}M_n/\ell_n \geq 1.$$ 

Then uniformly on $0 \leq x \leq n^{1/6}M_n/\ell_n - 1$, the quantities

$$S_{n,n} = \sum_{i=1}^{n} X_{i,n}, \quad V_{n,n}^2 = \sum_{i=1}^{n} X_{i,n}^2.$$

obey

$$\left| \frac{\Pr(|S_{n,n}/V_{n,n}| \geq x)}{2\Phi(x)} - 1 \right| \leq \frac{A}{\ell_n^3} \to 0.$$

Proof. This follows by the application of the quoted theorem to the i.i.d. case with $\delta = 1$ and $d_{n,1} = n^{1/6}M_n$. The calculated error bound follows from the triangular inequalities and conditions on $\ell_n$ and $M_n$. \qed

**A.4. Data-dependent Probabilistic Inequality.** In this section we derive a data-dependent probability inequality for empirical processes indexed by a finite class of functions. In what follows, for a random variable $X$ let $q(X,1-\tau)$ denote its $(1-\tau)$-quantile. For a class of functions $\mathcal{F}$ we define $\|X\|_\mathcal{F} = \sup_{f \in \mathcal{F}} |f(X)|$. Also for random variables $Z_1,\ldots,Z_n$ and a function $f$ define $\|f\|_{P_n,2} = \sqrt{E_n[f(Z_i)^2]}$, $G_n(f) = (1/\sqrt{n}) \sum_{i=1}^{n} \{f(Z_i) - E[f(Z_i)]\}$, and $G_n^\circ(f) = (1/\sqrt{n}) \sum_{i=1}^{n} \xi_i f(Z_i)$ where $\xi_i$ are independent Rademacher random variables.

In order to prove a bound on tail probabilities of a general separable empirical process, we need to go through a symmetrization argument. Since we use a data-dependent threshold, we need an appropriate extension of the classical symmetrization lemma to allow for this. Let us call a threshold function $x : \mathbb{R}^n \to \mathbb{R}$ $k$-sub-exchangeable if for any $v,w \in \mathbb{R}^n$ and any vectors $\tilde{v},\tilde{w}$ created by the pairwise exchange of the components in $v$ with components in $w$, we have that $x(\tilde{v}) \vee x(\tilde{w}) \geq [x(v) \vee x(w)]/k$. Several functions satisfy this property, in particular $x(v) = \|v\|$ with $k = \sqrt{2}$, constant functions with $k = 1$, and $x(v) = \|v\|_\infty$ with $k = 1$. The following result generalizes the standard symmetrization lemma for probabilities.
Lemma 6 (Symmetrization with Data-dependent Thresholds). Consider arbitrary independent stochastic processes $Z_1, \ldots, Z_n$ and arbitrary functions $\mu_1, \ldots, \mu_n : \mathcal{F} \mapsto \mathbb{R}$. Let $x(Z) = x(Z_1, \ldots, Z_n)$ be a $k$-sub-exchangeable random variable and for any $\tau \in (0, 1)$ let $q_\tau$ denote the $\tau$ quantile of $x(Z)$, $\bar{p}_\tau := P(x(Z) \leq q_\tau) \geq \tau$, and $p_\tau := P(x(Z) < q_\tau) \leq \tau$. Then

$$
P \left( \left\| \sum_{i=1}^{n} Z_i \right\|_{\mathcal{F}} > x_0 \vee x(Z) \right) \leq \frac{4}{\bar{p}_\tau} P \left( \left\| \sum_{i=1}^{n} \varepsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z)}{4k} \right) + p_\tau$$

where $x_0$ is a constant such that $\inf_{f \in \mathcal{F}} P \left( \left| \sum_{i=1}^{n} Z_i(f) \right| \leq \frac{2\bar{p}_\tau}{\tau} \right) \geq 1 - \frac{\bar{p}_\tau}{2}$.

Theorem 13 (Maximal Inequality for Empirical Processes). Let

$$q_D(\mathcal{F}, 1 - \tau) = \sup_{f \in \mathcal{F}} q(\left| \mathbb{G}_n(f) \right|, 1 - \tau) \leq \sup_{f \in \mathcal{F}} \sqrt{\text{var} P(\mathbb{G}_n(f))/\tau}$$

and consider the data dependent quantity

$$e_n(\mathcal{F}, \mathbb{P}_n) = \sqrt{2 \log |\mathcal{F}|} \sup_{f \in \mathcal{F}} \|f\|_{\mathbb{P}_n, 2}.$$

Then, for any $C \geq 1$ and $\tau \in (0, 1)$ we have

$$\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| \leq q_D(\mathcal{F}, 1 - \tau/2) \vee 4\sqrt{2}Ce_n(\mathcal{F}, \mathbb{P}_n),$$

with probability at least $1 - \tau - 4 \exp(-(C^2 - 1) \log |\mathcal{F}|)/\tau$.

Proof. Step 1. (Main Step) In this step we prove the main result. First, recall $e_n(\mathcal{F}, \mathbb{P}_n) := \sqrt{2 \log |\mathcal{F}|} \sup_{f \in \mathcal{F}} \|f\|_{\mathbb{P}_n, 2}$. Note that $\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{P}_n, 2}$ is $\sqrt{2}$-sub-exchangeable by Step 2 below.

By the symmetrization Lemma 6 we obtain

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| > 4\sqrt{2}Ce_n(\mathcal{F}, \mathbb{P}_n) \vee q_D(\mathcal{F}, 1 - \tau/2) \right\} \leq \frac{4}{\tau} \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |\mathbb{G}_n^o(f)| > Ce_n(\mathcal{F}, \mathbb{P}_n) \right\} + \tau.$$

Thus a union bound yields

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| > 4\sqrt{2}Ce_n(\mathcal{F}, \mathbb{P}_n) \vee q_D(\mathcal{F}, 1 - \tau/2) \right\} \leq \frac{4|\mathcal{F}|}{\tau} \sup_{f \in \mathcal{F}} \mathbb{P} \left\{ |\mathbb{G}_n^o(f)| > Ce_n(\mathcal{F}, \mathbb{P}_n) \right\}.$$

(A.22)
We then condition on the values of $Z_1, \ldots, Z_n$, denoting the conditional probability measure as $\mathbb{P}_\varepsilon$. Conditional on $Z_1, \ldots, Z_n$, by the Hoeffding inequality the symmetrized process $G_n^\varepsilon$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$ norm, namely, for $f \in F$, $\mathbb{P}_\varepsilon\{|G_n^\varepsilon(f)| > x\} \leq 2\exp(-x^2/[2\|f\|_{F_n,2}^2])$. Hence, we can bound

$$\mathbb{P}_\varepsilon\{|G_n^\varepsilon(f)| \geq C_n(f, \mathbb{P}_n)|Z_1, \ldots, Z_n\} \leq 2\exp(-C^2_\varepsilon n(f, \mathbb{P}_n)^2/[2\|f\|_{F_n,2}^2]) \leq 2\exp(-C^2\log |F|).$$

Taking the expectation over $Z_1, \ldots, Z_n$ does not affect the right hand side bound. Plugging in this bound into relation (A.23) yields the result.

**Step 2. (Auxiliary calculations.)** To establish that $\sup_{f \in F} \|f\|_{F_n,2}$ is $\sqrt{2}$-sub-exchangeable, let $\tilde{Z}$ and $\tilde{Y}$ be created by exchanging any components in $Z$ with corresponding components in $Y$. Then

$$\sqrt{2}(\sup_{f \in F} \|f\|_{\mathbb{P}_n(\tilde{Z})}^2 \lor \sup_{f \in F} \|f\|_{\mathbb{P}_n(\tilde{Y})}^2) \geq (\sup_{f \in F} \|f\|_{\mathbb{P}_n(\tilde{Z})}^2 + \sup_{f \in F} \|f\|_{\mathbb{P}_n(\tilde{Y})}^2)^{1/2} \geq (\sup_{f \in F} \mathbb{E}_n[f(\tilde{Z})^2] + \mathbb{E}_n[f(\tilde{Y})^2])^{1/2} \geq (\sup_{f \in F} \mathbb{E}_n[f(Z_i)^2] + \mathbb{E}_n[f(Y_i)^2])^{1/2} \geq (\sup_{f \in F} \|f\|_{\mathbb{P}_n(Z)}^2 \lor \sup_{f \in F} \|f\|_{\mathbb{P}_n(Y)}^2)^{1/2} = \sup_{f \in F} \|f\|_{\mathbb{P}_n(Z)}^2 \lor \sup_{f \in F} \|f\|_{\mathbb{P}_n(Y)}^2.$$

\[\Box\]

**Corollary 6** (Data-Dependent Probability Inequality). Let $\varepsilon_i$ be i.i.d random variables such that $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ for $i = 1, \ldots, n$. Conditional on $x_1, \ldots, x_n \in \mathbb{R}^p$, we have that for any $C \geq 1$, with probability at least $1 - 1/[9C^2\log p]$,

$$\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty \leq C \cdot 24 \sqrt{\frac{\log p}{n}} \max_{j=1, \ldots, p} \sqrt{\mathbb{E}_n[\varepsilon_i^2x_{ij}^2]} \lor \sqrt{\sigma^2\mathbb{E}_n[x_{ij}^2]}.$$

**Proof of Corollary 6** Consider the class of separable empirical process induced by $\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty$, i.e. the class of functions $f \in F = \{\varepsilon_ix_{ij} : j \in \{1, \ldots, p\}\}$ so that $\sqrt{n}\|\mathbb{E}_n[x_i\varepsilon_i]\|_\infty = \sup_{f \in F} |G_n(f)|$. Define the data dependent quantity

$$e_n(F, \mathbb{P}_n) = \sqrt{2\log p} \max_{j=1, \ldots, p} \sqrt{\mathbb{E}_n[\varepsilon_i^2x_{ij}^2]}.$$

Then, by Theorem 13 for any constant $C \geq 1$

$$\sup_{f \in F} |G_n(f)| \leq q(F, 1 - \tau/2) \lor 4\sqrt{2}C e_n(F, \mathbb{P}_n).$$
with probability $1 - \tau - 4 \exp(-(C^2 - 1) \log p)/\tau$. Picking $\tau = 1/[2C^2 \log p]$, we have by the Chebyshev’s inequality

$$q(\mathcal{F}, 1 - \tau/2) \leq \max_{j=1,\ldots,p} \sqrt{\mathbb{E}[\epsilon^2_{ij}]/\tau} = 2C \sqrt{\log p} \max_{j=1,\ldots,p} \sqrt{\mathbb{E}[\epsilon^2_{ij}]/\tau}.$$ 

Setting $C \geq 3$ we have with probability $1 - 1/[C^2 \log p]$

$$\sup_{f \in \mathcal{F}} |G_n(f)| \leq \left(\frac{6C}{3} \sqrt{\log p} \max_{j=1,\ldots,p} \sqrt{\mathbb{E}[\epsilon^2_{ij}]/\tau} \right) \wedge \left(\frac{24C}{3} \sqrt{\log p} \max_{j=1,\ldots,p} \sqrt{\mathbb{E}[\epsilon^2_{ij}]}/\tau \right).$$

(Note that if $p \leq 2$ the statement is trivial since the probability is greater than 1.)

A.5. Bounds via Symmetrization. Next we proceed to use symmetrization arguments to bound the empirical process. Let $\|f\|_{\mathbb{P}_n,2} = \sqrt{\mathbb{E}_n[f(X_i)^2]}$, $G_n(f) = \sqrt{n} \mathbb{E}_n[f(X_i) - \mathbb{E}[X_i]]$, and for a random variable $Z$ let $q(Z, 1 - \tau)$ denote its $(1 - \tau)$-quantile.

**Theorem 14** (Maximal Inequality via Symmetrization). Let

$$q_S(\mathcal{F}, 1 - \tau) = q(\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{P}_n,2}, 1 - \tau).$$

Then, for any $\tau \in (0,1)$, and $\delta \in (0,1)$ we have

$$\sup_{f \in \mathcal{F}} |G_n(f)| \leq 4 \sqrt{2 \log(|\mathcal{F}|)/\delta} q_S(\mathcal{F}, 1 - \tau),$$

with probability at least $1 - \tau - \delta$.

**Proof.** Let $e_n(\mathcal{F}) = \sqrt{\log(|\mathcal{F}|)/\delta} q_S(\mathcal{F}, 1 - \tau)$ and the event $\mathcal{E} = \{\sup_{f \in \mathcal{F}} \sqrt{n} \mathbb{E}_n[f^2] \leq q_S(\mathcal{F}, 1 - \tau)\}$, by definition $P(\mathcal{E}) \geq 1 - \tau$. By the symmetrization Lemma 2.3.7 of [35] we obtain

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |G_n(f)| > 4e_n(\mathcal{F}) \right\} \leq 4 \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |G_n^o(f)| > e_n(\mathcal{F}) \right\} \leq \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |G_n^o(f)| > e_n(\mathcal{F}) | \mathcal{E} \right\} + \tau.$$

Thus a union bound yields

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |G_n(f)| > e_n(\mathcal{F}) \right\} \leq \tau + |\mathcal{F}| \mathbb{P} \left\{ \sup_{f \in \mathcal{F}} |G_n^o(f)| > e_n(\mathcal{F}) | \mathcal{E} \right\}.$$

(A.23)

We then condition on the values of $Z_1, \ldots, Z_n$ and $\mathcal{E}$, denoting the conditional probability measure as $\mathbb{P}_\varepsilon$. Conditional on $Z_1, \ldots, Z_n$, by the Hoeffding inequality the symmetrized
process $G^o_n$ is sub-Gaussian for the $L_2(P_n)$ norm, namely, for $f \in \mathcal{F}$, $P_\varepsilon\{|G^o_n(f)| > x\} \leq 2 \exp(-x^2/\|f\|_{L^2_n}^2)$. Hence, we can bound

$$P_\varepsilon \{|G^o_n(f)| \geq e_n(\mathcal{F})|Z_1, \ldots, Z_n, \mathcal{E}\} \leq 2 \exp(-e_n(\mathcal{F})^2/\|f\|_{L^2_n}^2) \leq 2 \exp(-\log(2|\mathcal{F}|/\delta)).$$

Taking the expectation over $Z_1, \ldots, Z_n$ does not affect the right hand side bound. Plugging in this bound yields the result. \qed

**Appendix B. Technical Lemmas**

We begin with two technical lemmas that control the impact of the approximation errors in the analysis.

**Lemma 7.** Under condition ASM we have

$$\|\tilde{E}_n[x_i r_i]\|_{\infty} \leq \min \left\{ \frac{\sigma}{\sqrt{n}} c_s \right\}.$$

**Proof.** First note that for every $j = 1, \ldots, p$, we have $|\tilde{E}_n[x_{ij} r_i]| \leq \sqrt{\tilde{E}_n[x_{ij}^2] \tilde{E}_n[r_i^2]} = c_s$. Next, by definition of $\beta_0$ in (2.7), for $j \in T$ we have $\tilde{E}_n[x_{ij}(f_i - x'_i \beta_0)] = \tilde{E}_n[x_{ij} r_i] = 0$ since $\beta_0$ is a minimizer over the support of $\mathcal{E}$. For $j \in T^c$ we have that for any $t \in \mathbb{R}$

$$\tilde{E}_n[(f_i - x'_i \beta_0)^2] + \frac{\sigma^2 s}{n} \leq \tilde{E}_n[(f_i - x'_i \beta_0 - t x_{ij})^2] + \frac{\sigma^2 s + 1}{n}.$$

Therefore, for any $t \in \mathbb{R}$ we have

$$-\sigma^2/n \leq \tilde{E}_n[(f_i - x'_i \beta_0 - t x_{ij})^2] - \tilde{E}_n[(f_i - x'_i \beta_0)^2] = -2t \tilde{E}_n[x_{ij}(f_i - x'_i \beta_0)] + t^2 \tilde{E}_n[x_{ij}^2].$$

Taking the minimum over $t$ in the right hand side at $t^* = \tilde{E}_n[x_{ij}(f_i - x'_i \beta_0)]$ we obtain

$$-\sigma^2/n \leq -(\tilde{E}_n[x_{ij}(f_i - x'_i \beta_0)])^2$$

or equivalently, $|\tilde{E}_n[x_{ij}(f_i - x'_i \beta_0)]| \leq \sigma/\sqrt{n}$. \qed

**Lemma 8.** Assume that condition ASM holds and that there is $q > 2$ such that $E[|\varepsilon|^q] < \infty$. Then we have

$$P \left( \sqrt{\tilde{E}_n[\sigma^2 e_i^2]} > (1 + u_n) \sqrt{\tilde{E}_n[\sigma \varepsilon_i + r_i]^2} \right) \leq \min \left\{ \frac{\psi(v)}{v^{q/4}} \right\} \psi(v) + \frac{2(1 + u_n)}{(1 - v)u_n n}.$$

where $\psi(v) := \frac{C_q E[|\varepsilon|^{q+4}]}{\psi_{qq}^{q/4}} \wedge \frac{2E[|\varepsilon|]}{n^{1/(q-2)-1/q} v^{1/2}}.$
Proof. Let \( a_n = [ (1 + u_n)^2 - 1 ]/(1 + u_n)^2 \). We have that
\[
P(\mathbb{E}[\sigma_i^2 | \mathbf{X}] > (1 + u_n)^2 \mathbb{E}[(\sigma_i + r_i)^2] \mid \mathbf{X}) = P(2\mathbb{E}[\epsilon_i r_i] < -c_s^2 - a_n \mathbb{E}[\sigma_i^2 \epsilon_i^2] \mid \mathbf{X}). \tag{B.24}
\]
By Lemma 9 we have
\[
\Pr(\sqrt{\mathbb{E}[\epsilon_i^2]} < 1 - v \mid F_0 = F) \leq \Pr(|\mathbb{E}[\epsilon_i^2] - 1| > v \mid F_0 = F) \leq \psi(v).
\]
Thus,
\[
P(\mathbb{E}[\sigma_i^2 \epsilon_i^2] > (1 + u_n)^2 \mathbb{E}[(\sigma_i + r_i)^2] \mid \mathbf{X}) \leq \psi(v) + P(2\mathbb{E}[\sigma_i r_i] < -c_s^2 - a_n \sigma_i^2 (1 - v) \mid \mathbf{X}).
\]
Since \( \mathbb{E}[(2\mathbb{E}[\sigma_i r_i])^2] = 4\sigma_i^2 c_s^2 / n \), by Chebyshev inequality we have
\[
P \left( \sqrt{\mathbb{E}[\sigma_i^2 \epsilon_i^2]} \leq (1 + u_n) \sqrt{\mathbb{E}[(\sigma_i + r_i)^2]} \mid \mathbf{X} \right) \leq \psi(v) + \frac{4\sigma_i^2 c_s^2 / n}{(c_s^2 + a_n \sigma_i^2 (1 - v))^2} \leq \psi(v) + \frac{4\sigma_i^2 c_s^2 / n}{(1 - v)u_n n}.
\]
The result follows by minimizing over \( v \in (0, 1) \). \( \square \)

Next we focus on controlling the deviations of empirical second moments of the noise.

**Lemma 9.** Assume that condition ASM holds and that there is \( q > 2 \) such that \( E[|\epsilon|^q] < \infty \). Then there is a constant \( C_q \), that depends on \( q \) only, such that for \( v > 0 \) we have
\[
\Pr(|\mathbb{E}[\epsilon_i^2] - 1| > v \mid F_0 = F) \leq \psi(v) := \frac{C_q E[|\epsilon_i|^q/4]}{v^{q/2(1-1/q^2)} / n^{1/(q/2)}}. \]

**Proof.** By the application of either Rosenthal’s inequality \( \text{[26]} \) for the case of \( q > 4 \) or Vonbahr-ESseen’s inequalities \( \text{[37]} \) for the case of \( 2 < q \leq 4 \),
\[
\Pr(|\mathbb{E}[\epsilon_i^2] - 1| > v \mid F_0 = F) \leq \psi(v) := \frac{C_q E[|\epsilon_i|^q/4]}{v^{q/2(1-1/q^2)} / n^{1/(q/2)}}. \]
\( \square \)

**Lemma 10.** Under conditions ASM, \( M' \) and \( R' \), we have that with probability \( 1 - \tau_1 - \tau_2 \) we have
\[
\max_{1 \leq j \leq p} \mathbb{E}_n[\epsilon_i^2 (\epsilon_i^2 - 1)] \leq 4 \sqrt{\frac{\log(2p/\tau_1)}{n}} \left( \frac{E[\epsilon_i]}{\tau_2} \right)^{4/q} \max_{1 \leq j \leq p} \left( \mathbb{E}_n[\epsilon_i^8] \right)^{1/4}.
\]
Lemma 11. Under conditions ASM, M’ and R’, assume that

\[ n^{1/6} \geq \ell_n \max_{1 \leq j \leq p} (E_n[|x_{ij}|^3]E[|\epsilon_i|^3])^{1/3}. \]

For \( \tau \in (0, 1) \), let the constants \( c_1, c_2 > 1 \) satisfy

\begin{enumerate}[(i)]
\item \( c_1 \geq 1 + 4 \sqrt{\frac{\log(\theta/p)}{n}} \left( \frac{3E[\epsilon_i^8]}{\tau} \right)^{4/3} \max_{1 \leq j \leq p} (E_n[x_{ij}^8])^{1/4}; \)
\item \( c_2 \geq 1/(1 - \frac{3c_1^2E[|\epsilon_i|^4]}{\tau^{1/4}n^{1/4}}). \)
\end{enumerate}

Then, there is an universal constant \( A \) such that uniformly in \( t \geq 0 \)

\[ t + 1 \leq n^{1/6}/[\ell_n \max_{1 \leq j \leq p} (E_n[|x_{ij}|^3]E[|\epsilon_i|^3])^{1/3}], \]

we have

\[ P \left( \left\| \frac{\sqrt{n}E_n[x_{ij}\epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} \right\|_\infty > \sqrt{c_1c_2} t \right) \leq 2p \Phi(t) \left( 1 + \frac{A}{\ell_n^3} \right) + \tau. \]

Proof. Let \( \ell = 1/[\ell_n \max_{1 \leq j \leq p} (E_n[|x_{ij}|^3]E[|\epsilon_i|^3])^{1/3}] > 0 \), \( \tau_1 = \tau_2 = \tau_3 = \tau/3. \)

\begin{align*}
P \left( \left\| \frac{\sqrt{n}E_n[x_{ij}\epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} \right\|_\infty > \sqrt{c_1c_2} t \right) & \leq P \left( \max_{1 \leq j \leq p} \frac{\sqrt{n}E_n[x_{ij}\epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} > t \right) + P \left( \max_{1 \leq j \leq p} \frac{E_n[x_{ij}^2]}{E_n[\epsilon_i^2]} > c_1c_2 \right).
\end{align*}

By Lemma 10 above, provided that \( t \leq \ell n^{1/6} - 1 \), we have that there is an universal constant \( A \), such that

\[ P \left( \max_{1 \leq j \leq p} \frac{\sqrt{n}E_n[x_{ij}\epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} > t \right) \leq p \max_{1 \leq j \leq p} P \left( \frac{\sqrt{n}E_n[x_{ij}\epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} > t \right) \leq 2p \Phi(t) \left( 1 + \frac{A}{\ell_n^3} \right). \]

Next note that

\begin{align*}
P \left( \max_{1 \leq j \leq p} \frac{E_n[x_{ij}^2]}{E_n[\epsilon_i^2]} > c_1c_2 \right) & \leq P \left( \max_{1 \leq j \leq p} E_n[x_{ij}^2(\epsilon_i^2 - 1)] > c_1 - 1 \right) + P (E_n[\epsilon_i^2] < 1/c_2).
\end{align*}
By Lemma 9 we have there is a constant $C_q$ such that

$$P \left( \mathbb{E}_n[\epsilon_i^2] < 1/c_2 \right) \leq \frac{C_q \mathbb{E}[|\epsilon_i|^{q\vee 4}]}{(c_2 - 1)/c_2)^{4q/4}} \leq \tau_3$$

where the last inequality follows from the choice of $c_2$.

Finally, by Lemma 10 and the choice of $c_1$, we have

$$P \left( \max_{1 \leq j \leq p} \mathbb{E}_n[x_{ij}^2(\epsilon_i^2 - 1)] > c_1 - 1 \right) \leq \tau_1 + \tau_2.$$ 

Since $\tau_1 + \tau_2 + \tau_3 = \tau$ the result follows. \hfill \Box

**Appendix C. Results under Conditions M and R**

**Condition M.** There exist a finite constant $q > 2$ such that the disturbance obeys $\sup_{n \geq 1} \mathbb{E}_0[|\epsilon|^q] < \infty$, and the design $X$ obeys $\sup_{n \geq 1} \max_{1 \leq j \leq p} \mathbb{E}_n[|x_{ij}|^q] < \infty$.

**Condition R.** We have that $p \leq \alpha n^{q(q-2)/2}/2$ for some constant $0 < \eta < 1$, $c_s^2 \Phi^{-1}(1 - \alpha/2p) \leq \sigma^2$, and $\Phi^{-1}(1 - \alpha/2p)^2 \alpha^{-2} n^{-[(1-2/q)/2]} \log n \leq 1/3$. For convenience we also assume that $\Phi^{-1}(1 - \alpha/2p) \geq 1$.

Conditions M and R were considered in [5] in the parametric setting. The analysis relied on concentration of measure and moderate deviation bounds. Below we provide the results that validate the choice of $\lambda$ in the current nonparametric setting not covered in [5].

**Theorem 15** (Properties of the Penalty Level under Conditions M and R). Suppose that conditions M and R hold, and let $u_n = 2/\Phi^{-1}(1 - \alpha/2p)$. Then, there is a constant $C(q)$, that depends only on $q$, and a deterministic sequence $\gamma_n$ converging to 1, such that

(i) the exact option in (3.17) implements $\lambda > c\Lambda$ with probability at least

$$1 - \alpha - C(q) \left( \frac{\mathbb{E}[|\epsilon|^{q\vee 4}]}{n^{q/4}} \wedge \frac{\mathbb{E}[|\epsilon|^q]}{n^{1/4(q-2)}} \right) - \frac{16}{n};$$

(ii) the asymptotic option in (3.17) implements $\lambda > c\Lambda$ with probability at least

$$1 - \alpha \gamma_n - C(q) \left( \frac{\mathbb{E}[|\epsilon|^{q\vee 4}]}{n^{q/4}} \wedge \frac{\mathbb{E}[|\epsilon|^q]}{n^{1/4(q-2)}} \right) - \frac{25}{n} - 24\gamma_n \frac{\Phi^{-1}(1 - \alpha/2p)^2 \log n}{\alpha^2 n^{(1-2/q)/2}};$$

(iii) the approximate score $\Lambda(1 - \alpha |X)$ in (3.17) satisfies

$$\Lambda(1 - \alpha(1 + C(q)/\log^{-q/2} n) |X) \leq \sqrt{2n} \Phi^{-1}(1 - \alpha/2p)(1 + \sqrt{2\log \gamma_n/n})/(1 - \gamma_n - t_n) \leq \sqrt{2n} \log(p/\alpha)(1 + \sqrt{2\log \gamma_n/n})/(1 - \gamma_n - t_n)$$
where $r_n = \alpha - \frac{2}{q} n^{-[(1-2/q)\wedge 1/2]} \log n$, and $t_n = \Phi^{-1}(1 - \alpha/2p)$.

Proof of Theorem 15. Under conditions R and M, and using Lemma 8 we have with probability 1, we have with probability 1

Thus it suffices to prove that, conditional on $X$, setting $v = 1/2$, and condition R that ensures $c^2 \Phi^{-1}(1 - \alpha/2p) \leq \sigma^2$.

To show statement (ii), we define $t_n = \Phi^{-1}(1 - \alpha/2p)$ and $r_n = \alpha - \frac{2}{q} n^{-[(1-2/q)\wedge 1/2]} \log n$. Note that $u_n = 2/t_n$, $t_n \geq 1$, and $t_n^2 r_n \leq 1/3$.

By the same argument used to establish (i), under our conditions, we have $\tilde{\Lambda} \leq (1+u_n/4)\Lambda$ with probability

Thus it suffices to prove that, conditional on $X$, $\Lambda \leq (1+u_n/2)\sqrt{n} \Phi^{-1}(1 - \alpha/2p)$ with probability at least $1 - \alpha \gamma_n - \alpha (2) \gamma_n (\exp(t^2 r_n) - 1) - \psi(r_n)$.

Then for any $F = F_n$ and $X = X_n$ that obey Condition M:

Next note that by Condition M and Slastnikov’s theorem on moderate deviations, (20) (see also [27]) we have that uniformly in $0 \leq |t| \leq k \sqrt{\log n}$ for some $k^2 < q - 2$ and a sequence $\gamma_n \to 1$, uniformly in $1 \leq j \leq p$,

We apply the result above for $t = t_n(1 - r_n - r_n/t_n) \leq \sqrt{2 \log(2p/\alpha)} \leq \sqrt{\eta(q - 2) \log n}$ for $\eta < 1$ by assumption. Note that we apply Slastnikov’s theorem to $n^{-1/2} \sum^n_{i=1} z_{i,n}$ for
$z_{i,n} = x_{ij} \varepsilon_i$, where we allow the design $X$, the law $F$, and index $j$ to be (implicitly) indexed by $n$. Slavnikov’s theorem then applies provided

$$\sup_{n,j \leq p} \mathbb{E}_n |E_{n}|z_n|^q = \sup_{n,j \leq p} \mathbb{E}_n |x_{ij}|^q |E_{n}| |r|^q < \infty,$$

which is implied by our Condition M, and where we used the condition that the design is uniformly in $1 \leq j \leq p$ that obey our Condition M.

Thus, we obtained the moderate deviation result uniformly in $1 \leq j \leq p$ and for any sequence of distributions $F = F_n$ and designs $X = X_n$ that obey our Condition M.

$$\Pr(\Lambda > (1 + u_n/2) \sqrt{n} \gamma_n | X) \leq 2p \Phi(t_n(1 - r_n - r_n/t_n)) \gamma_n + \psi(r_n)$$

$$= 2p \Phi(t_n) \gamma_n + 2p \gamma_n \int_{t_n(1 - r_n - r_n/t_n)}^{t_n} \phi(u) du + \psi(r_n)$$

$$\leq \alpha \gamma_n + 2p \gamma_n \frac{\phi(t_n(1 - r_n - r_n/t_n)) - \phi(t_n)}{t_n(1 - r_n - r_n/t_n)} + \psi(r_n)$$

$$\leq \alpha \gamma_n + 2p \gamma_n \frac{\phi(t_n) \exp(t_n^2 r_n - 1)}{t_n - r_n - r_n/t_n} + \psi(r_n)$$

$$\leq \alpha \gamma_n + 24 \gamma_n t_n^2 r_n + \psi(r_n),$$

where we used that $\Phi(t) \leq \phi(t)/t$ and $\phi(t)/t \leq 2\Phi(t)$ for $t \geq 1$, $r_n \leq 1/3$, and that $\exp(m) \leq 1 + 2m$ for $0 \leq m \leq 1/3$.

To show statement (iii) of the lemma, let $\nu' > (1 + \sqrt{2 \log \gamma_n}/t_n)/(1 - r_n - 1/t_n),$$

$$\Pr(\Lambda > \nu' \sqrt{n} \gamma_n | X) = \Pr(\sqrt{n} \|\mathbb{E}_n [x_i \varepsilon_i]\|_\infty > \nu' \sqrt{n} \sqrt{\mathbb{E}_n [\varepsilon_i^2]} | X)$$

$$\leq \Pr(\sqrt{n} \|\mathbb{E}_n [x_i \varepsilon_i]\|_\infty > \nu' \sqrt{n} (1 - r_n) | X) + \Pr(\sqrt{n} \sqrt{\mathbb{E}_n [\varepsilon_i^2]} < (1 - r_n))$$

$$\leq \Pr(\sqrt{n} \|\mathbb{E}_n [x_i \varepsilon_i]\|_\infty > \nu' \sqrt{n} (1 - r_n - 1/t_n) | X) + \psi(r_n)$$

$$\leq \Pr(\sqrt{n} \|\mathbb{E}_n [x_i \varepsilon_i]\|_\infty > \nu' \sqrt{n} + \sqrt{2 \log \gamma_n} | X) + \psi(r_n)$$

where the last step follows by Lemma 9. Thus the bound

$$\Pr(\sqrt{n} \|\mathbb{E}_n [x_i \varepsilon_i]\|_\infty > \nu' \sqrt{n} + \sqrt{2 \log \gamma_n} | X) \leq \gamma_n \Phi(t_n + \sqrt{2 \log \gamma_n}) \leq \alpha$$

follows analogously to the proof of statement (ii); we omit the details for brevity. Therefore, under our conditions, $\Lambda(1 - \alpha - \psi(r_n)|X) \leq \nu' \sqrt{n} \Phi^{-1}(1 - \alpha/2p)$. Note that by construction $\psi(r_n) \leq \alpha C(q) \log^{-q/2} n$. The last bound follows from standard tail bounds for Gaussian random variables. \qed
Appendix D. Proofs of Section 3

Proof of Theorem 1. The exact option in (3.17) is \( \lambda = (1 + u_n)\alpha(1 - \alpha|X|) \) so that

\[
P(c\Lambda \geq \lambda|X) = P(\Lambda \geq (1 + u_n)\alpha(1 - \alpha|X|)|X)
\]

\[
\leq P(\Lambda \geq \lambda(1 - \alpha|X|)|X) + P(\Lambda \geq (1 + u_n)\Lambda|X)
\]

\[
\leq \alpha + P(\Lambda \geq (1 + u_n)\Lambda|X).
\]

By (3.16) and Lemma 8 with \( v = r_n \), the last term satisfies

\[
P(\Lambda \geq (1 + u_n)\Lambda|X) \leq P(\sqrt{\mathbb{E}_n[\sigma^2 \epsilon_i^2]} \geq (1 + u_n)\sqrt{\mathbb{E}_n[(\sigma \epsilon_i + r_i)^2]}|X)
\]

\[
\leq \psi(r_n) + \frac{2(1 + u_n)}{n(1 - r_n)u_n}
\]

by the choice of \( r_n = (\alpha^{-1} \log n \epsilon_q \mathbb{E}[|\epsilon|^q])^{1/2} / n^{1/4} < 1/2 \) under condition R’.

\[\Box\]

Proof of Theorem 2. Let \( t_n = \Phi^{-1}(1 - \alpha/2p) \) and recall \( r_n = (\alpha^{-1} \log n \epsilon_q \mathbb{E}[|\epsilon|^q])^{1/2} < 1/2 \) under Condition R’. Next, note that for \( u_n \geq 0 \) it follows that \( 1 + u_n \geq (1 + u_n/2)(1+[u_n \wedge 1]/4) \), so that

\[
P(\Lambda \geq (1 + u_n)\sqrt{nt_n}|X) \leq P(\Lambda \geq (1 + u_n/2)\sqrt{nt_n}|X) + P(\Lambda \geq (1 + u_n^4/4)\Lambda|X).
\]

Regarding the second term, by (3.16) and Lemma 8 with \( v = r_n \), it follows that

\[
P(\Lambda \geq (1 + u_n^4/4)\Lambda|X) \leq P(\sqrt{\mathbb{E}_n[\sigma^2 \epsilon_i^2]} \geq (1 + u_n^4/4)\sqrt{\mathbb{E}_n[(\sigma \epsilon_i + r_i)^2]}|X)
\]

\[
\leq \psi(r_n) + \frac{2(1 + u_n^4/4)}{n(1 - r_n)(u_n \wedge 1)/4}
\]

Next note that

\[
\Pr(\Lambda > (1 + u_n/2)\sqrt{nt_n}|X) \leq \Pr\left(\frac{n^{1/2}||\mathbb{E}_n[\epsilon_i^2]|X||}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} > (1 + u_n/4)t_n|X\right) + \Pr\left(\frac{\sqrt{n}}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} > \sqrt{nt_n}u_n/4|X\right).
\]

To control the second term above, note that under Condition M’ and R’, \( 4/nt_n \leq 1 - r_n \), we have that by Lemma 9

\[
\Pr\left(\frac{\sqrt{n}}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} > \sqrt{nt_n}u_n/4\right) = \Pr\left(\mathbb{E}_n[\epsilon_i^2] < \frac{4}{u_n t_n}|X\right) \leq \psi(r_n) \leq \alpha/\log n.
\]
\[\sqrt{\text{lasso}}\]

Next, under Condition M’ and R’, letting \(c_1, c_2\) as in Lemma 11 with \(\tau = \alpha / \log n\), we have \((1 + u_n/4) \geq \sqrt{c_1 c_2}\). Thus, since \(t_n + 1 \leq n^{1/6}/\max_{1 \leq j \leq p} (E_n[|x_{ij}|^3]E[|\epsilon_i|^3])^{1/3}\) by Condition R’, Lemma 11 yields

\[
\Pr\left(\frac{n^{1/2}\|E_n[x_i \epsilon_i]\|_\infty}{\sqrt{E_n[\epsilon_i^2]}} > (1 + u_n/4)t_n | X\right) \leq 2p\Phi(t_n) \left(1 + \frac{A}{\ell_n^3}\right) + \frac{\alpha}{\log n} \\
\leq \alpha \left(1 + \frac{A}{\ell_n^3} + \frac{1}{\log n}\right).
\]

\[\square\]

**Proof of Theorem 3** Under Conditions M’ and R’, note that by triangle inequality and by Lemma 11 with \(\tau = \alpha / \log n\) we have

\[
\Pr(A > (\nu + \nu')\sqrt{n}t_n | X) \leq \Pr\left(\frac{n^{1/2}\|E_n[x_i \epsilon_i]\|_\infty}{\sqrt{E_n[\epsilon_i^2]}} > \nu t_n | X\right) + \Pr\left(\sqrt{E_n[\epsilon_i^2]} < 1/|\nu' t_n| | X\right) \\
\leq 2p\Phi(\nu t_n/\sqrt{c_1 c_2})(1 + A/\ell_n^3) + 2\alpha / \log n < \alpha
\]

provided that \(\nu > \sqrt{c_1 c_2}(1 + \sqrt{2\log(3 + 3A/\ell_n^3)})/t_n\) we have \(2p\Phi(\nu t_n/c_1 c_2)(1 + A/\ell_n^3) \leq \alpha/3\), \(\nu' \geq 1/[(1 - r_t) t_n]\) we have \(\psi((\nu' t_n - 1)/|\nu' t_n|) < \alpha / \log n\), and \(\log n > 2\).

\[\square\]

**Appendix E. Proofs of Section 4**

**Proof of Theorem 4** We prove the upper bound on the prediction norm in two steps.

Step 1. In this step we show that \(\delta = \beta - \beta_0 \in \Delta_{\bar{e}}\) under the prescribed penalty level. By definition of \(\hat{\beta}\)

\[
\sqrt{\bar{Q}(\beta)} - \sqrt{\bar{Q}(\beta_0)} \leq \frac{\lambda}{n} ||\beta_0||_1 - \frac{\lambda}{n} ||\hat{\beta}||_1 \leq \frac{\lambda}{n} (||\hat{\delta_T||_1 - ||\hat{\delta}_{T^c}||_1),
\]

where the last inequality holds because

\[
||\beta_0||_1 - ||\hat{\beta}||_1 = ||\beta_0||_1 - ||\hat{\delta_T||_1 - ||\hat{\delta}_{T^c}||_1 \leq ||\hat{\delta_T||_1 - ||\hat{\delta}_{T^c}||_1.
\]

Note that using the convexity of \(\sqrt{\bar{Q}}\), \(-\bar{S} \in \partial \sqrt{\bar{Q}(\beta_0)}\), and if \(\lambda \geq cn\|\bar{S}\|_\infty\), we have

\[
\sqrt{\bar{Q}(\beta)} - \sqrt{\bar{Q}(\beta_0)} \geq -\bar{S}^T \delta \geq -\|\bar{S}\|_\infty ||\delta||_1
\]

\[
\geq -\frac{\lambda}{cn} (||\hat{\delta}_T||_1 + ||\hat{\delta}_{T^c}||_1).
\]

(E.27)
Combining (E.25) with (E.28) we obtain
\[- \frac{\lambda}{cn} (\|\hat{\delta}_T\|_1 + \|\hat{\delta}_{T^c}\|_1) \leq \frac{\lambda}{n} (\|\hat{\delta}_T\|_1 - \|\hat{\delta}_{T^c}\|_1), \tag{E.29}\]

that is
\[\|\hat{\delta}_{T^c}\|_1 \leq \frac{c + 1}{c - 1} \|\hat{\delta}_T\|_1, \quad \text{or} \quad \hat{\delta} \in \Delta_\varepsilon. \tag{E.30}\]

Step 2. In this step we derive bounds on the estimation error. We shall use the following relations:
\[\hat{Q}(\beta) - \hat{Q}(\beta_0) = \|\hat{\delta}\|_{2,n}^2 - 2E_n[(\sigma \varepsilon_i + r_i)x_i'] \hat{\delta}, \tag{E.31}\]
\[\hat{Q}(\beta) - \hat{Q}(\beta_0) = \left( \sqrt{\hat{Q}(\beta)} + \sqrt{\hat{Q}(\beta_0)} \right) \left( \sqrt{\hat{Q}(\beta)} - \sqrt{\hat{Q}(\beta_0)} \right), \tag{E.32}\]
\[2|E_n(\sigma \varepsilon_i + r_i)x_i'\hat{\delta}| \leq 2\|E_n[(\sigma \varepsilon_i + r_i)x_i]\|_\infty \|\hat{\delta}\|_1 \leq 2\sqrt{\hat{Q}(\beta_0)} \lambda \|\hat{\delta}\|_1/[cn], \tag{E.33}\]
\[\|\hat{\delta}_T\|_1 \leq \frac{\sqrt{s} \|\hat{\delta}\|_{2,n}}{\kappa_\varepsilon} \quad \text{for} \quad \hat{\delta} \in \Delta_\varepsilon, \tag{E.34}\]

where (E.28) holds by Holder inequality and the last inequality holds by the definition of \(\kappa_\varepsilon\).

Note that by (E.31), if \(\sqrt{\hat{Q}(\beta)} + \sqrt{\hat{Q}(\beta_0)} = 0\), we have \(\|\delta\|_{2,n} = 0\) and we are done. So we can assume \(\sqrt{\hat{Q}(\beta)} + \sqrt{\hat{Q}(\beta_0)} > 0\). Using (E.25) and (E.31)-(E.34) we obtain
\[\|\hat{\delta}\|_{2,n}^2 \leq \frac{2\lambda}{cn} \sqrt{\hat{Q}(\beta_0)} \|\hat{\delta}\|_1 + \left( \sqrt{\hat{Q}(\beta)} + \sqrt{\hat{Q}(\beta_0)} \right) \frac{\lambda}{n} \left( \frac{\sqrt{s} \|\hat{\delta}\|_{2,n}}{\kappa_\varepsilon} - \|\hat{\delta}_{T^c}\|_1 \right). \tag{E.35}\]

Also using (E.25) and (E.34) we obtain
\[\sqrt{\hat{Q}(\beta)} \leq \sqrt{\hat{Q}(\beta_0)} + \frac{\lambda}{n} \left( \frac{\sqrt{s} \|\hat{\delta}\|_{2,n}}{\kappa_\varepsilon} - \|\hat{\delta}_{T^c}\|_1 \right) \leq \sqrt{\hat{Q}(\beta_0)} + \frac{\lambda \sqrt{s} \|\hat{\delta}\|_{2,n}}{\kappa_\varepsilon}. \tag{E.36}\]

Combining inequalities (E.30) and (E.35), we obtain
\[\|\hat{\delta}\|_{2,n}^2 \leq \frac{2\lambda}{cn} \sqrt{\hat{Q}(\beta_0)} \|\hat{\delta}\|_1 + 2\sqrt{\hat{Q}(\beta_0)} \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \|\hat{\delta}\|_{2,n} + \left( \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \|\hat{\delta}\|_{2,n} \right)^2 - 2\sqrt{\hat{Q}(\beta_0)} \frac{\lambda}{n} \|\hat{\delta}_{T^c}\|_1. \]

Simplifying further the expression we obtain
\[\|\hat{\delta}\|_{2,n}^2 \leq \frac{2\lambda}{cn} \sqrt{\hat{Q}(\beta_0)} \|\hat{\delta}\|_1 + 2\sqrt{\hat{Q}(\beta_0)} \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \|\hat{\delta}\|_{2,n} + \left( \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \|\hat{\delta}\|_{2,n} \right)^2, \]
and then using (E.34) we obtain
\[\left[ 1 - \left( \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \right)^2 \right] \|\hat{\delta}\|_{2,n}^2 \leq 2 \left( \frac{1}{c} + 1 \right) \sqrt{\hat{Q}(\beta_0)} \frac{\lambda \sqrt{s}}{n \kappa_\varepsilon} \|\hat{\delta}\|_{2,n}. \]
Provided that $\frac{\lambda \sqrt{s}}{n \kappa_c} < 1$ and solving the inequality above we obtain the bound stated in the theorem.

Proof of Theorem 5. Let $\delta := \hat{\beta} - \beta_0$. Under the condition on $\lambda$ above, we have that $\delta \in \Delta_{\bar{c}}$. Thus, we have

$$\|\delta\|_1 \leq (1 + \bar{c}) \|\delta_{T}\|_1 \leq (1 + \bar{c}) \frac{\sqrt{s} \|\delta\|_{2,n}}{\kappa_c},$$

by the restricted eigenvalue condition.

Proof of Theorem 6. Let $\delta := \hat{\beta} - \beta_0$. We have that

$$\|\delta\|_{\infty} \leq \|E_n[x_i x_{i}^T \delta]\|_{\infty} + \|E_n[x_i x_{i}^T] - \delta\|_{\infty}.$$

Note that by the first order optimality conditions of $\hat{\beta}$ and the assumption on $\lambda$

$$\|E_n[x_i x_{i}^T \delta]\|_{\infty} \leq \|E_n[x_i (y_i - x_{i}^T \hat{\beta})]\|_{\infty} + \|S\|_{\infty} \sqrt{\bar{Q}(\delta_0)}$$

$$\leq \frac{\lambda \sqrt{Q(\beta_0)}}{n} + \frac{\lambda \sqrt{Q(\beta_0)}}{c n}$$

by the first order conditions and the condition on $\lambda$.

Next let $e_j$ denote the $j$th-canonical direction.

$$|E_n[e_j x_{i}^T \delta] - \delta_j| \leq |E_n[e_j (x_{i}^T I) \delta]|$$

$$\leq \|\delta\|_1 \|E_n[x_i x_{i}^T] - I\|_{\infty}.$$

Therefore, using the optimality of $\hat{\beta}$ that implies $\sqrt{\bar{Q}(\hat{\beta})} \leq \sqrt{\bar{Q}(\beta_0)} + (\lambda/n)\|\delta\|_1$, we have

$$\|\delta\|_{\infty} \leq \left( \sqrt{\bar{Q}(\hat{\beta})} + \frac{\sqrt{Q(\beta_0)}}{c} \right) \frac{\lambda}{n} + \|E_n[x_i x_{i}^T] - I\|_{\infty} \|\delta\|_1$$

$$\leq (1 + \frac{1}{c}) \frac{\lambda \sqrt{Q(\beta_0)}}{n} + \left( \frac{\lambda^2}{n^2} + \|E_n[x_i x_{i}^T] - I\|_{\infty} \right) \|\delta\|_1.$$

Proof of Theorem 7. Let $\delta := \hat{\beta} - \beta_0 \in \Delta_{\bar{c}}$ under the condition that $\lambda \geq c \|\bar{\Lambda}\|_{\infty}$ by Step 1 in Theorem 4.

First we establish the upper bound. By optimality of $\hat{\beta}$

$$\sqrt{\bar{Q}(\hat{\beta})} - \sqrt{\bar{Q}(\beta_0)} \leq \frac{\lambda}{n} ((\|\beta_0\|_1 - \|\hat{\beta}\|_1) \leq \frac{\lambda}{n} (\|\delta_{T}\|_1 - \|\delta_{T'}\|_1).$$
Thus, if $\delta \notin \Delta_1$ we have $\hat{Q}(\hat{\beta}) \leq \hat{Q}(\beta_0)$. So we can assume $\delta \in \Delta_1$ in the bound above, so that $\|\delta_T\|_1 \leq \sqrt{s}\|\delta\|_{2,n}/\kappa_1$. Thus, setting $\rho_1 = \lambda\sqrt{s}/[n\kappa_1]$, we have $\|\delta\|_{2,n} \leq 2(1 + 1/c)\rho_1\sqrt{\hat{Q}(\beta_0)/[1 - \rho_2^2]}$ if $\delta \in \Delta_1$ by inspection of Theorem 4.

To establish the lower bound, by convexity of $\sqrt{\hat{Q}}$ and the condition that $\lambda \geq cn\|\tilde{S}\|_\infty$ we have

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta_0)} \geq -S^\alpha\delta \geq -\frac{\lambda\|\delta\|_1}{cn}.$$ 

Thus, combining these bounds with Theorems 4 and 5, letting $\rho := \lambda\sqrt{s}/[n\kappa_1] < 1$, we obtain

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta_0)} \geq -\frac{1 + c}{c} \frac{\rho^2}{1 - \rho^2} 2(1 + \frac{1}{c}) \sqrt{\hat{Q}(\beta_0)} = -\frac{4\rho}{c} \frac{\rho^2}{1 - \rho^2} \sqrt{\hat{Q}(\beta_0)}.$$ 

where the last equality follows from simplifications. □

**Proof of Lemma 7** First note that by strong duality we have

$$\mathbb{E}_n [y_i\hat{a}_i] = \frac{\|Y - X\hat{\beta}\|}{\sqrt{n}} + \frac{\lambda}{n} \sum_{j=1}^p |\hat{\beta}_j|.$$ 

Since $\mathbb{E}_n [x_{ij}\hat{a}_i] \hat{\beta}_j = \lambda|\hat{\beta}_j|/n$ for every $j = 1, \ldots, p$, we have

$$\mathbb{E}_n [y_i\hat{a}_i] = \frac{\|Y - X\hat{\beta}\|}{\sqrt{n}} + \sum_{j=1}^p \mathbb{E}_n [x_{ij}\hat{a}_i] \hat{\beta}_j = \frac{\|Y - X\hat{\beta}\|}{\sqrt{n}} + \mathbb{E}_n \left[\hat{a}_i \sum_{j=1}^p x_{ij}\hat{\beta}_j\right].$$ 

Rearranging the terms we have $\mathbb{E}_n \left[(y_i - x_i^t\hat{\beta})\hat{a}_i\right] = \|Y - X\hat{\beta}\|/\sqrt{n}$.

If $\|Y - X\hat{\beta}\| = 0$, we have $\sqrt{\hat{Q}(\hat{\beta})} = 0$ and the statement of the lemma trivially holds.

If $\|Y - X\hat{\beta}\| > 0$, since $\|\hat{a}\| \leq \sqrt{n}$ the equality can only hold for $\hat{a} = \sqrt{n}(Y - X\hat{\beta})/\|Y - X\hat{\beta}\| = (Y - X\hat{\beta})/\sqrt{\hat{Q}(\hat{\beta})}$.

Next, note that for any $j \in \hat{T}$ we have $\mathbb{E}_n [x_{ij}\hat{a}_i] = \text{sign}(\hat{\beta}_j)\lambda/n$. Therefore, we have

$$\sqrt{\hat{Q}(\hat{\beta})} \sqrt{|\hat{T}|}\lambda = \|(X'(Y - X\hat{\beta}))_{\hat{T}}\| \leq \|(X'(Y - X\beta_0))_{\hat{T}}\| + \|(X'X(\hat{\beta}_0 - \hat{\beta}))_{\hat{T}}\| \leq \sqrt{|\hat{T}|} n\mathbb{E}_n [x_i(\sigma_e r_i)]\|\sqrt{n} + n \phi_{\max}(\tilde{m})\|\hat{\beta} - \beta_0\|_{2,n},$$

where $\hat{T}$ is the set of indices of non-zero elements of $\hat{\beta}$.
where we used
\[ \| (X'X(\hat{\beta} - \beta_0))_T \| \leq \sup_{\|\alpha_T\|=\hat{m}_n, \|\alpha\|=1} |\alpha'X(\hat{\beta} - \beta_0)| \]
\[ \leq \sup_{\|\alpha_T\|=\hat{m}_n, \|\alpha\|=1} \|\alpha'X\| \|X(\hat{\beta} - \beta_0)\| \]
\[ = \sup_{\|\alpha_T\|=\hat{m}_n, \|\alpha\|=1} \sqrt{\|\alpha'X'\alpha\|} \|X(\hat{\beta} - \beta_0)\| \]
\[ = n\sqrt{\phi_{\max}(\hat{m})}\|\hat{\beta} - \beta_0\|_{2,n}. \]

\[ \square \]

**Proof of Theorem**\[11\] We can assume that \( \sqrt{\tilde{Q}(\beta_0)} > 0 \) otherwise the result is trivially true.

In the event \( \lambda \geq c\Lambda \), by Lemma \[1\]
\[ \left( \sqrt{\frac{\tilde{Q}(\beta)}{\tilde{Q}(\beta_0)}} - \frac{1}{c} \right) \lambda \sqrt{\tilde{Q}(\beta_0)} \sqrt{\hat{T}} \leq \sqrt{\phi_{\max}(\hat{m})}\|\hat{\beta} - \beta_0\|_{2,n}. \] (E.37)

Under the condition \( \lambda \sqrt{s} < n\kappa_c \rho, \rho < 1 \), we have by Theorem \[7\] that
\[ \left( 1 - \frac{\rho^2}{1 - \rho^2} \frac{4\hat{c}}{c} - \frac{1}{c} \right) \lambda \sqrt{\tilde{Q}(\beta_0)} \sqrt{\hat{T}} \leq \|\hat{\beta} - \beta_0\|_{2,n}. \]

\[ \square \]

**Proof of Theorem**\[12\] We can assume that \( \sqrt{\tilde{Q}(\beta_0)} > 0 \) otherwise the result follows by Theorem \[11\] which establish \( \hat{\beta} = \beta_0 \).

In the event \( \lambda \geq c\Lambda \), by Lemma \[1\]
\[ \left( \sqrt{\frac{\tilde{Q}(\beta)}{\tilde{Q}(\beta_0)}} - \frac{1}{c} \right) \lambda \sqrt{\tilde{Q}(\beta_0)} \sqrt{\hat{T}} \leq n\sqrt{\phi_{\max}(\hat{m})}\|\hat{\beta} - \beta_0\|_{2,n}. \] (E.38)

Under the condition \( \lambda \sqrt{s} \leq n\kappa_c \rho, \rho < 1 \), we have by Theorem \[4\] and Theorem \[7\] that
\[ \left( 1 - \frac{\rho^2}{1 - \rho^2} \frac{4\hat{c}}{c} - \frac{1}{c} \right) \lambda \sqrt{\tilde{Q}(\beta_0)} \sqrt{\hat{T}} \leq n\sqrt{\phi_{\max}(\hat{m})}2 \left( 1 + \frac{1}{c} \right) \sqrt{\tilde{Q}(\beta_0)} \frac{1}{1 - \rho^2} \frac{\lambda \sqrt{s}}{n\kappa_c}. \]

Since \( \rho^2 < (c - 1)/(c - 1 + 4\hat{c}) \) we have
\[ \sqrt{\hat{T}} \leq \frac{\sqrt{s}}{\kappa_c} \cdot \frac{2 \sqrt{\phi_{\max}(\hat{m})}(c + 1)}{c - 1 - \rho^2(c - 1 + 4\hat{c})} = \sqrt{\phi_{\max}(\hat{m})} \cdot \frac{2\hat{c}}{\kappa_c} \left( 1 + \frac{\rho^2(c - 1 + 4\hat{c})}{c - 1 - \rho^2(c - 1 + 4\hat{c})} \right). \]

Under the condition \( 2\rho^2 \leq (c - 1)/(c - 1 + 4\hat{c}) \), the expression above simplifies to
\[ |\hat{T}| \leq \phi_{\max}(\hat{m}) \left( \frac{4\hat{c}}{\kappa_c} \right)^2 \] (E.39)

since \( \rho^2(c - 1 + 4\hat{c}) \leq c - 1 - \rho^2(c - 1 + 4\hat{c}) \).
Consider any \( m \in \mathcal{M} \), and suppose \( \hat{m} > m \). Therefore by sublinearity of sparse eigenvalues

\[
\hat{m} \leq s \cdot \left[ \frac{\hat{m}}{m} \right] \phi_{\max}(m) \left( \frac{4\tilde{c}}{\kappa_{c}} \right)^2.
\]

Thus, since \( \lceil k \rceil < 2k \) for any \( k \geq 1 \) we have \( m < s \cdot 2\phi_{\max}(m)(4\tilde{c}/\kappa_{c})^2 \) which violates the condition of \( m \in \mathcal{M} \) and \( s \). Therefore, we must have \( \hat{m} \leq m \). In turn, applying (E.39) once more with \( \hat{m} \leq m \) we obtain \( \hat{m} \leq s \cdot \phi_{\max}(m)(4\tilde{c}/\kappa_{c})^2 \). The result follows by minimizing the bound over \( m \in \mathcal{M} \).

**Lemma 12** (Performance of a generic second-step estimator). Let \( \hat{\beta} \) be any first-step estimator with support \( \hat{T} \), define

\[
B_n := \hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0), \quad C_n := \hat{Q}(\beta_{0T}) - \hat{Q}(\beta_0), \quad \text{and} \quad D_m := \sup_{\|\delta_{Tc}\|_0 \leq m, \|\delta\|_{2,n} > 0} \left| E_n \left[ \sigma_{\epsilon,x} \delta \right] \right|,
\]

and let \( \tilde{\beta} \) the second-step estimator. Then, we have that for \( \hat{m} := |\hat{T} \setminus T| \)

\[
\|\hat{\beta} - \beta_0\|_{2,n} \leq D_m + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}.
\]

**Proof of Lemma 12.** Let \( \tilde{\delta} := \hat{\beta} - \beta_0 \). By definition of the second-step estimator, it follows that \( \hat{Q}(\hat{\beta}) \leq \hat{Q}(\tilde{\beta}) \) and \( \hat{Q}(\tilde{\beta}) \leq \hat{Q}(\beta_{0T}) \). Thus,

\[
\hat{Q}(\tilde{\beta}) - \hat{Q}(\beta_0) \leq \left( \hat{Q}(\tilde{\beta}) - \hat{Q}(\beta_0) \right) \wedge \left( \hat{Q}(\beta_{0T}) - \hat{Q}(\beta_0) \right) \leq B_n \wedge C_n.
\]

Moreover, since

\[
\left| \hat{Q}(\tilde{\beta}) - \hat{Q}(\beta_0) - \|\tilde{\delta}\|_{2,n}^2 \right| \leq \frac{\|S'\tilde{\delta}\|}{\|\hat{\delta}\|_{2,n}} \|\hat{\delta}\|_{2,n} + 2c_s \|\tilde{\delta}\|_{2,n}
\]

we have

\[
\|\tilde{\delta}\|_{2,n} \leq \left( \frac{\|S'\tilde{\delta}\|}{\|\hat{\delta}\|_{2,n}} + 2c_s \right) \|\hat{\delta}\|_{2,n} + B_n \wedge C_n.
\]

Solving which we obtain the stated result:

\[
\|\tilde{\delta}\|_{2,n} \leq D_m + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}
\]

using the definition of \( D_m \). \( \square \)

**Lemma 13** (Control of \( B_n \) and \( C_n \) for \( \sqrt{\text{LASSO}} \)). Under condition ASM, if \( \lambda \geq c\bar{A} \) and \( \rho_1 = \lambda\sqrt{s}/[n\kappa_1] < 1 \), we have that

\[
B_n \wedge C_n \leq 1\{T \not\subset \hat{T}\} \cdot \hat{Q}(\beta_0)(1 + 1/c) \frac{4\rho^2_1}{1 - \rho^2_1} \left( 1 + \frac{(1 + 1/c)\rho^2_1}{1 - \rho^2_1} \right).
\]
Proof of Lemma 13. First note that \( C_n = 0 \) if \( T \subseteq \hat{T} \).

Let \( \delta = \hat{\beta} - \beta_0 \). By optimality of \( \hat{\beta} \) for (1.3), we have

\[
\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) \leq \left( \sqrt{\hat{Q}(\hat{\beta})} + \sqrt{\hat{Q}(\beta_0)} \right) \frac{1}{n} \left( \| \delta_T \|_1 - \| \delta_T^c \|_1 \right)
\]

\[
\leq 2 \sqrt{\hat{Q}(\beta_0)} \frac{1}{n} \left( \| \delta_T \|_1 - \| \delta_T^c \|_1 \right) + \frac{\lambda^2}{n} \left( \| \delta_T \|_1 - \| \delta_T^c \|_1 \right)^2.
\]

If \( \| \delta_T^c \|_1 > \| \delta_T \|_1 \), we have \( B_n \leq 0 \). Otherwise we have \( \| \delta_T \|_1 \leq \sqrt{s} \| \delta \|_{2,n}/\kappa_1 \).

Moreover, under \( \| \delta_T \|_1 \leq \| \delta_T^c \|_1 \) we have that \( \| \delta \|_{2,n} \leq 2(1 + 1/c) \sqrt{\hat{Q}(\beta_0)\rho_1/(1 - \rho_1^2)} \), for \( \lambda \sqrt{s} \leq n \kappa_1 \rho_1, \rho_1 < 1 \), by Theorem 4.

Thus,

\[
\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) \leq 2 \hat{Q}(\beta_0) \frac{2(1 + 1/c)\rho_1^2}{1 - \rho_1^2} + \rho_1^2 \frac{2(1 + 1/c)\hat{Q}(\beta_0)\rho_1^2}{(1 - \rho_1^2)^2}.
\]

\[
\square
\]

Lemma 14 (Control of \( D_m \)). Suppose condition ASM holds. We have that uniformly over \( m \leq p \) with probability at least \( 1 - 1/C^2 - 1/[9C^2 \log p] \)

\[
D_m \leq \frac{C\sigma}{\sqrt{\phi_{\min}(m)}} \sqrt{\frac{s}{n}} + \frac{24C\sigma}{\sqrt{\phi_{\min}(m)}} \sqrt{1 + \max_{j=1,...,p} \mathbb{E}_n \left[ x_{ij}^2 \gamma_j^2 \right]} \sqrt{\frac{m \log p}{n}}.
\]

Proof of Lemma 14. By the triangle inequality we have

\[
D_m := \sup_{\| \delta_T \|_0 \leq m, \| \delta \|_{2,n} > 0} \left| \mathbb{E}_n \left[ \sigma \epsilon_i x_i^T \delta \| \delta \|_{2,n} \right] \right| \leq \sup_{\| \delta_T \|_0 \leq m, \| \delta \|_{2,n} > 0} \mathbb{E}_n \left[ \sigma \epsilon_i x_i^T x_i \delta \| \delta \|_{2,n} \right] + \frac{1}{\sqrt{n}} \mathbb{E}_n \left[ \sigma \epsilon_i x_i^T \delta \| \delta \|_{2,n} \right] \leq D_m^T + D_m^c.
\]

To bound \( D_m^T \), note that \( \| \delta \|_{2,n} \geq \sqrt{\phi_{\min}(m)} \| \delta \| \) and \( x_i^T \delta = x_i^T \delta_T \), so that

\[
D_m^T \leq \frac{\sigma \| \mathbb{E}_n[\epsilon_i x_i x_i^T] \| \| \delta_T \|}{\sqrt{\phi_{\min}(m)} \| \delta \|}.
\]

We proceed to bound the second moment of \( \| \mathbb{E}_n[\epsilon_i x_i x_i^T] \| \). Thus, since \( \mathbb{E}[\epsilon_i] = 0 \) and \( \mathbb{E}[\epsilon_i^2] \),

\[
\mathbb{E} \left[ \| \mathbb{E}_n[\epsilon_i x_i x_i^T] \|^2 \right] = \mathbb{E} \left[ \mathbb{E}_n[\epsilon_i x_i x_i^T]^2 \right] = \mathbb{E} \left[ \mathbb{E}_n[\epsilon_i x_i x_i^T] \right] = \text{trace} \left( \mathbb{E}_n[\epsilon_i x_i x_i^T] / n \right) = s/n.
\]

Therefore, by Chebyshev inequality we have with probability at least \( 1 - 1/C^2 \)

\[
D_m^T \leq \frac{C\sigma}{\sqrt{\phi_{\max}(m)}} \sqrt{\frac{s}{n}}.
\]
Next, to bound $D_m^T$, we note that the support of $\text{supp}(\delta) \subseteq T \cup \hat{T}$. Thus, we have $||\delta||_{2,n} \geq \sqrt{\phi_{\text{min}}(m)}||\delta||$ and

$$E_n[\epsilon_i x_i^T \cdot \delta] = E_n[\epsilon_i x_i^T \cdot \delta^{T_e}] \leq \|E_n[\epsilon_i x_i^{T_e}]\|_\infty \|\delta^{T_e}\|_1 \leq \|E_n[\epsilon_i x_i^{T_e}]\|_\infty \sqrt{m} \|\delta^{T_e}\|_1.$$ 

Thus, we have

$$D_m^T \leq \sqrt{m} \sigma \|E_n[\epsilon_i x_i^{T_e}]\|_\infty / \sqrt{\phi_{\text{min}}(m)}.$$ 

Finally, by Corollary 6 with probability at least $1 - 1/\lceil 9C^2 \log p \rceil$, we have

$$\|E_n[\epsilon_i x_i^{T_e}]\|_\infty \leq 24C \sqrt{1 \vee \max_{j=1, \ldots, p} E_n \left[ x_{ij}^2 \epsilon_i^2 \right]} \sqrt{\log p / n}.$$ 

Proof of Theorem 10. The theorem follows by Lemma 12 and applying Lemma 13 to bound $B_n \wedge C_n$ and Lemma 14 to bound $D_{\hat{m}}$.

Proof of Theorem 11. Note that because $\sigma = 0$ and $c_s = 0$, we have $\sqrt{Q(\beta_0)} = 0$ and $\sqrt{Q(\hat{\beta})} = \|\hat{\beta} - \beta_0\|_{2,n}$. Thus, by optimality of $\hat{\beta}$ we have

$$\|\hat{\beta} - \beta_0\|_{2,n} + \frac{\lambda}{n} \|\hat{\beta}\|_1 \leq \frac{\lambda}{n} \|\beta_0\|_1.$$ 

Thus $\|\hat{\beta}\|_1 \leq \|\beta_0\|_1$ and $\delta = \hat{\beta} - \beta_0$ satisfies $\|\delta^{T_e}\|_1 \leq \|\delta_T\|_1$. This implies that

$$\|\delta\|_{2,n} \leq \frac{\lambda}{n} \|\delta_T\|_1 \leq \frac{\lambda \sqrt{s}}{n \kappa_1} \|\delta\|_{2,n}.$$ 

Since $\lambda \sqrt{s} < n \kappa_1$ we have $\|\delta\|_{2,n} = 0$. In turn, $0 = \sqrt{s} \|\delta\|_{2,n} \geq \kappa_1 \|\delta_T\|_1 \geq \kappa_1 \|\delta\|_{1,1}/2$. This shows that $\delta = 0$ and $\hat{\beta} = \beta_0$.

Lemma 15. Consider $\epsilon_i \sim t(2)$. Then, for $\tau \in (0, 1)$ we have that:

(i) $P(E_n[\epsilon_i^2] \geq 2\sqrt{2}/\tau + \log(4n/\tau)) \leq \tau/2$.

(ii) Let $\bar{x} := \max_{i,j} |x_{ij}|$ and $0 < a < 1/6$. Then, with probability at least $1 - \tau - 1/n^{3/2(1/6-a)^2}$ we have

$$\left\| \frac{E_n[x_i \epsilon_i]}{\sqrt{E_n[\epsilon_i^2]}} \right\|_\infty \leq \frac{4\bar{x} \sqrt{2 \log(4p/\tau) \log(4n/\tau)} + 2\sqrt{2}/\tau}{\sqrt{n} \sqrt{a \log n}}.$$
(iii) For $u_n \geq 0$ and $0 < a < 1/6$, we have

$$P(\sqrt{\mathbb{E}_n[\sigma^2 \epsilon_i^2]} \leq (1+u_n)\sqrt{\mathbb{E}_n[(\sigma \epsilon_i + r_i)^2]}) \leq \frac{4\sigma^2 c_2^2 \log(4n/\tau)}{n(c_2^2 + |u_n|(1+u_n)\sigma^2 a \log n)^2} + \frac{1}{n^{1/2}(1/6 - a)^2} + \tau/2.$$ 

Proof of Lemma 15. To show (i) we will establish a bound on $q(\mathbb{E}_n[\epsilon_i^2], 1 - \tau)$. Recall that for a $t(2)$ random variable, the cumulative distribution function and the density function are given by:

$$F(x) = \frac{1}{2} \left( 1 + \frac{x}{\sqrt{2 + x^2}} \right) \quad \text{and} \quad f(x) = \frac{1}{(2 + x^2)^{3/2}}.$$ 

For any truncation level $t_n \geq \sqrt{2}$ we have

$$\mathbb{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}] = 2 \int_0^{\sqrt{2}} \frac{x^2 dx}{(2+x^2)^{3/2}} + 2 \int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^2 dx}{(2+x^2)^{3/2}} \leq 2 \int_0^{\sqrt{2}} \frac{x^2 dx}{(2+x^2)^{3/2}} + 2 \int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^2 dx}{x^3} \leq \log t_n.$$ 

$$\mathbb{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}] \leq 2 \int_0^{\sqrt{2}} \frac{x^2 dx}{3^{3/2}} + 2 \int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^2 dx}{x^{3/2}} \leq t_n.$$ 

(E.40)

Also, because $1 - \sqrt{1 - v} \leq v$ for every $0 \leq v \leq 1$,

$$P(|\epsilon_i|^2 > t_n) = \left( 1 - \sqrt{\frac{t_n}{2 + t_n}} \right) \leq 2/(2 + t_n).$$ 

(E.41)

Thus, by setting $t_n = 4n/\tau$ and $t = 2\sqrt{2}/\tau$ we have (14), relation (7.5),

$$P(|\mathbb{E}_n[\epsilon_i^2] - \mathbb{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}]| \geq t) \leq \frac{\mathbb{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}]}{nt^2} + nP(|\epsilon_i^2| > t_n) \leq \frac{t_n}{nt^2} + \frac{2n}{2 + t_n} \leq \tau/2.$$ 

(E.42)

Thus, (i) is established and we have

$$q \left( \max_{1 \leq j \leq p} \mathbb{E}_n[x_{ij}^2 \epsilon_i^2], 1 - \tau/2 \right) \leq \bar{x}^2 q(\mathbb{E}_n[\epsilon_i^2], 1 - \tau/2) \leq \bar{x}^2 \left( \log(4n/\tau) + 2\sqrt{2}/\tau \right).$$

To show (ii), using the bound above and Theorem (14) we have that with probability $1 - \tau$ we have

$$\|G_n [x_i \epsilon_i]\|_\infty \leq 4\bar{x} \sqrt{2 \log(4p/\tau)} \sqrt{\log(4n/\tau) + 2\sqrt{2}/\tau}.$$
Moreover, for $0 < a < 1/6$, we have

$$P(\mathbb{E}_n[\epsilon_i^2] \leq a \log n) \leq P(\mathbb{E}_n[\epsilon_i^21\{\epsilon_i^2 \leq n^{1/2}\}] \leq a \log n)$$

$$\leq P(\mathbb{E}_n[\epsilon_i^21\{\epsilon_i^2 \leq n^{1/2}\}] - \mathbb{E}[\epsilon_i^21\{\epsilon_i^2 \leq n^{1/2}\}] \geq (1/6) \log n - a \log n)$$

$$\leq \frac{1}{n^{1/2}(1/6-a)^2}$$

(E.43)

by Chebyshev inequality and since $\mathbb{E}[\epsilon_i^21\{\epsilon_i^2 \leq n^{1/2}\}] \geq (1/6) \log n$. Thus, with probability at least $1 - \tau - \frac{1}{n^{1/2}(1/6-a)^2}$ we have

$$\left\| \frac{\mathbb{E}_n[x_i\epsilon_i]}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} \right\|_\infty \leq 4\bar{x} \sqrt{2 \log(4p/\tau)} \sqrt{\log(4n/\tau) + 2\sqrt{2/\tau}} \sqrt{a \log n}.$$

To establish (iii), let $a_n = [(1 + u_n^2 - 1)/(1 + u_n)^2 = u_n(2 + u_n)/(1 + u_n)^2 \geq u_n/(1 + u_n)$ and note that by (E.40), (E.42), and (E.43) we have

$$P\left(\sqrt{\mathbb{E}_n[\sigma^2\epsilon_i^2]} > (1 + u_n)\sqrt{\mathbb{E}_n[(\sigma_i + r_i)^2]}\right) = P\left(2\sigma \mathbb{E}_n[\epsilon_i r_i] > c_s^2 + a_n \mathbb{E}_n[\sigma^2\epsilon_i^2]\right)$$

$$\leq P\left(2\sigma \mathbb{E}_n[\epsilon_i r_i1\{\epsilon_i^2 \leq t_n\}] > c_s^2 + a_n \sigma^2 a \log n\right) + P\left(\mathbb{E}_n[\epsilon_i^2] \leq a \log n\right) + nP(\epsilon_i^2 \leq t_n)$$

$$\leq \frac{4\sigma^2 c_s^2 \log t_n}{n(c_s^2 + a_n \sigma^2 a \log n)^2} + \frac{1}{n^{1/2}(1/6-a)^2} + \tau/2.$$

\[\blacksquare\]

**Proof of Theorem 12.** The result follows from Theorem 11 and Lemma 15 which provides bounds for $\lambda$ and $\sqrt{Q(\beta_0)} \leq c_s + \sigma \sqrt{\mathbb{E}_n[\epsilon_i^2]}$. Note the simplification that

$$\frac{4\sigma^2 c_s^2 \log t_n}{n(c_s^2 + a_n \sigma^2 a \log n)^2} \leq \frac{2 \log t_n}{na_n a \log n}.$$

The result follows with $t_n = 4n/\tau$, $a = 1/24$, and $a_n = u_n/[1 + u_n]$. \[\blacksquare\]

**Appendix F. Comparing Computational Methods for LASSO and \sqrt{LASSO}**

Below we discuss in more detail the applications of these methods for lasso and $\sqrt{\text{LASSO}}$. For each method, the similarities between the lasso and $\sqrt{\text{LASSO}}$ formulations derived below provide theoretical justification for the similar computational performance.

**Interior point methods.** Interior point methods (IPMs) solvers typically focus on solving conic programming problems in standard form,

$$\min_w c'w : Aw = b, w \in K.$$  (F.44)
The main difficulty of the problem arises because the conic constraint will be binding at the optimal solution.

IPMs regularize the objective function of the optimization with a barrier function so that the optimal solution of the regularized problem naturally lies in the interior of the cone. By steadily scaling down the barrier function, an IPM creates a sequence of solutions that converges to the solution of the original problem (F.44).

In order to formulate the optimization problem associated with the lasso estimator as a conic programming problem (F.44), we let \( \beta = \beta^+ - \beta^- \), and note that for any vector \( v \in \mathbb{R}^n \) and any scalar \( t \geq 0 \) we have that

\[
v'v \leq t \quad \text{is equivalent to} \quad \| (v, (t-1)/2) \|_2 \leq (t + 1)/2.
\]

Thus, we have that lasso optimization problem can be cast

\[
\min_{t, \beta^+, \beta^-, a_1, a_2, v} \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \sum_{j=1}^{p} \beta_j^+ + \beta_j^- \\
v = Y - X\beta^+ + X\beta^- \\
t = -1 + 2a_1 \\
t = 1 + 2a_2 \\
(v, a_2, a_1) \in Q^{n+2}, \ t \geq 0, \beta^+ \in \mathbb{R}^p_+, \ \beta^- \in \mathbb{R}^p_+.
\]

The \( \sqrt{\text{LASSO}} \) optimization problem can be cast by similarly but without auxiliary variables \( a_1, a_2 \):

\[
\min_{t, \beta^+, \beta^-, v} \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \sum_{j=1}^{p} \beta_j^+ + \beta_j^- \\
v = Y - X\beta^+ + X\beta^- \\
(v, t) \in Q^{n+1}, \beta^+ \in \mathbb{R}^p_+, \ \beta^- \in \mathbb{R}^p_+.
\]

**First order methods.** The new generation of first order methods focus on structured convex problems that can be cast as

\[
\min_w f(A(w) + b) + h(w) \quad \text{or} \quad \min_w h(w) : A(w) + b \in K.
\]

where \( f \) is a smooth function and \( h \) is a structured function that is possibly non-differentiable or with extended values. However it allows for an efficient proximal function to be solved, see [1]. By combining projections and (sub)gradient information these methods construct a sequence of iterates with strong theoretical guarantees. Recently these methods have
been specialized for conic problems which includes lasso and $\sqrt{\text{LASSO}}$. It is well known that several different formulations can be made for the same optimization problem and the particular choice can impact the computational running times substantially. We focus on simple formulations for lasso and $\sqrt{\text{LASSO}}$.

Lasso is cast as

$$\min_w f(A(w) + b) + h(w)$$

where $f(\cdot) = \| \cdot \|^2 / n$, $h(\cdot) = (\lambda / n) \cdot \| \cdot \|_1$, $A = X$, and $b = -Y$. The projection required to be solved on every iteration for a given current point $\beta^k$ is

$$\beta(\beta^k) = \arg \min_{\beta} 2\mathbb{E}[x_i(y_i - x'_i\beta^k)]'\beta + \frac{1}{2} \mu \|\beta - \beta^k\|^2 + \frac{\lambda}{n} \|\beta\|_1.$$  

It follows that the minimization in $\beta$ above is separable and can be solved by soft-thresholding as

$$\beta_j(\beta^k) = \text{sign}(\beta^k_j + 2\mathbb{E}[x_{ij}(y_i - x'_i\beta^k)]/\mu) \max \left\{ \frac{1}{\mu} \left| \beta^k_j + 2\mathbb{E}[x_{ij}(y_i - x'_i\beta^k)]/\mu \right| - \lambda / [n\mu], 0 \right\}.$$  

For $\sqrt{\text{LASSO}}$ the “conic form” is given by

$$\min_w h(w) : A(w) + b \in K.$$  

Letting $Q^{n+1} = \{(z, t) \in \mathbb{R}^n \times \mathbb{R} : t \geq \|z\|\}$ and $h(w) = f(\beta, t) = t/\sqrt{n} + (\lambda / n) \|\beta\|_1$ we have that

$$\min_{\beta,t} \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \|\beta\|_1 : A(\beta, t) + b \in Q^{n+1}$$  

where $b = (-Y', 0)'$ and $A(\beta, t) \mapsto (\beta'X', t')$.

In the associated dual problem, the dual variable $z \in \mathbb{R}^n$ is constrained to be $\|z\| \leq 1/\sqrt{n}$ (the corresponding dual variable associated with $t$ is set to $1/\sqrt{n}$ to obtain a finite dual value). Thus we obtain

$$\max_{\|z\| \leq 1/\sqrt{n}} \inf_{\beta} \frac{\lambda}{n} \|\beta\|_1 + \frac{1}{2} \mu \|\beta - \beta^k\|^2 - z'(Y - X\beta).$$  

Given iterates $\beta^k, z^k$, as in the case of lasso that the minimization in $\beta$ is separable and can be solved by soft-thresholding as

$$\beta_j(\beta^k, z^k) = \text{sign}(\beta^k_j + (X'z^k/\mu)_j) \max \left\{ \frac{1}{\mu} (\beta^k_j + (X'z^k/\mu)_j) - \lambda / [n\mu], 0 \right\}.$$
The dual projection accounts for the constraint $\|z\| \leq 1/\sqrt{n}$ and solves

$$z(\beta^k, z^k) = \arg \min_{\|z\| \leq 1/\sqrt{n}} \frac{\theta_k}{2t_k} \|z - z^k\|^2 + (Y - X\beta^k)'z$$

which yields

$$z(\beta^k, z^k) = \frac{z_k + (t_k/\theta_k)(Y - X\beta^k)}{\|z_k + (t_k/\theta_k)(Y - X\beta^k)\|} \min \left\{ \frac{1}{\sqrt{n}}, \frac{1}{\|z_k + (t_k/\theta_k)(Y - X\beta^k)\|} \right\}.$$ 

Componentwise Search. A common approach to solve unconstrained multivariate optimization problems is to (i) pick a component, (ii) fix all remaining components, (iii) minimize the objective function along the chosen component, and loop steps (i)-(iii) until convergence is achieved. This is particularly attractive in cases where the minimization over a single component can be done very efficiently. Its simple implementation also contributes for the widespread use of this approach.

Consider the following lasso optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i'\beta)^2] + \frac{\lambda}{n} \sum_{j=1}^p \gamma_j |\beta_j|.$$ 

Under standard normalization assumptions we would have $\gamma_j = 1$ and $\mathbb{E}_n[x_{ij}^2] = 1$ for $j = 1, \ldots, p$. Below we describe the rule to set optimally the value of $\beta_j$ given fixed the values of the remaining variables. It is well known that lasso optimization problem has a closed form solution for minimizing a single component.

For a current point $\beta$, let $\beta_{-j} = (\beta_1, \beta_2, \ldots, \beta_{j-1}, 0, \beta_{j+1}, \ldots, \beta_p)'$.

- If $2\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] > \lambda \gamma_j / n$ it follows that the optimal choice for $\beta_j$ is
  $$\beta_j = \left( -2\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] + \lambda \gamma_j / n \right) / \mathbb{E}_n[x_{ij}^2].$$

- If $2\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] < -\lambda \gamma_j / n$ it follows that the optimal choice for $\beta_j$ is
  $$\beta_j = \left( 2\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] - \lambda \gamma_j / n \right) / \mathbb{E}_n[x_{ij}^2].$$

- If $2|\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]| \leq \lambda \gamma_j / n$ we would set $\beta_j = 0$.

This simple method is particularly attractive when the optimal solution is sparse which is typically the case of interest under choices of penalty levels that dominate the noise like $\lambda \geq cn\|S\|_\infty$. 

\(\sqrt{\text{LASSO \ 49}}\)
Despite of the additional square-root, which creates a non-separable criterion function, it turns out that the componentwise minimization for \(\sqrt{\text{LASSO}}\) also has a closed form solution. Consider the following optimization problem:

\[
\min_{\beta \in \mathbb{R}^p} \sqrt{\mathbb{E}_n[(y_i - x_i'\beta)^2]} + \frac{\lambda}{n} \sum_{j=1}^{p} \gamma_j |\beta_j|.
\]

As before, under standard normalization assumptions we would have \(\gamma_j = 1\) and \(\mathbb{E}_n[x_{ij}^2] = 1\) for \(j = 1, \ldots, p\). Below we describe the rule to set optimally the value of \(\beta_j\) given fixed the values of the remaining variables.

- If \(\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] > (\lambda/n)\gamma_j \sqrt{\hat{Q}(\beta_{-j})}\), we have
  \[
  \beta_j = -\frac{\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]}{\mathbb{E}_n[x_{ij}^2]}\gamma_j \sqrt{\hat{Q}(\beta_{-j})} - \frac{\lambda\gamma_j}{\mathbb{E}_n[x_{ij}^2]} \sqrt{\hat{Q}(\beta_{-j})} - \frac{(\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]^2/\mathbb{E}_n[x_{ij}^2])}{\sqrt{n^2 - (\lambda^2\gamma_j^2/\mathbb{E}_n[x_{ij}^2])}}
  \]

- If \(\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] < - (\lambda/n)\gamma_j \sqrt{\hat{Q}(\beta_{-j})}\), we have
  \[
  \beta_j = -\frac{\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]}{\mathbb{E}_n[x_{ij}^2]}\gamma_j \sqrt{\hat{Q}(\beta_{-j})} - \frac{\lambda\gamma_j}{\mathbb{E}_n[x_{ij}^2]} \sqrt{\hat{Q}(\beta_{-j})} - \frac{(\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]^2/\mathbb{E}_n[x_{ij}^2])}{\sqrt{n^2 - (\lambda^2\gamma_j^2/\mathbb{E}_n[x_{ij}^2])}}
  \]

- If \(|\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]| \leq (\lambda/n)\gamma_j \sqrt{\hat{Q}(\beta_{-j})}\), we have \(\beta_j = 0\).

References


