PIVOTAL ESTIMATION OF NONPARAMETRIC FUNCTIONS VIA SQUARE-ROOT LASSO*

By Alexandre Belloni and Victor Chernozhukov and Lie Wang

We proposed a self-tuning $\sqrt{\text{lasso}}$ method for estimating nonparametric regression under non-Gaussian heteroskedastic noise. An attractive feature of the approach is that it does not rely on the knowledge of the scale of the noise, unlike many other ℓ_1-regularized methods. In turn this translates into a robustness property that allows the same penalty level to be appropriate for a variety of design conditions.

Our analysis is based on new identification conditions that allow for repeated regressors. We derive various non-asymptotic performance bounds for $\sqrt{\text{lasso}}$ including prediction norm rate, ℓ_1-rate, ℓ_∞-rate, and sharp sparsity bound. In order to cover heteroskedastic non-Gaussian noise, we rely on moderate deviation theory for self-normalized sums to achieve Gaussian-like results under weak conditions. Moreover, we derive 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. Under mild conditions the rate of convergence of ols post $\sqrt{\text{lasso}}$ is no worse than $\sqrt{\text{lasso}}$ even with a misspecified selected model and possibly better otherwise. We show that the robustness properties of $\sqrt{\text{lasso}}$ also extend to the parametric noiseless case and unbounded variance case. In the first case, $\sqrt{\text{lasso}}$ recovers the true parameter value exactly, in sharp contrast to lasso. In the second case under symmetric disturbances, $\sqrt{\text{lasso}}$ can be applied with similar penalty choices and still achieve near Gaussian rates in several cases, in contrast to lasso which would require a substantially larger penalty level.

Key Words: square-root lasso, high-dimensional sparse regression, imperfect model selection

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

$$(1.1) \qquad y_i = f(z_i) + \sigma \epsilon_i, \quad i = 1, \ldots, n,$$

where y_i's are the outcomes, z_i's are vectors of fixed basic covariates, ϵ_i's are independent disturbances, f is the regression function, and σ is a scaling parameter. The 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 $p > n$, 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 above can be written as $y_i = x_i'\beta_0 + r_i + \sigma\epsilon_i$, where $f_i = f(z_i)$ and $r_i := f_i - x_i'\beta_0$ is the approximation error. The vector β_0 is a solution of an oracle problem that balances bias and variance, and the cardinality of the support of coefficient β_0 is denoted by $s := \|\beta_0\|_0$. It is well known that ordinary least square (ols) 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, ℓ_1-based penalization methods have been playing a central role. The ℓ_1-norm penalized methods have been extensively studied in the high dimensional settings [5, 9, 14, 18, 29, 34, 33]. It was demonstrated that, under appropriate choice of penalty level, the ℓ_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. Importantly, in the context of linear regression, these ℓ_1-regularized problems can be cast as convex optimization problems which make them computationally efficient (polynomial time). We refer to [5, 7, 8, 6, 11, 15, 16, 23, 29] for a more detailed review of the existing literature which has been focusing on the homoskedastic case.

In this paper, we focus on the case of nonparametric regression under non-Gaussian heteroskedastic errors. We propose to use a self-tuning $\sqrt{\text{lasso}}$ which is pivotal with respect to the scaling parameter σ. Such pivotallity is in sharp contrast to many others ℓ_1-regularized methods, for example lasso. The penalization parameter in those methods should majorate the non-negligible spurious correlation between the noise terms and the large number of additional regressors. Typically, such penalty level scales linearly with the unknown scaling parameter σ of the noise. Simple upper bounds for σ can be derived based on the empirical variance of the response variable. However, upper bounds on σ 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 the homoskedastic parametric model studied in [4], the choice of the

penalty parameter becomes pivotal given the covariates and the distribution of the error term. In contrast, in the nonparametric heteroskedastic setting we need to account for the impact of the approximation error and the loadings 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 [4]. We perform a thorough non-asymptotic theoretical analysis of the choice of the penalty parameter.

In order to allow for more general designs we propose two new condition measures. Unlike previous conditions, they are tailored for establishing prediction norm. This is appealing because the rates in the prediction norm is the relevant metric in nonparametric estimation, and can be established under weaker conditions. (For instance, the results for prediction rates remain unaffected if repeated regressors are included.) Nonetheless, the proposed quantities alone cannot yield results in ℓ_p-norm (although they can be used to potentially sharp ℓ_p-norm rates in combination with previous conditions). Interestingly, we also show that the restricted eigenvalue proposed in [5] can be used to bound these new quantities.

The second set of contributions is to derive new finite sample upper bounds for prediction norm rate of convergence, ℓ_1-rate of convergence, ℓ_∞-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.

The third contribution aims to remove the potentially significant bias towards zero introduced by the ℓ_1-norm regularization employed in (2.3). We consider the post model selection estimator that applies ordinary least squares (ols) to the model selected by $\sqrt{\text{lasso}}$. It follows that if the model selection works perfectly then the ols post $\sqrt{\text{lasso}}$ estimator is simply the oracle estimator whose properties are well known. However, perfect model selection might be unlikely for many designs of interest. This is usually the case in a nonparametric setting. Thus, we are also interested in the properties of ols post $\sqrt{\text{lasso}}$ when perfect model selection fails, including cases where the oracle model is not completely selected by $\sqrt{\text{lasso}}$.

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 with penalty choice that does not

depend on the standard deviation of the noise. We develop the necessary modifications on the penalty loadings and derive finite-sample bounds for the case of symmetric noise. For bounded designs the results match the Gaussian-noise rates up to a factor of $(\mathbb{E}_n[\epsilon_i^2])^{1/2}$ which tends to grow slowly in this case. We provide specific bounds to the case of Student's t-distribution with 2 degrees of freedom where $\mathbb{E}_n[\epsilon_i^2] \lesssim_P \log n$.

Notation. In making asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we also allow for $s = s_n \to \infty$. In what follows, all parameter values are indexed by the sample size n, but we omit the index whenever this does not cause confusion. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2-norm is denoted by $\|\cdot\|$, the ℓ_1-norm is denoted by $\|\cdot\|_1$, the ℓ_∞-norm is denoted by $\|\cdot\|_\infty$, and the ℓ_0-norm $\|\cdot\|_0$ denotes the number of non-zero components of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$, and by $|T|$ the cardinality of T. The symbol $\mathrm{E}[\cdot]$ denotes the expectation. We also use standard empirical process notation $\mathbb{E}_n[f(z_i)] := \sum_{i=1}^n f(z_i)/n$ and $\mathbb{G}_n(f(z_i)) := \sum_{i=1}^n (f(z_i) - \mathrm{E}[f(z_i)])/\sqrt{n}$. We also denote $\bar{\mathbf{E}}[\cdot] = \mathbb{E}_n \mathrm{E}[\cdot]$ and the $L^2(\mathbb{P}_n)$-norm by $\|f\|_{\mathbb{P}_n,2} = (\mathbb{E}_n[f_i^2])^{1/2}$. Given covariate values $x_1, \ldots, x_n$, we define the prediction norm of a vector $\delta \in \mathbb{R}^p$ as $\|\delta\|_{2,n} = \{\mathbb{E}_n[(x_i'\delta)^2]\}^{1/2}$, and given values $y_1, \ldots, y_n$ we define $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$. We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$ that does not depend on n (and therefore does not depend on quantities indexed by n like p or s); and $a \lesssim_P b$ to denote $a = O_P(b)$.

2. Nonparametric regression model and Estimators. Consider the nonparametric regression model:

$$(2.1) \quad y_i = f(z_i) + \sigma \epsilon_i, \quad \epsilon_i \sim F_i, \quad \mathrm{E}[\epsilon_i] = 0, \quad \bar{\mathbf{E}}[\epsilon_i^2] = 1, \quad i = 1, \ldots, n,$$

where z_i are vectors of fixed regressors, ϵ_i are independent errors, and σ is the scaling factor of the errors. In order to recover the regression function f we consider linear combinations of the covariates $x_i = P(z_i)$ which are p-vectors of transformation of z_i normalized so that $\mathbb{E}_n[x_{ij}^2] = 1$ $(j = 1, \ldots, p)$.

The goal is to estimate the nonparametric regression function f at the design points, namely the values $f_i = f(z_i)$ for $i = 1, \ldots, n$. In many applications of interest, especially in the nonparametric settings, 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 (2.1). The target coefficients β_0 that we consider solves the following oracle risk minimization

problem:

$$(2.2) \qquad \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(f_i - x_i'\beta)^2] + \frac{\sigma^2 \|\beta\|_0}{n},$$

where the problem above yields an upper bound on the risk of the best k-sparse least squares estimator in the case of homoskedastic Gaussian errors, i.e. the best estimator among all least squares estimators that use k out of p components of x_i to estimate f_i, for $i = 1, \ldots, n$. The solution β_0 of the oracle achieves a balance between the mean square of the approximation error $r_i := f_i - x_i'\beta_0$ and the variance, where the latter is determined by the complexity of the model (number of non-zero components of β_0).We consider the case that the support of the best sparse approximation $T = \text{supp}(\beta_0)$ is unknown.

We call β_0 the oracle target value, $T := \text{support}(\beta_0)$ the oracle model, $s := |T| = \|\beta_0\|_0$ the dimension of the oracle model, and $x_i'\beta_0$ the oracle approximation to f_i. We summarize the previous setting in the following condition.

Condition ASM. *We have data $\{(y_i, z_i) : i = 1, \ldots, n\}$ that for each n obey the regression model (2.1), where y_i are the outcomes, z_i are vectors of fixed regressors, and ϵ_i are i.n.i.d. errors. The vector β_0 is defined by (2.2) where the regressors $x_i := P(z_i)$ are normalized so that $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \ldots, p$.*

2.1. *Heteroskedastic $\sqrt{lasso}$.* In this section we formally define the estimators which are tailored to deal with heteroskedasticity.

We propose to consider the $\sqrt{\text{lasso}}$ estimator defined as

$$(2.3) \qquad \widehat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} + \frac{\lambda}{n} \|\Gamma\beta\|_1,$$

where $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$, $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_p)$, γ_j, $j = 1, \ldots, p$, is a penalty loading. The scaled ℓ_1-penalty allows to sharp adjustments to efficiently deal with heteroskedasticity. Indeed, every penalty loading can be taken equal to 1 in the traditional case of homoskedastic errors[1].

In order to reduce the shrinkage bias intrinsic from $\sqrt{\text{lasso}}$, we consider the post model selection estimator that applies ordinary least squares (ols) to the model $\widehat{T}$ selected by $\sqrt{\text{lasso}}$. Formally, set

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

---

[1]In the heteroskedastic case, if $\{\lambda, \Gamma\}$ are appropriate choices, then $\{\lambda\|\Gamma\|_\infty, I_p\}$ is also an appropriate choice but potentially conservative, i.e. leading to overpenalization. Throughout we assume $\Gamma_{jj} \geq 1$ for $j = 1, \ldots, p$.

and define the ols post $\sqrt{\text{lasso}}$ estimator $\widetilde{\beta}$ as

$$(2.4) \qquad \widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} \quad : \quad \beta_j = 0 \ \ \text{if} \ \ j \in \widehat{T}^c.$$

2.2. *Conditions on the Gram Matrix.* It is known that the Gram matrix $\mathbb{E}_n[x_i x_i']$ plays an important role in the analysis of estimators in this setup. In our case, the smallest eigenvalue of the Gram matrix is 0 if $p > n$ which creates identification problems. Thus, to restore identification, one needs to restrict the type of deviation vectors δ from β_0 that we will consider. Because of the ℓ_1 regularization, it will be important to consider vectors δ that belong to the restricted set $\Delta_{\bar{c}}$ defined as

$$\Delta_{\bar{c}} = \{\delta \in \mathbb{R}^p : \|\Gamma \delta_{T^c}\|_1 \leq \bar{c}\|\Gamma \delta_T\|_1, \delta \neq 0\}, \quad \text{for} \ \ \bar{c} \geq 1.$$

We will state the bounds in terms of the following restricted eigenvalues of the Gram matrix $\mathbb{E}_n[x_i x_i']$:

$$(2.5) \qquad \kappa_{\bar{c}} := \min_{\delta \in \Delta_{\bar{c}}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma \delta_T\|_1}.$$

The restricted eigenvalues can depend on n, T, and Γ, but we suppress the dependence in our notations. The restricted eigenvalues (2.5) are variants of the restricted eigenvalue introduced in Bickel, Ritov and Tsybakov [5].

Next consider the minimal and maximal m-sparse eigenvalues of a matrix M,

$$(2.6)$$

$$\phi_{\min}(m, M) := \min_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|_2^2}, \ \ \text{and} \ \ \phi_{\max}(m, M) := \max_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|_2^2}.$$

Typically we consider minimal and maximal m-sparse eigenvalues associated with the Gram matrix $\mathbb{E}_n[x_i x_i']$, and the rescaled Gram matrix $\Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1}$. For convenience, when the matrix is omitted from the notation we refer to the Gram matrix, namely $\phi_{\min}(m) = \phi_{\min}(m, \mathbb{E}_n[x_i x_i'])$ and $\phi_{\max}(m) = \phi_{\max}(m, \mathbb{E}_n[x_i x_i'])$. These quantities play an important role in the sparsity and post model selection analysis. Moreover, sparse eigenvalues provide a simple sufficient condition to bound restricted eigenvalues. Indeed, following [5], we can bound $\kappa_{\bar{c}}$ from below by

$$\kappa_{\bar{c}} \ \geqslant \ \max_{m \geqslant 0} \ \frac{\sqrt{\phi_{\min}(m)}}{\|\Gamma\|_\infty} \left(1 - \sqrt{\frac{\phi_{\max}(m)}{\phi_{\min}(m)}} \bar{c}\sqrt{s/m}\right).$$

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

$$(2.7) \quad 0 < k \leq \phi_{\min}(m) \leq \phi_{\max}(m) \leq k' < \infty, \ \ \text{for all} \ \ m \leq 4(k'/k)^2 \bar{c}^2 s,$$

then $\kappa_{\bar{c}} \geq \sqrt{\phi_{\min}(4(k'/k)^2 \bar{c}^2 s)}/[2\|\Gamma\|_\infty]$. We note that (2.7) only requires the eigenvalues of certain "small" $(m+s) \times (m+s)$ submatrices of the large $p \times p$ Gram matrix to be bounded from above and below.

For standard arbitrary bounded dictionaries arising in the nonparametric estimations, for example regression splines, orthogonal polynomials, and trigonometric series (see [26]), the following lemma proved in [2] provides primitive conditions under which the sparse eigenvalues well behaved with high probability when the values of x_i, $i = 1, \ldots, n$ were generated randomly.

LEMMA 1 (Sparse eigenvalues, bounded regressors case). *Suppose $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. vectors, such that the population design matrix $\mathrm{E}[\tilde{x}_i \tilde{x}_i']$ has ones on the diagonal, and its $s \log n$-sparse eigenvalues are bounded from above by $\varphi_{\max} < \infty$ and bounded from below by $\varphi_{\min} > 0$. Define x_i as a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij}/(\mathbb{E}_n[\tilde{x}_{ij}^2])^{1/2}$. Suppose that $\max_{1 \leq i \leq n} \|\tilde{x}_i\|_\infty \leq K_n$ a.s., and $K_n^2 s \log^4(n) \log(p \vee n) = o(n\varphi_{\min}^2/\varphi_{\max})$. Then, for any $m \geq 0$ such that $m + s \leq s \log n$, the empirical maximum and minimal m-sparse eigenvalues obey: $\phi_{\max}(m) \leq 4\varphi_{\max}$, and $\phi_{\min}(m) \geq \varphi_{\min}/4$, with probability approaching 1 as $n \to \infty$.*

Other sufficient conditions for (2.7) are provided by [5], [34], and [18]. [5] and others also provide different sets of sufficient primitive conditions for $\kappa_{\bar{c}}$ to be bounded away from zero.

3. Overview of Asymptotic Results and Comparisons under Heteroskedasticity.

3.1. *Rates of Convergence of $\sqrt{lasso}$ and post-$\sqrt{lasso}$.* In this section we formally state the main algorithm to compute the estimators and we provide rates of convergence results under simple primitive conditions. We defer the finite sample analysis under significantly weaker conditions to Section 4.

We propose setting the penalty level as

$$(3.1) \qquad \lambda = (1 + u_n)c\sqrt{n}(\Phi^{-1}(1 - \alpha/2p) + 1 + u_n)$$

and the penalty loadings according to the following iterative algorithm.

ALGORITHM 1 (Estimation of Square-root Lasso Loadings). *Choose $\alpha \in (0, 1)$, $\nu \geq 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations.*

Step 0. Set $k = 0$, λ as defined in (3.1). For $w > (\bar{\mathbf{E}}[\epsilon_i^4])^{1/4}/(\bar{\mathbf{E}}[\epsilon_i^2])^{1/2}$
define $\widehat{\gamma}_{j,0} = w(\mathbb{E}_n[x_{ij}^4])^{1/4}$, $j = 1, \ldots, p$.
Step 1. Compute the $\sqrt{lasso}$ estimator $\widehat{\beta}$ based on the current
penalty loadings $\{\widehat{\gamma}_{j,k}, j = 1, \ldots, p\}$.
Step 2. Set $\widehat{\gamma}_{j,k+1} := 1 \vee \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}/\sqrt{\mathbb{E}_n[(y_i - x_i'\widehat{\beta})^2]}$.
Step 3. If $\max_{1 \leq j \leq p} |\widehat{\gamma}_{j,k} - \widehat{\gamma}_{j,k+1}| \leqslant \nu$ or $k > K$, stop;
otherwise set $k \leftarrow k + 1$ and go to Step 1.

The parameter $1-\alpha$ is a confidence level which guarantees near-oracle performance with probability at least $1-\alpha$; we recommend $\alpha = 0.05/\log n$. The constant $c > 1$ is the slack parameter used in [5]; we recommend $c = 1.01$. The parameter u_n is intended to account for the approximation errors; we recommend $u_n = 0.1/\log n$. The parameter w is pivotal to the scaling parameter σ and its goal is to simply bound the ratio of moments; we recommend $w = 2$ (which covers several interesting distributions). Finally, we recommend iterating the procedure to avoid unnecessary overpenalization since at each iteration more precise estimates of the penalty loadings tend to be achieved. 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 E.

REMARK 1. *Algorithm 1 relies on the $\sqrt{lasso}$ estimator $\widehat{\beta}$. Another possibility is to use the post $\sqrt{lasso}$ estimator $\widetilde{\beta}$. Asymptotically, the analysis would be conceptually very similar.*

The following is a set of simple sufficient conditions which is used to clearly communicate the results.

CONDITION P. *There exist a finite constant $q \geq 6$ such that the noise obeys $\sup_{n \geq 1} \bar{\mathbf{E}}[|\epsilon_i^q|] < \infty$, the covariates obey $\sup_{n \geq 1} \max_{1 \leq j \leq p} \mathbb{E}_n[|x_{ij}^q|] < \infty$, and we have that $\inf_{n \geq 1} \min_{1 \leq j \leq p} \mathbb{E}_n[x_{ij}^2 \mathrm{E}[\epsilon_i^2]] > 0$. Moreover, we have that $\sup_{n \geq 1} \phi_{\max}(s \log n)/\phi_{\min}(s \log n) < \infty$, $s \log(p \vee n) = o(n)$, and $\log p = o(n^{1/3})$.*

Based on this choice of penalty level and loadings, the following corollary summarizes the asymptotic performance of $\sqrt{lasso}$ for commonly used designs.

COROLLARY 1 (Asymptotic performance of $\sqrt{lasso}$). *Suppose Conditions ASM and P hold, let $c > 1$ and $\bar{c} = (c + 1)/(c - 1)$. Let the penalty level λ be set as in (3.1) with $\alpha = 0.05/\log n$, and penalty loadings as in*

Algorithm 1 with $u_n = 0.1/\log n$. Then we have that

$$\|\widehat{\beta} - \beta_0\|_{2,n} \lesssim_P (c_s + \sigma)\sqrt{\frac{s \log p}{n}}, \quad and \quad \|\widehat{\beta} - \beta_0\|_1 \lesssim_P (c_s + \sigma)\sqrt{\frac{s^2 \log p}{n}}.$$

If in addition $\|\mathbb{E}_n[x_i x_i'] - I\|_\infty = o(1/s)$ we have

$$\|\widehat{\beta} - \beta_0\|_\infty \lesssim_P (c_s + \sigma)\sqrt{\frac{\log p}{n}}.$$

The result above establishes that $\sqrt{\text{lasso}}$ achieves the same near oracle rate of convergence of lasso despite of not knowing the scaling parameter σ. The results above allows for heteroskedastic errors with mild restrictions on its moments. It also substantially improve the restrictions on the growth of p relative to n with respect to [4]. We note that the theory allows for any choice of iterations K in Algorithm 1.

The following corollary summarizes the performance of ols post $\sqrt{\text{lasso}}$ under commonly used designs.

COROLLARY 2 (Asymptotic performance of ols post $\sqrt{\text{lasso}}$).   *Under the conditions of Corollary 1 let $\widehat{m} = |\widehat{T} \setminus T|$. We have that*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P c_s + \sigma\sqrt{\frac{s \log p}{n}} \quad and \quad \widehat{m} \lesssim_P s.$$

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 $T \not\subseteq \widehat{T}$. 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 $\widehat{m} = o_P(s)$ and $T \subseteq \widehat{T}$ with probability going to 1 or in the case of perfect model selection,[2] when $T = \widehat{T}$ with probability going to 1. Moreover, under mild conditions, the upper bound for the prediction norm rate of $\sqrt{\text{lasso}}$ is sharp, i.e. in general the rate of convergence cannot be faster than $\sigma\sqrt{\log p}\sqrt{s/n}$. Thus the use of the post model selection estimator leads to a strict improvement in the rate of convergence on these well-behaved models.

---

[2]Results on lasso's model selection performance derived on Wainright [33] can be extended to the $\sqrt{\text{lasso}}$ estimator based on Theorem 3 and 4.

3.2. *A Benchmark: Oracle Projection Estimators under Orthonormal Random Design.* Next we discuss examples of nonparametric estimation. Later, we will compare the results derived here for $\sqrt{\text{lasso}}$ and ols post $\sqrt{\text{lasso}}$ with projection estimators.

Consider the nonparametric model (2.1) where f is a function from $[0, 1]$ to $\mathbb{R}$, $\epsilon_i \sim N(0, 1)$ and $z_i \sim \text{Uniform}(0, 1), i = 1, \ldots, n$. Given a basis $\{P_j(\cdot)\}_{j=1}^{\infty}$ the projection estimator with k terms is defined as

$$\widehat{f}^{(k)}(z) = \sum_{j=1}^{k} \widehat{\theta}_j P_j(z) \text{ where } \widehat{\theta}_j = \mathbb{E}_n[y_i P_j(z_i)] \text{ and } \widehat{\theta}^{(k)} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k, 0, \ldots)'.$$

Projection estimators are particularly appealing in orthonormal designs.

EXAMPLE 1 (Series Approximations in Sobolev Balls). *Let the basis* $\{P_j(\cdot)\}_{j=1}^{\infty}$ *be the trigonometric basis for* $L^2[0, 1]$ *and suppose that* f *belongs to the periodic Sobolev class* $W^{per}(\alpha, L)$, *that is,* $f(0) = f(1)$ *and*

$$f \in W(\alpha, L) = \left\{ f \in [0, 1] \to \mathbb{R} : \begin{array}{l} f^{(\alpha-1)} \text{ is absolutely continuous and} \\ \int_0^1 [f^{(\alpha)}(z)]^2 dz \le L^2 \end{array} \right\}.$$

It follows that the Fourier coefficients $\theta_j = \int_0^1 f(z) P_j(z) dz$ *of* f *satisfy* $\sum_{j=1}^{\infty} |\theta_j| < \infty$ *and* $\theta \in \Theta(\alpha, L) = \{\theta \in \ell^2(\mathbf{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \le L^2/\pi^{2\alpha}\}$ *where* $a_j = j^{\alpha}$ *for even* j *and* $a_j = (j-1)^{\alpha}$ *for odd* j *represents the* L_2-*norm of the* α-*derivative of the* jth *base function,* $\alpha \ge 1$ *and* $L > 0$. *Thus, for each* $z \in [0, 1]$

$$f(z) = \sum_{j=1}^{\infty} \theta_j P_j(z).$$

Now consider the oracle problem of choosing the best s-*dimensional projection/series estimator. This oracle problem solves*

$$\min_{0 \le k \le n} c_k^2 + \sigma^2 \frac{k}{n}.$$

Here c_k^2 *is an upper bound on the approximation error*

$$\bar{\mathbf{E}} \left[\{f_i - \sum_{j=1}^{k} \theta_j P_j(z_i)\}^2 \right]$$

of the projection estimator. By Lemma 12, we have $c_k^2 \le Ck^{-2\alpha}$ *where the constant* C *is uniform in* $f \in W^{per}(\alpha, L)$. *A rate-optimal choice of the number of series terms satisfies* $k = s \le \lfloor Vn^{\frac{1}{2\alpha+1}} \rfloor$, *for some* $V > 0$ *uniformly*

over $f \in W^{per}(\alpha, L)$, and implies an upper bound on the oracle risk given by

$$c_s^2 + \frac{\sigma^2 s}{n} \lesssim \sigma^2 n^{-\frac{2\alpha}{2\alpha+1}}.$$

□

EXAMPLE 2 (p-Rearranged α-Ellipsoids). *Define the set of coefficients*

$$\Theta^S(\alpha, p, L) = \left\{ \theta \in \ell^2(\mathbf{N}) : \begin{array}{l} \exists \ permutation \ \gamma : \{1, \ldots, p\} \to \{1, \ldots, p\} \\ \sum_{j=1}^p j^{2\alpha} \theta_{\gamma(j)}^2 + \sum_{j=p+1}^\infty j^{2\alpha} \theta_j^2 \leq L^2 \end{array} \right\}.$$

We consider functions f such that for some $\theta \in \Theta^S(\alpha, p, L)$ we have for each $z \in [0, 1]$ that

$$f(z) = \sum_{j=1}^\infty \theta_j P_j(z)$$

where $\{P_j(\cdot), j \geq 1\}$ is a bounded orthonormal basis. In this setting, we will consider a sparse series approximation and the associated sparse projection estimator based on a support $\widetilde{T} \subset \{1, \ldots, p\}$ as

$$f^{\widetilde{T}}(z) = \sum_{j \in \widetilde{T}} \theta_j P_j(z), \quad \widehat{f}^{\widetilde{T}}(z) = \sum_{j \in \widetilde{T}} \widehat{\theta}_j P_j(z) \quad where \ \widehat{\theta}_j = \mathbb{E}_n[P_j(z_i) y_i].$$

Thus, the approximation error associated with $f^{\widetilde{T}}$ is

$$c_{\widetilde{T}}^2 = \sum_{j \in \widetilde{T}^c} \theta_j^2 = \sum_{j \in \{1, \ldots, p\} \setminus \widetilde{T}} \theta_j^2 + \sum_{j \geq p+1} \theta_j^2 \leq \sum_{j \in \{1, \ldots, p\} \setminus \widetilde{T}} \theta_j^2 + C p^{-2\alpha}.$$

The class of p-rearranged α-ellipsoids reduces significantly the relevance of the order of the basis. In this case the oracle chooses the best s-dimensional projection/series with support $T = \{\gamma_f(1), \ldots, \gamma_f(s)\} \subset \{1, \ldots, p\}$ where γ_f is a permutation that makes the sequence $\{|\theta_{\gamma_f(j)}|\}_{j=1}^p$ non-increasing. In particular, this oracle weakly improves upon the conventional series estimator described in Example 1 since

$$\sum_{j=s+1}^p \theta_j^2 \geq \sum_{j \in \{1, \ldots, p\} \setminus T} \theta_j^2.$$

In general, the rate-optimal choice of the number of series terms is at least as good as in Example 1, $|T| = s \leq \lfloor V n^{\frac{1}{2\alpha+1}} \rfloor$, which implies an upper bound on the oracle risk given by

$$c_s^2 + \sigma^2 \frac{s}{n} \lesssim \sigma^2 n^{-\frac{\alpha}{2\alpha+1}}.$$

However, in many cases the sparse approximation can improve substantially over the standard series approximation. For example, suppose that Fourier coefficients feature the following pattern $\theta_j = 0$ for $j \leq j_0$ and $|\theta_j| \leq Kj^{-a}$ for $j > j_0$. In this case, the standard series approximation based on the first $k \leq j_0$ terms, $\sum_{j=1}^{k} \theta_j P_j(z)$, fails to provide any predictive power for $f(z)$, and the corresponding standard series estimator based on k terms therefore also fails completely. On the other hand, series approximation based on $k > j_0$ terms carry unnecessary j_0 terms which increase the variance of the series estimator. For instance, if $\theta_{n+1} = 1$ and $\theta_j = 0$ for $j \neq n + 1$, the standard series estimator fails to be consistent. In contrast, the sparse series approximation avoids the first unnecessary n term to achieve consistency. □

REMARK 2 (Comparison between $\sqrt{\text{lasso}}$ and Oracle Projection Estimators under orthogonal random design). *Consider the case where the regression function f belongs to the Sobolev class $W(\alpha, L)$, $\alpha \geq 1$, and we have an orthonormal random design. Example 1 yields that the rate-optimal choice for the size of the support of β_0 is $s \lesssim n^{1/[2\alpha+1]}$. Based on Lemma 12 we have that the oracle projection estimator achieves*

$$\|\widehat{\theta}^{(s)} - \beta_0\| \lesssim_P \sigma\sqrt{s/n} \lesssim n^{-\alpha/[2\alpha+1]}.$$

Under this random design, and mild regularities conditions (see Corollary 1), without knowing the exact support, $\sqrt{\text{lasso}}$ achieves

$$\|\widehat{\beta} - \beta_0\| \lesssim_P (\sigma + c_s)\sqrt{s \log p/n} \lesssim n^{-\alpha/[2\alpha+1]}\sqrt{\log p}.$$

However, in the case of a sparse model in which the first components are no longer relevant, like in the p-rearranged α-ellipsoids, the adaptivity of $\sqrt{\text{lasso}}$ allows it to preserve its rate while the oracle series projection estimator is not consistent.

4. Finite-sample analysis of $\sqrt{\text{lasso}}$. Next we establish several finite-sample results regarding the $\sqrt{\text{lasso}}$ estimator. Importantly, these results are based on new conditions on the design matrix. Such conditions are invariant to the introduction of repeated regressors and well behaved if the restricted eigenvalue discussed in Section 2.2 is well behaved.

We highlight that most of the analysis in this section is pure geometric. That 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

$$\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty, \quad \text{where} \quad \widetilde{S} = \mathbb{E}_n[x_i(\sigma\epsilon_i + r_i)]/\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}$$

is the score of $\sqrt{\widehat{Q}}$ at β_0. Therefore, by choosing λ and Γ such that the event above holds with high probability (as discussed in Section 4.5) the stated results hold with high probability.

4.1. *New Identification Conditions.* In Section 2.2 we discussed typical high-level and primitive conditions on the design matrix $\mathbb{E}_n[x_i x_i']$ used in the recent literature [5]. Although previous proposed quantities like restrictive eigenvalues seem appropriate to the development of rates of convergence in ℓ_p-norms, at least in some designs of interest, they still have a gap for establishing the rate of convergence in the prediction norm.

In an attempt to (at least partially) fill this gap we propose the following new quantities

(4.1)

$$
\varrho_{\bar{c}} := \sup_{\substack{\delta \in \Delta_{\bar{c}}, \ \|\delta\|_{2,n} > 0 \\ \|\Gamma(\delta + \beta_0)\|_1 \le \bar{c}\|\Gamma\beta_0\|_1}} \frac{|\widetilde{S}'\delta|}{\|\delta\|_{2,n}} \quad \text{and} \quad \bar{\kappa} := \inf_{\|\Gamma\delta_{T^c}\|_1 < \|\Gamma\delta_T\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1}.
$$

These quantities depend on n, T, and Γ; in what follows, we suppress this dependence whenever this is convenient.

The quantity $\bar{\kappa}$ in (4.1) is motivated by the original restricted eigenvalue (2.5) conditions imposed in Bickel, Ritov and Tsybakov [5] and the compatibility condition of van de Geer [28], and van de Geer and Bülhmann [30]. Either condition implies that $\bar{\kappa} > 0$ holds, and that is what ultimately is used to derive rates. Indeed, in the case of the compatibility condition in [28] we can take $\nu(T) = 0$ which is weaker than required $\nu(T) > 0$ in [28]. Thus (4.1) is an interesting condition since it was shown in [5] and [30] that the RE and the compatibility assumptions are relatively weak and are implied by many other assumptions in the literature.

The control of $\varrho_{\bar{c}}$ also plays a critical role in our analysis. In our view it is a novel concept since $\varrho_{\bar{c}}$ depends not only on the design but also on the error and approximation terms. Fundamentally, it can be controlled via empirical process techniques based on entropy functions since the vectors δ are required to be in the restricted set $\Delta_{\bar{c}}$ and to have an ℓ_1-norm not much larger than $\|\Gamma\beta_0\|_1$.

It is interesting to note that the restricted eigenvalue condition of Bickel, Ritov and Tsybakov [5] yields a possible way to control both $\bar{\kappa}$ and $\varrho_{\bar{c}}$, namely,

$$
\bar{\kappa} := \inf_{\|\Gamma\delta_{T^c}\|_1 < \|\Gamma\delta_T\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} \ge \min_{\delta \in \Delta_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1} \ge \min_{\delta \in \Delta_{\bar{c}}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1} = \kappa_{\bar{c}},
$$

$$
\varrho_{\bar{c}} \le \sup_{\delta \in \Delta_{\bar{c}}} \frac{\|\Gamma^{-1}\widetilde{S}\|_\infty \|\Gamma\delta\|_1}{\|\delta\|_{2,n}} \le \sup_{\delta \in \Delta_{\bar{c}}} \frac{\|\Gamma^{-1}\widetilde{S}\|_\infty (1 + \bar{c})\|\Gamma\delta_T\|_1}{\|\delta\|_{2,n}} \le \frac{(1 + \bar{c})\sqrt{s}}{\kappa_{\bar{c}}} \|\Gamma^{-1}\widetilde{S}\|_\infty,
$$

as formally stated in the results below .

Thus, an analysis based on the quantities $\varrho_{\bar{c}}$ and $\bar{\kappa}$ will be more general than the one relying on (2.5) only. Notably, the quantities $\bar{\kappa}$ and $\varrho_{\bar{c}}$ can be well behaved even in the presence of repeated regressors.

LEMMA 2. *Assume that Condition ASM holds. We have $\bar{\kappa} \geq \kappa_1$. If $|T| = 1$ we have that $\bar{\kappa} \geq 1/\|\Gamma\|_\infty$. Moreover, if copies of regressors are included with the same corresponding penalty loadings, we have that $\bar{\kappa}$ does not change.*

LEMMA 3. *Assume that Condition ASM holds. We have $\varrho_{\bar{c}} \leq (1 + \bar{c})\sqrt{s}\|\Gamma^{-1}\widetilde{S}\|_\infty/\kappa_{\bar{c}}$. We have that $\varrho_{\bar{c}}$ does not change if we include copies regressors.*

We close this section with the result establishing that the $\sqrt{\text{lasso}}$ estimator satisfies the two constraints in the definition of $\varrho_{\bar{c}}$ provided the penalty level λ is set appropriately. That encompass the usual restricted set $\Delta_{\bar{c}}$ and an additional condition on the rescaled ℓ_1-norm of the estimator.

LEMMA 4. *Assume that for some $c > 1$ we have $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, then we have for $\bar{c} = (c+1)/(c-1)$ that*

$$(4.2) \qquad \|\Gamma\widehat{\beta}_{T^c}\|_1 \leq \bar{c}\|\Gamma(\widehat{\beta}_T - \beta_0)\|_1 \quad and \quad \|\Gamma\widehat{\beta}\|_1 \leq \bar{c}\|\Gamma\beta_0\|_1.$$

REMARK 3. *The quantities above are particularly suitable for the analysis based on the criterion function conducted in this work. Another potential interesting measure which is tailored for an analysis based on first order conditions is*

$$\upsilon_{\bar{c}} := \min_{\delta \in \Delta_{\bar{c}}} \frac{\|\mathbb{E}_n[x_i x_i']\delta\|_\infty}{\|\delta\|_{2,n}}$$

which will also be invariant if repeated regressors are included.

REMARK 4. *Although we apply these definitions to $\sqrt{lasso}$ we note that they also apply to lasso and other ℓ_1-penalized estimators. A natural generalization of $\varrho_{\bar{c}}$ to other penalized estimators would replace $\widetilde{S}$ with $\nabla\widehat{Q}(\beta_0)$ in its definition.*

4.2. *Finite-sample bounds on rates.* We start establishing a finite-sample bound for the prediction norm for the $\sqrt{\text{lasso}}$ estimator. We note that this bound is established under heteroskedasticity, without knowledge of the scaling parameter σ, and under the weak design conditions described in Section 4.1.

THEOREM 1 (Finite Sample Bounds on Estimation Error). *Under Condition ASM, let $c > 1$, $\bar{c} = (c+1)/(c-1)$, and suppose that λ obeys the growth restriction $\bar{\rho} := \lambda\sqrt{s}/[n\bar{\kappa}] < 1$. If $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, then*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}\frac{(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2}.$$

We recall that the choice of λ does not depend on the scaling parameter σ. The impact of σ in the bound above comes through the factor

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

Thus, this result leads to the same rate of convergence as in the case of the lasso estimator that knows σ since $\mathbb{E}_n[\epsilon_i^2]$ concentrates around one under (2.1) and the law of large numbers.

The analysis of $\sqrt{\text{lasso}}$ raises several different issues from that of lasso, and so the proof of Theorem 1 is involved. In particular, we need to invoke the additional growth restriction $\bar{\rho} < 1$, which is not present in the lasso analysis that treats σ as known. This is required because the introduction of the square-root removes the quadratic growth which would eventually dominates the ℓ_1 penalty for large enough deviations from β_0. This condition ensures that the penalty its not too large so identification of β_0 is still possible. Note however that when this side condition fails and σ is bounded away from zero, lasso is not guaranteed to be consistent since its rate of convergence is typically given by $\sigma\lambda\sqrt{s}/[n\kappa_{\bar{c}}]$.

Also, the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ accounts for the approximation errors $r_1, \ldots, r_n$. That has two implications. First, the impact of c_s on the estimation of β_0 is diminished by a factor of $(\varrho_{\bar{c}} + \bar{\rho})/(1 - \bar{\rho}^2)$. Second, despite of the approximation errors, we have $\widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$. This is in contrast to the analysis that relied on $\lambda \geq cn\|\mathbb{E}_n[\epsilon_i x_i]\|_\infty$ instead, see [5, 2]. We build on the latter to establish ℓ_1-rate and ℓ_∞-rate of convergence.

THEOREM 2 (ℓ_1-rate of convergence). *Under Condition ASM, if $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, for $c > 1$ and $\bar{c} := (c+1)/(c-1)$, then*

$$\|\Gamma(\widehat{\beta} - \beta_0)\|_1 \leq (1 + \bar{c})\sqrt{s}\|\widehat{\beta} - \beta_0\|_{2,n}/\kappa_{\bar{c}}.$$

Moreover, if $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have

$$\|\Gamma(\widehat{\beta} - \beta_0)\|_1 \leq \frac{2(1 + \bar{c})\sqrt{s}}{\kappa_{\bar{c}}}\sqrt{\widehat{Q}(\beta_0)}\frac{(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2}.$$

The results above highlight that, in general, $\bar{\kappa}$ alone is not suitable to bound ℓ_1 and ℓ_2 rates of convergence. This is expected since repeated regressors are allowed in the design.

THEOREM 3 (ℓ_∞-rate of convergence).   *Under Condition ASM, if $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, for $c > 1$ and $\bar{c} = (c+1)/(c-1)$, then for $\digamma = \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\|_\infty$ we have*

$$\frac{\|\Gamma^{-1}(\widehat{\beta} - \beta_0)\|_\infty}{\sqrt{\widehat{Q}(\beta_0)}} \leq \frac{(1+c)\lambda}{cn} + \frac{\lambda^2}{n^2}\frac{\sqrt{s}}{\bar{\kappa}}\frac{\|\widehat{\beta} - \beta_0\|_{2,n}}{\sqrt{\widehat{Q}(\beta_0)}} + \digamma\frac{\|\Gamma(\widehat{\beta} - \beta_0)\|_1}{\sqrt{\widehat{Q}(\beta_0)}}.$$

Moreover, if $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$ we have

$$\frac{\|\Gamma^{-1}(\widehat{\beta} - \beta_0)\|_\infty}{\sqrt{\widehat{Q}(\beta_0)}} \leq \frac{(1+c)\lambda}{cn} + \frac{2\lambda\bar{\rho}}{n}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2} + 2(1+\bar{c})\digamma\frac{\sqrt{s}}{\kappa_{\bar{c}}}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2}.$$

The ℓ_∞-rate is bounded based on the prediction norm and the ℓ_1-rate of convergence. Since we have $\|\cdot\|_\infty \leq \|\cdot\|_1$, the result is meaningful for nearly orthogonal designs so that $\|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\|_\infty$ is small. In fact, near orthogonality also allows to bound the restricted eigenvalue $\kappa_{\bar{c}}$ from below. In the homoskedastic case for Lasso (which corresponds to $\Gamma = I$) [5] and [15] established that if for some $u \geq 1$ we have $\|\mathbb{E}_n[x_i x_i'] - I\|_\infty \leq 1/(u(1+\bar{c})s)$ then $\kappa_{\bar{c}} \geq \sqrt{1 - 1/u}$. In that case, the first term determines the rate of convergence in the ℓ_∞-norm.

We close this subsection establishing relative finite-sample bound on the estimation of $\widehat{Q}(\beta_0)$ based on $\widehat{Q}(\widehat{\beta})$ under the assumptions of Theorem 1.

THEOREM 4 (Estimation of $\widehat{Q}(\beta_0)$).   *Under Condition ASM, if $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ and $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, for $c > 1$ and $\bar{c} := (c+1)/(c-1)$ we have*

$$-\varrho_{\bar{c}}\|\widehat{\beta} - \beta_0\|_{2,n} \leq \quad \sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \quad \leq \bar{\rho}\|\widehat{\beta} - \beta_0\|_{2,n}.$$

Moreover, if $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$ we have

$$-2\varrho_{\bar{c}}\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2} \leq \quad \sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \quad \leq 2\bar{\rho}\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2}.$$

Thus, under the mild condition $\varrho_{\bar{c}} + \bar{\rho} = o(1)$, Theorem 4 establishes that

$$\sqrt{\widehat{Q}(\widehat{\beta})} = (1 + o(1))\sqrt{\widehat{Q}(\beta_0)}.$$

The quantity $\widehat{Q}(\widehat{\beta})$ is particularly relevant in the analysis of $\sqrt{\text{lasso}}$ since it appears in the first-order condition which is the key to establish sparsity properties.

4.3. *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 optimality conditions. We start with a technical lemma.

LEMMA 5 (Relating Sparsity and Prediction Norm). *Under Condition ASM, let $\widehat{T} = \text{supp}(\widehat{\beta})$ and $\widehat{m} = |\widehat{T} \setminus T|$. For any $\lambda > 0$ we have*

$$\frac{\lambda}{n}\sqrt{\widehat{Q}(\widehat{\beta})}\sqrt{|\widehat{T}|} \leq \sqrt{|\widehat{T}|}\|\Gamma^{-1}\widetilde{S}\|_{\infty}\sqrt{\widehat{Q}(\beta_0)} + \sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}.$$

The proof of the above lemma relies 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 technical lemma above and Theorem 4, we establish the following result.

THEOREM 5 (Lower Bound on Prediction Norm). *Under Condition ASM, $\widehat{T} = \text{supp}(\widehat{\beta})$ and $\widehat{m} = |\widehat{T} \setminus T|$, if $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_{\infty}$, $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, where $c > 1$ and $\bar{c} = (c+1)/(c-1)$, we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \geq \frac{\lambda\sqrt{|\widehat{T}|}\,\sqrt{\widehat{Q}(\beta_0)}}{n\sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})}}\left(1 - \frac{1}{c} - \frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2}\right).$$

It is interesting to contrast the lower bound on the prediction norm above with the corresponding lower bound for lasso. In the case of lasso, as derived in [16], the lower bound does not have the term $\sqrt{\widehat{Q}(\beta_0)}$ since the impact of the scaling parameter σ is accounted in the penalty level λ. Thus, under Condition ASM and σ bounded away from zero and above, the lower bounds for lasso and $\sqrt{\text{lasso}}$ are very close.

Next we proceed to bound the size of the selected support $\widehat{T} = \text{supp}(\widehat{\beta})$ for the $\sqrt{\text{lasso}}$ estimator relative to the size s of the support of the oracle estimator β_0.

THEOREM 6 (Sparsity bound for $\sqrt{\text{lasso}}$). *Under Condition ASM, $\widehat{T} = \text{supp}(\widehat{\beta})$, and $\widehat{m} := |\widehat{T} \setminus T|$. If $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1/\sqrt{2}$, and $\frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}}+\bar{\rho})}{1-\bar{\rho}^2} \leq 1/\bar{c}$, for $c > 1$ and $\bar{c} = (c+1)/(c-1)$, we have*

$$|\widehat{T}| \leq 4\bar{c}^2 \left(\frac{n}{\lambda}\frac{\varrho_{\bar{c}}+\bar{\rho}}{1-\bar{\rho}^2}\right)^2 \phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})$$

Moreover, if $\kappa_{\bar{c}} > 0$ we have

$$\widehat{m} \leq s \cdot (4\bar{c}^2/\kappa_{\bar{c}})^2 \min_{m \in \mathcal{M}} \phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > s\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1}) \cdot 2(4\bar{c}^2/\kappa_{\bar{c}})^2\}$.

The slightly more stringent side condition ensures that the right hand side of the bound in Theorem 5 is positive. Asymptotically, mild conditions, for example the design condition that $\phi_{\max}(s \log n)/\phi_{\min}(s \log n) \lesssim 1$, the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ 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

$$\widehat{m} \lesssim s.$$

REMARK 5. *The first sparsity result in the theorem above relates to the prediction norm rate of convergence, under conditions of Theorem 1*

$$(4.3) \qquad \left(\frac{n}{\lambda}\frac{\|\widehat{\beta} - \beta_0\|_{2,n}}{2\sqrt{\widehat{Q}(\beta_0)}}\right)^2 \leq \left(\frac{n}{\lambda}\frac{\varrho_{\bar{c}}+\bar{\rho}}{1-\bar{\rho}^2}\right)^2.$$

Typically, the term on the right hand side of (4.3) will be of the order of s. This can be the case even if $\kappa_{\bar{c}} = 0$. For instance, well behaved designs discussed in Lemma 1 with a single repeated regressor.

REMARK 6. *Consider the case that $f(z) = 1$ and p repeated regressors $x_i = (1, \ldots, 1)'$ are used (which allows us to set $\Gamma = I$). In this setting there is a sparse solution $\sqrt{\text{lasso}}$ but also there is a solution which has p nonzero regressors. Nonetheless, the prediction norm can be well behaved since it is invariant under repeated regressors, $\bar{\kappa} = 1$ and $\varrho_{\bar{c}} \leq \mathbb{E}_n[\epsilon_i] \lesssim_P 1/\sqrt{n}$. Thus, the sparsity bound above will become trivial not because of the prediction norm rate but because of the maximum sparse eigenvalue. Indeed, in this case $\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1}) = m + 1$ and the set $\mathcal{M}$ becomes empty leading to the trivial bound $\widehat{m} \leq p$.*

4.4. *Finite-sample bounds on the estimation error of ols post* $\sqrt{lasso}$. Based on the model selected by $\sqrt{lasso}$ estimator, $\widehat{T} := \text{supp}(\widehat{\beta})$, 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, we are more interested on the case when model selection does not work perfectly, as occurs for many designs in applications.

The following theorem establishes bounds on the prediction error of the ols post $\sqrt{lasso}$ estimator. The analysis accounts for the data-driven choice of components and for the possibly having a misspecified selected model (i.e. $T \not\subseteq \widehat{T}$). In what follows we let $\vartheta := \max_{j=1,\ldots,p} \mathbb{E}_n[x_{ij}^2 \text{E}[\epsilon_i^2]]$.

THEOREM 7 (Performance of ols post $\sqrt{lasso}$). *Under Condition ASM, let* $\widehat{T} = \text{supp}(\widehat{\beta})$ *denote the support selected by* $\sqrt{lasso}$, $\widehat{m} = |\widehat{T}|$. *Then we have that the post-*$\sqrt{lasso}$ *estimator satisfies for any* $C \geq 1$, *with probability at least* $1 - 1/C^2 - 1/[9C^2 \log p]$, *we have*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq C\sigma\sqrt{\frac{\vartheta}{\phi_{\min}(s)}\frac{s}{n}} + c_s + 24C\sigma\sqrt{\frac{\widehat{m}\log p}{n\phi_{\min}(\widehat{m})}}\sqrt{\vartheta \vee \max_{j=1,\ldots,p}\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} + c_{\widehat{T}}$$

where $c_{\widehat{T}} = \min_{\beta \in \mathbb{R}^p}\sqrt{\mathbb{E}_n[(f_i - x_i'\beta_{\widehat{T}})^2]}$. *Moreover, if* $\lambda/n \geq c\|\Gamma^{-1}\tilde{S}\|_\infty$ *for* $c > 1$, $\bar{c} = (c+1)/(c-1)$, *and* $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, *then we have*

$$c_{\widehat{T}} = \min_{\beta \in \mathbb{R}^p}\sqrt{\mathbb{E}_n[(f_i - x_i'\beta_{\widehat{T}})^2]} \leq c_s + 2\sqrt{\widehat{Q}(\beta_0)}\frac{(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2}.$$

The analysis builds upon the sparsity and prediction rate of the $\sqrt{lasso}$ estimator, and on a data-dependent empirical process inequality derived in [3]. The heteroskedasticity of the noise is bounded through the factor $\vartheta = \max_{j=1,\ldots,p}\mathbb{E}_n[x_{ij}^2\text{E}[\epsilon_i^2]]$ and the random term $\max_{1\leq j\leq p}\mathbb{E}_n[x_{ij}^2\epsilon_i^2]$.

REMARK 7. *We note that the random term in the bound above can be controlled in a variety of ways. For example, if the fourth moment of the regressors and noise are uniformly bounded we have* $\max_{j=1,\ldots,p}\mathbb{E}_n[x_{ij}^2\epsilon_i^2] \leq (\mathbb{E}_n[\epsilon_i^4])^{1/2}\max_{j=1,\ldots,p}(\mathbb{E}_n[x_{ij}^4])^{1/2} \lesssim_P 1$. *Alternatively, under other moment conditions and* $\log p = o(n)$ *we have* $\max_{j=1,\ldots,p}\mathbb{E}_n[x_{ij}^2\epsilon_i^2] \leq \vartheta + o_P(1)$. *In the homoskedastic case,* $\text{E}[\epsilon_i^2] = 1$ *for all* $i = 1, \ldots, n$, *we have that* $\vartheta = 1$.

4.5. *Penalty Level and Loadings for* $\sqrt{lasso}$. Here we analyze the data-driven choice for the penalty level and loadings proposed in Algorithm 1

which are pivotal with respect the scaling parameter σ. Our focus is on establishing that λ/n dominates the rescaled score, namely

$$(4.4) \qquad \lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty, \quad \text{where} \ c > 1,$$

which implies that $\widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$, $\bar{c} = (c+1)/(c-1)$, so that the results in the previous sections hold. We note that the principle of setting λ/n to dominate the score of the criterion function is motivated by [5]'s choice of penalty level for lasso under homoskedasticity and known σ. Here, in order to account for heteroskedasticity the penalty level λ/n needs to majorate the score rescaled by the penalty loadings.

REMARK 8.   *In the parametric case, $r_i = 0$, $i = 1, \ldots, n$, the score does not depend on σ nor β_0. Under the homoskedastic Gaussian assumption, namely $F_i = \Phi$ and $\Gamma = I$, the score is in fact completely pivotal conditional on the covariates. This means that in principle we know the distribution of $\|\Gamma^{-1}\widetilde{S}\|_\infty$, or at least we can compute it by simulation. Therefore the choice of λ can be directly made by the quantiles of the $\|\Gamma^{-1}\widetilde{S}\|_\infty$, see [4].*

In order to achieve Gaussian-like behavior under heteroskedastic non-Gaussian disturbances we have to rely on certain conditions on the moment of the noise, the growth of p relative to n, and also consider α to be either bounded away from zero or approaches zero not too rapidly. In this section we focus on the following set of conditions.

CONDITION D. *There exist a finite constant $q > 4$ such that the disturbance obeys $\sup_{n \geq 1}\bar{\mathbf{E}}[|\epsilon_i|^q] < \infty$, and the covariates obey $\sup_{n \geq 1} \max_{1 \leq j \leq p} \mathbb{E}_n[|x_{ij}|^q] < \infty$.*

CONDITION R.   *Let $w_n = \left(\alpha^{-1}\log n C_q \bar{\mathbf{E}}[|\epsilon_i|^{q\vee4}]\right)^{1/q}/n^{1/4} < 1/2$, and set u_n such that $u_n/[1 + u_n] \geq w_n$, $u_n \leq 1/2$. Moreover, for $1 \leq \ell_n \to \infty$, assume that*

$$n^{1/6}/\ell_n \geq (\Phi^{-1}(1 - \alpha/2p) + 1)\max_{1 \leq j \leq p}(\mathbb{E}_n[|x_{ij}^3|\mathrm{E}[|\epsilon_i^3|]])^{1/3}/(\mathbb{E}_n[x_{ij}^2\mathrm{E}[\epsilon_i^2]])^{1/2}.$$

In the following theorem we provide sufficient conditions for the validity of the penalty level and loadings proposed. For convenience, we use the notation that $\widehat{\Gamma}_k = \mathrm{diag}(\widehat{\gamma}_{1,k}, \ldots, \widehat{\gamma}_{p,k})$ and $\Gamma^* = \mathrm{diag}(\gamma_1^*, \ldots, \gamma_p^*)$ where $\gamma_j^* = 1 \vee \sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}/\sqrt{\mathbb{E}_n[\epsilon_i^2]}$, $j = 1, \ldots, p$.

THEOREM 8.   *Suppose that Conditions ASM, D and R hold. Consider the choice of penalty level λ in (3.1) and penalty loadings Γ_k, $k \geq 0$, in*

Algorithm 1. For $k = 0$ we have that

$$P\left(\frac{\lambda}{n} \geq c\|\widehat{\Gamma}_0^{-1}\widetilde{S}\|_\infty\right) \leq 1 - \alpha\left(1 + \frac{A}{\ell_n^3} + \frac{3}{\log n}\right) - \frac{4(1 + u_n)\bar{\mathbf{E}}[|\epsilon_i|^q]}{u_n \ n^{1-[2/q]}}$$

$$- \frac{C_q\bar{\mathbf{E}}[|\epsilon_i|^{q\vee 8}]}{(w^4 - \bar{\mathbf{E}}[\epsilon_i^4])^{q/4}n^{q/8}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon_i|^q]}{n^{1\wedge(q/4-1)}(w^4 - \bar{\mathbf{E}}[\epsilon_i^4])^{q/4}}.$$

Moreover, conditioned on $\lambda/n \geq c\|\widehat{\Gamma}_0^{-1}\widetilde{S}\|_\infty$, provided

$$2 \max_{1 \leq i \leq n} \|x_i\|_\infty \left(2\sqrt{\widehat{Q}(\beta_0)} \max_{\Gamma = \widehat{\Gamma}^0, \Gamma^*} \left\{\frac{\varrho_{\bar{c}}(\Gamma) + \bar{\rho}(\Gamma)}{1 - \bar{\rho}^2(\Gamma)}\right\} + c_s\right) \leq \sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]}(\sqrt{1 + u_n} - 1),$$

we have $\lambda/n \geq c\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty$ for all $k \geq 1$.

The main insight of the analysis is the use of the theory of moderate deviation for self normalized sums, [13] and [10]. The growth condition depends on the number of bounded moments q of regressors and of the noise term. Under condition D and α fixed, condition R is satisfied for n sufficiently large if $\log p = o(n^{1/3})$. This is asymptotically less restrictive that the condition $\log p \leq (q - 2)\log n$ required in [4]. However, condition D is more stringent than some conditions in [4] thus neither set of condition dominates the other.

Under conditions on the growth of p relative to n, Theorem 8 establishes the validity of the penalty level and loadings in Algorithm 1. It also shows that many nice properties of the penalty level in the homoskedastic Gaussian case continue to hold in many non-Gaussian settings. The following corollary summarizes the asymptotic behavior of the penalty choices.

COROLLARY 3. *Suppose that Conditions ASM, D and R hold, and penalty level λ is chosen as (3.1) and the loadings $\widehat{\Gamma}_k$ by Algorithm 1. If $\log p = o(n^{1/3})$, $1/\kappa_{2\bar{c}} \lesssim 1$, and $\max_{1 \leq i \leq n} \|x_i\|_\infty(c_s + \sqrt{(s\log p)/n}) = o(1)$, then there exists $u_n = o(1)$ such that for every $k \geq 0$*

$$P(\lambda/n \geq c\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty) \geq 1 - \alpha(1 + o(1)), \quad \|\widehat{\Gamma}_k\|_\infty \lesssim_P 1$$

and $\lambda \leq (1 + o(1))c\sqrt{2n\log(p/\alpha)}$.

4.6. *Extreme cases.* In this section we show that the robustness advantage of $\sqrt{\text{lasso}}$ extends to two extreme cases. Such robustness arises because the score is normalized by $\sqrt{\widehat{Q}(\beta_0)}$ avoiding the dependence of σ in the penalty level. This self-normalization allows for similar choices of λ to be valid in many more settings.

4.6.1. *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 $\widehat{Q}(\beta_0) = 0$, allows for exact recovery under less stringent restrictions.

THEOREM 9 (Exact recovery for the parametric noiseless case). *Under Condition ASM, let $\sigma = 0$ and $c_s = 0$. Suppose that $\lambda > 0$ obeys the growth restriction $\lambda\sqrt{s} < n\bar{\kappa}$. Then we have that $\|\widehat{\beta} - \beta_0\|_{2,n} = 0$. Moreover, if $\kappa_1 > 0$, we have $\widehat{\beta} = \beta_0$.*

REMARK 9. *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, $\sqrt{lasso}$ intrinsically adjusts the penalty λ by a factor of $\sqrt{\widehat{Q}(\widehat{\beta})}$. Under mild conditions Theorem 4 ensures that $\sqrt{\widehat{Q}(\widehat{\beta})} = \sqrt{\widehat{Q}(\beta_0)} = 0$ which allows for the perfect recovery. Also note that the lower bound derived in Theorem 5 becomes trivially zero.*

4.6.2. *Nonparametric unbounded variance.* Next we turn to the unbounded variance case. We note that the theory developed in Section 4 does not rely on the assumption that $\bar{\mathbf{E}}[\epsilon_i^2] = 1$. In particular, Theorem 1 relies only on the choice of penalty level and penalty loadings to satisfy the assumed condition $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$. Under symmetric errors we further exploit the self-normalized theory to develop a choice of penalty level and loadings,

$$(4.5) \qquad \lambda = (1 + u_n)c\sqrt{n}(1 + \sqrt{2\log(2p/\alpha)}) \quad \text{and} \quad \gamma_j = \max_{1 \leq i \leq n} |x_{ij}|,$$

where as before we typically can take $u_n = o(1)$.

THEOREM 10 (Bounds on the $\sqrt{lasso}$ prediction norm for symmetric errors). *Consider a nonparametric regression model with data $\{(y_i, z_i) : i = 1, \ldots, n\}$, $y_i = f(z_i) + \epsilon_i$, $x_i = P(z_i)$ such that $\mathbb{E}_n[x_{ij}^2] = 1$ $(j = 1, \ldots, p)$, ϵ_i's are independent symmetric errors, and β_0 defined as any solution to (2.2). Let the penalty level and loadings as in (4.5) where u_n is such that $P(\mathbb{E}_n[\sigma\epsilon_i^2] > (1 + u_n)\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]) \leq \gamma$. Moreover let $P(\mathbb{E}_n[\epsilon_i^2] \leq 1) \leq \eta$. If $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, then with probability at least $1 - \alpha - \gamma - \eta$ we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \frac{2(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2} \left(c_s + \sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]} \right).$$

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

COROLLARY 4 (Bounds on the $\sqrt{\text{lasso}}$ prediction norm for $\epsilon_i \sim t(2)$). *Under the setting of Theorem 10, suppose that $\epsilon_i \sim t(2)$ are i.i.d. disturbances. Then for any $\tau \in (0, 1/2)$, with probability at least $1 - \alpha - \tau - \frac{2\log(4n/\tau)}{nu_n/[1+u_n]} - \frac{72\log^2 n}{n^{1/2}(\log n - 6)^2}$, we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\left(c_s + \sigma\sqrt{\log(4n/\tau) + 2\sqrt{2}/\tau}\right)\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2}.$$

Asymptotically, if $1/\alpha = o(\log n)$ and $s\log(p/\alpha) = o(n\bar{\kappa})$, considering $\tau = 1/\log n$, the result above yields that with probability $1 - \alpha(1 + o(1))$

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

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.

APPENDIX A: PROOFS OF SECTION 4.1

PROOF OF LEMMA 2. The first result holds by definition. Note that for a diagonal matrix with positive entries, $\|v\|_{2,n} \geq \|\Gamma v\|_{2,n}/\|\Gamma\|_\infty$ and, since $\mathbb{E}_n[x_{ij}^2] = 1$, $\|v\|_{2,n} \leq \|v\|_1$ for any $v \in \mathbb{R}^p$. For any δ such that $\|\Gamma\delta_{T^c}\|_1 < \|\Gamma\delta_T\|_1$ we have that

$$\frac{\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} \geq \frac{\|\Gamma\|_\infty^{-1}\|\Gamma\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1}$$
$$\geq \frac{\|\Gamma\|_\infty^{-1}(\|\Gamma\delta_T\|_{2,n} - \|\Gamma\delta_{T^c}\|_{2,n})}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} \geq \frac{\|\Gamma\|_\infty^{-1}(\|\Gamma\delta_T\|_{2,n} - \|\Gamma\delta_{T^c}\|_1)}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1}.$$

The result follows since $\|\Gamma\delta_T\|_{2,n} = \|\Gamma\delta_T\|_1$ if $|T| = 1$.

To show the third statement note that T does not change by including repeated regressors. Next let δ^1 and δ^2 denote the vectors in each copy of the regressors so that $\delta = \delta^1 + \delta^2$. It follows that

$$\frac{\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} = \frac{\|\delta\|_{2,n}}{\|\Gamma\delta_T^1\|_1 - \|\Gamma\delta_{T^c}^1\|_1 - \|\Gamma\delta_T^2\|_1 - \|\Gamma\delta_{T^c}^2\|_1}$$

which is minimized in the case that $\tilde{\delta}^1 = \delta$, $\tilde{\delta}_T^1 = \delta_T^1 + \delta_T^2$, $\tilde{\delta}_{T^c}^1 = \delta_{T^c}^1 + \delta_{T^c}^2$, and $\tilde{\delta}^2 = 0$. $\square$

PROOF OF LEMMA 3. Note that T does not change by including repeated regressors. Next let δ^1 and δ^2 denote the vectors in each copy of the regressors so that $\delta = \delta^1 + \delta^2$. It follows that $|\widetilde{S}'\delta|/\|\delta\|_{2,n} = |\widetilde{S}'\tilde{\delta}|/\|\tilde{\delta}\|_{2,n}$ where $\tilde{\delta}_T = \delta_T^1 + \delta_T^2$, and $\tilde{\delta}_{T^c} = \delta - \tilde{\delta}_T$. This transformation also increases the ℓ_1-norm of the coefficients over T and is considered so that $\tilde{\delta} \in \Delta_{\bar{c}}$. Finally, the restriction of $\tilde{\delta}$ to its first p components is also considered into the definition of $\varrho_{\bar{c}}$ without the repeated regressors. $\qquad\square$

APPENDIX B: PROOFS OF SECTION 4.2-4.4

PROOF OF THEOREM 1. First note that by Lemma 4 we have $\widehat{\delta} := \widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$. By optimality of $\widehat{\beta}$ and definition of $\bar{\kappa}$, $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}]$ we have

(B.1)
$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq \frac{\lambda}{n}\|\Gamma\beta_0\|_1 - \frac{\lambda}{n}\|\Gamma\widehat{\beta}\|_1 \leq \frac{\lambda}{n}(\|\Gamma\widehat{\delta}_T\|_1 - \|\Gamma\widehat{\delta}_{T^c}\|_1) \leq \bar{\rho}\|\widehat{\delta}\|_{2,n}.$$

Multiplying both sides by $\sqrt{\widehat{Q}(\widehat{\beta})} + \sqrt{\widehat{Q}(\beta_0)}$ and using that $(a+b)(a-b) = a^2 - b^2$

(B.2) $\qquad \|\widehat{\delta}\|_{2,n}^2 \leq 2\mathbb{E}_n[(\sigma\epsilon_i + r_i)x_i'\widehat{\delta}] + \left(\sqrt{\widehat{Q}(\widehat{\beta})} + \sqrt{\widehat{Q}(\beta_0)}\right)\bar{\rho}\|\widehat{\delta}\|_{2,n}.$

From (B.1) we have $\sqrt{\widehat{Q}(\widehat{\beta})} \leq \sqrt{\widehat{Q}(\beta_0)} + \bar{\rho}\|\widehat{\delta}\|_{2,n}$ so that

$$\|\widehat{\delta}\|_{2,n}^2 \leq 2\mathbb{E}_n[(\sigma\epsilon_i + r_i)x_i'\widehat{\delta}] + 2\sqrt{\widehat{Q}(\beta_0)}\bar{\rho}\|\widehat{\delta}\|_{2,n} + \bar{\rho}^2\|\widehat{\delta}\|_{2,n}^2.$$

Since $|\mathbb{E}_n[(\sigma\epsilon_i + r_i)x_i'\widehat{\delta}]| = \sqrt{\widehat{Q}(\beta_0)}|\widetilde{S}'\widehat{\delta}| \leq \sqrt{\widehat{Q}(\beta_0)}\varrho_{\bar{c}}\|\widehat{\delta}\|_{2,n}$ we obtain

$$\|\widehat{\delta}\|_{2,n}^2 \leq 2\sqrt{\widehat{Q}(\beta_0)}\varrho_{\bar{c}}\|\widehat{\delta}\|_{2,n} + 2\sqrt{\widehat{Q}(\beta_0)}\bar{\rho}\|\widehat{\delta}\|_{2,n} + \bar{\rho}^2\|\widehat{\delta}\|_{2,n}^2,$$

and the result follows provided $\bar{\rho} < 1$. $\qquad\square$

PROOF OF THEOREM 4. Let $\delta := \widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$ under the condition that $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ by Lemma 4.

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

$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq \frac{\lambda}{n}(\|\Gamma\beta_0\|_1 - \|\Gamma\widehat{\beta}\|_1) \leq \frac{\lambda}{n}(\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1) \leq \frac{\lambda\sqrt{s}}{n\bar{\kappa}}\|\delta\|_{2,n}$$

by definition of $\bar{\kappa}$ (note that if $\delta \notin \Delta_1$ we have $\widehat{Q}(\widehat{\beta}) \leq \widehat{Q}(\beta_0)$). The result follows from Theorem 1 to bound $\|\delta\|_{2,n}$.

To establish the lower bound, by convexity of $\sqrt{\widehat{Q}}$ and the definition of $\varrho_{\bar{c}}$ we have

$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -\widetilde{S}'\delta \geq -\varrho_{\bar{c}}\|\delta\|_{2,n}.$$

Thus, by Theorem 1, letting $\bar{\rho} := \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we obtain

$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -2\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}}^2 + \varrho_{\bar{c}}\bar{\rho}}{1 - \bar{\rho}^2}.$$

$\square$

PROOF OF THEOREM 5. We can assume that $\sqrt{\widehat{Q}(\beta_0)} > 0$ otherwise the result is trivially true. In the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Lemma 5

$$(B.3)$$
$$\left(\sqrt{\frac{\widehat{Q}(\widehat{\beta})}{\widehat{Q}(\beta_0)}} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\widehat{T}|} \leq \sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}.$$

Under the condition $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have by the lower bound in Theorem 4

$$\left(1 - \frac{1}{c} - \frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2}\right)\frac{\lambda\sqrt{\widehat{Q}(\beta_0)}}{n\sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})}}\sqrt{|\widehat{T}|} \leq \|\widehat{\beta} - \beta_0\|_{2,n}.$$

$\square$

PROOF OF THEOREM 6. For notational convenience we denote $\phi_n(m) = \phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[x_i x_i']\Gamma^{-1})$. We can assume that $\sqrt{\widehat{Q}(\beta_0)} > 0$ otherwise the result follows by Theorem 9 which establish $\widehat{\beta} = \beta_0$.

In the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Lemma 5

$$(B.4) \qquad \left(\sqrt{\frac{\widehat{Q}(\widehat{\beta})}{\widehat{Q}(\beta_0)}} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\widehat{T}|} \leq \sqrt{\phi_n(\widehat{m})}\|\widehat{\beta} - \beta_0\|_{2,n}.$$

Under the condition $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have by Theorem 1 and Theorem 4 that

$$\left(1 - \frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\widehat{T}|} \leq \sqrt{\phi_n(\widehat{m})}2\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2}.$$

Since we assume $\frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\rho})}{1 - \bar{\rho}^2} \leq 1/\bar{c}$ we have

$$\sqrt{|\widehat{T}|} \leq 2\bar{c}\sqrt{\phi_n(\widehat{m})}\,\frac{n}{\lambda}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2}.$$

By Lemma 3 and $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, we have $\varrho_{\bar{c}} \leq [\lambda/n][\sqrt{s}/\kappa_{\bar{c}}](1+\bar{c})/c$. Lemma 2 yields $\bar{\kappa} \geq \kappa_{\bar{c}}$ so that $\bar{\rho} \leq \lambda\sqrt{s}/[n\kappa_{\bar{c}}]$. Thus, under the condition $\bar{\rho} \leq 1/\sqrt{2}$,

$$\text{(B.5)} \qquad |\widehat{T}| \leq s \; \phi_n(\widehat{m}) \left(\frac{4\bar{c}^2}{\kappa_{\bar{c}}}\right)^2,$$

since $1 + [1/c] + [\bar{c}/c] = \bar{c}$.

Consider any $m \in \mathcal{M}$, and suppose $\widehat{m} > m$. Therefore by sublinearity of sparse eigenvalues

$$\widehat{m} \leq s \cdot \left\lceil \frac{\widehat{m}}{m} \right\rceil \phi_n(m) \left(\frac{4\bar{c}^2}{\kappa_{\bar{c}}}\right)^2.$$

Thus, since $\lceil k \rceil < 2k$ for any $k \geq 1$ we have $m < s \cdot 2\phi_n(m)(4\bar{c}^2/\kappa_{\bar{c}})^2$ which violates the condition of $m \in \mathcal{M}$ and s. Therefore, we must have $\widehat{m} \leq m$. In turn, applying (B.5) once more with $\widehat{m} \leq m$ we obtain $\widehat{m} \leq s \cdot \phi_n(m)(4\bar{c}^2/\kappa_{\bar{c}})^2$. The result follows by minimizing the bound over $m \in \mathcal{M}$. □

PROOF OF THEOREM 7. Let $X = [x_1; \ldots; x_n]'$ denote a n by p matrix and for a set of indices $S \subset \{1, \ldots, p\}$ we define $\mathcal{P}_S = X[S](X[S]'X[S])^{-1}X[S]'$ denote the projection matrix on the columns associated with the indices in S. We have that $f - X\widetilde{\beta} = (I - \mathcal{P}_{\widehat{T}})f - \mathcal{P}_{\widehat{T}}\epsilon$ where I is the identity operator. Therefore we have
(B.6)
$$\begin{aligned}\sqrt{n}\|\beta_0 - \widetilde{\beta}\|_{2,n} &= \|X\beta_0 - X\widetilde{\beta}\|_2 \leq \sqrt{n}c_s + \|f - X\widetilde{\beta}\|_2 \\ &\leq \sqrt{n}c_s + \|(I - \mathcal{P}_{\widehat{T}})f\|_2 + \sigma\|\mathcal{P}_T\epsilon\|_2 + \sigma\|\mathcal{P}_{\widehat{T}\setminus T}\epsilon\|_2.\end{aligned}$$

Since $\|X[\widehat{T} \setminus T]/\sqrt{n}(X[\widehat{T} \setminus T]'X[\widehat{T} \setminus T]/n)^{-1}\| \leq \sqrt{1/\phi_{\min}(\widehat{m})}$, $\widehat{m} = |\widehat{T} \setminus T|$, the last term in (B.6) satisfies

$$\|\mathcal{P}_{\widehat{T}\setminus T}\epsilon\|_2 \leq \sqrt{1/\phi_{\min}(\widehat{m})}\|X[\widehat{T} \setminus T]'\epsilon/\sqrt{n}\|_2 \leq \sqrt{\widehat{m}/\phi_{\min}(\widehat{m})}\|X'\epsilon/\sqrt{n}\|_\infty.$$

By Corollary 8, with probability at least $1 - 1/[9C^2 \log p]$, we have

$$\|X'\epsilon/\sqrt{n}\|_\infty = \sqrt{n}\|\mathbb{E}_n[\epsilon_i x_i]\|_\infty \leq 24C\sqrt{\vartheta \vee \max_{j=1,\ldots,p} \mathbb{E}_n\left[x_{ij}^2\epsilon_i^2\right]}\sqrt{\log p}.$$

We proceed to bound $\|\mathcal{P}_T\epsilon\|_2$. Since $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{E}[\epsilon_i^2] \leq \vartheta$,

$$\mathrm{E}\left[\|X[T]'\epsilon/\sqrt{n}\|_2^2\right] = n\mathrm{E}[\mathbb{E}_n[\epsilon_i x_{iT}]'\mathbb{E}_n[\epsilon_i x_{iT}]] = \sum_{j\in T}\mathbb{E}_n[\mathrm{E}[\epsilon_i^2]x_{ij}^2] \leq \vartheta s.$$

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

$$\|\mathcal{P}_T \epsilon\|_2 \leq \frac{\|X[T]'\epsilon/\sqrt{n}\|_2}{\sqrt{\phi_{\min}(s)}} \leq \frac{C\sqrt{\vartheta}}{\sqrt{\phi_{\min}(s)}}\sqrt{s}.$$

These relations yield the first result.

The second result follows from Theorem 1 and

$$\min_{\beta} \sqrt{\mathbb{E}_n[(f_i - x_i'\beta_{\widehat{T}})^2]} \leq \sqrt{\mathbb{E}_n[(f_i - x_i'\widehat{\beta})^2]} \leq c_s + \|\beta_0 - \widehat{\beta}\|_{2,n}.$$

$\square$

APPENDIX C: PROOFS OF SECTION 4.5

PROOF OF THEOREM 8. Let $t_n = \Phi^{-1}(1 - \alpha/2p)$ and recall we have $w_n = (\alpha^{-1}\log n C_q \bar{\mathbf{E}}[|\epsilon_i|^{q\vee 4}])^{1/q} < 1/2$ under Condition R. Thus

(C.1)
$$
\begin{aligned}
P\left(\lambda/n \geq c\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty\right) \ &= P\left((1+u_n)(t_n + 1 + u_n) \geq \sqrt{n}\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty\right) \\
&\leq P(\sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]} \leq \sqrt{1+u_n}\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}) + \\
&\quad + P(1 + u_n \geq \sqrt{n}\|\widehat{\Gamma}_k^{-1}\mathbb{E}_n[x_ir_i]\|_\infty/\sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]}) + \\
&\quad + P(t_n \geq \max_{1\leq j\leq p}\sqrt{n}|\mathbb{E}_n[x_{ij}\epsilon_i]|/\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}) + \\
&\quad + P(\sqrt{1+u_n}\widehat{\gamma}_{j,k} \geq \sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]/\mathbb{E}_n[\epsilon_i^2]}, j = 1, \ldots, p).
\end{aligned}
$$

Next we proceed to bound each term.

First Term of (C.1). By Lemma 7 with $v = w_n$ we have that

$$
\begin{aligned}
P(\sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]} \leq \sqrt{1+u_n}\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}) \ &\leq \psi(w_n) + \frac{2(1+u_n)\max_{1\leq i\leq n}\mathrm{E}[\epsilon_i^2]}{(1-w_n)u_n \; n} \\
&\leq \frac{\alpha}{\log n} + \frac{4(1+u_n)(\bar{\mathbf{E}}[|\epsilon_i|^q])^{2/q}}{u_n \; n^{1-[2/q]}}.
\end{aligned}
$$

Second Term of (C.1). By Lemma 6 and using that $\widehat{\gamma}_{j,k} \geq 1$,

$$\|\widehat{\Gamma}_k^{-1}\mathbb{E}_n[x_ir_i]\|_\infty \leq \|\mathbb{E}_n[x_ir_i]\|_\infty \leq \sigma/\sqrt{n}.$$

Thus, since $[2u_n + u_n^2]/[1+u_n]^2 \geq u_n/[1+u_n] \geq w_n$, we have

$$
\begin{aligned}
P((1+u_n)\sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]} \geq \sqrt{n}\|\widehat{\Gamma}_k^{-1}\mathbb{E}_n[x_ir_i]\|_\infty) \ &\leq P(\sqrt{\mathbb{E}_n[\epsilon_i^2]} \geq 1/(1+u_n)) \\
&\leq P(|\mathbb{E}_n[\epsilon_i^2] - 1| \geq [2u_n + u_n^2]/[1+u_n]^2) \\
&\leq \psi(w_n) \leq \alpha/\log n.
\end{aligned}
$$

Third Term of (C.1). Let $\bar{t} = \min_{1 \le j \le p}(\mathbb{E}_n[x_{ij}^2 \mathrm{E}[\epsilon_i^2]])^{1/2}/(\mathbb{E}_n[|x_{ij}^3|\mathrm{E}[|\epsilon_i^3|]])^{1/3} > 0$. By Lemma 16, since $t_n \le \bar{t}\, n^{1/6} - 1$ by Condition R, we have that there is an universal constant A, such that

$$P\left(\max_{1 \le j \le p} \frac{\sqrt{n}|\mathbb{E}_n[x_{ij}\epsilon_i]|}{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}} > t_n\right) \le p\, \max_{1 \le j \le p} P\left(\frac{\sqrt{n}|\mathbb{E}_n[x_{ij}\epsilon_i]|}{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}} > t_n\right)$$
$$\le 2p\,\bar{\Phi}(t_n)\left(1 + \frac{A}{\ell_n^3}\right) \le \alpha\left(1 + \frac{A}{\ell_n^3}\right)$$

where the last inequality follows from the definition of t_n.

Fourth Term of (C.1). Let $\widehat{\Gamma}_k = \mathrm{diag}(\widehat{\gamma}_{1,k}, \ldots, \widehat{\gamma}_{p,k})$. First we consider the initial choice of $\widehat{\gamma}_{j,0} = w(\mathbb{E}_n[x_{ij}^4])^{1/4}$. Then we have

$$\sqrt{1+u_n}\widehat{\gamma}_{j,0} \ge \sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}/\sqrt{\mathbb{E}_n[\epsilon_i^2]} \quad \text{for all } j = 1, \ldots, p$$

provided that $\sqrt{1+u_n}w\sqrt{\mathbb{E}_n[\epsilon_i^2]} \ge (\mathbb{E}_n[\epsilon_i^4])^{1/4}$. We bound this probability

$$P(\sqrt{1+u_n}w\sqrt{\mathbb{E}_n[\epsilon_i^2]} < (\mathbb{E}_n[\epsilon_i^4])^{1/4}) \le P(\mathbb{E}_n[\epsilon_i^4] > w^4) + P\left(\mathbb{E}_n[\epsilon_i^2] < \frac{1}{1+u_n}\right)$$
$$\le \frac{C_q\bar{\mathbf{E}}[|\epsilon_i|^{q\vee 8}]}{v^q n^{q/8}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon_i|^q]}{n^{1\wedge(q/4-1)}v^{q/4}} + \psi\left(\frac{u_n}{1+u_n}\right)$$

where $v^4 = (w^4 - \bar{\mathbf{E}}[\epsilon_i^4]) \vee 0$. The result follows since $u_n/[1+u_n] \ge w_n$ so that $\psi(u_n/[1+u_n]) \le \alpha/\log n$.

To show the second result of the theorem, consider the iterations of Algorithm 1 for $k \ge 1$ conditioned on $\lambda/n \ge c\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty$ for $k = 0$. First we establish a lower bound on $\widehat{\gamma}_{j,k}$. Let $x_{\infty j} = \max_{1 \le i \le n}|x_{ij}|$,

$$\widehat{\gamma}_{j,k} = 1 \vee \frac{\sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}}{\sqrt{\mathbb{E}_n[(y_i - x_i'\widehat{\beta})^2]}} \ge \frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} - \sqrt{\mathbb{E}_n[x_{ij}^2\{x_i'(\widehat{\beta}-\beta_0)\}^2]}/\sigma - \sqrt{\mathbb{E}_n[x_{ij}^2 r_i^2]}/\sigma}{\sqrt{\mathbb{E}_n[\epsilon_i^2]} + \|\widehat{\beta}-\beta_0\|_{2,n}/\sigma + \sqrt{\mathbb{E}_n[r_i^2]}/\sigma}$$
$$\ge \frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} - x_{\infty j}(\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)/\sigma}{\sqrt{\mathbb{E}_n[\epsilon_i^2]} + (\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)/\sigma}.$$

Since $\widehat{\gamma}_{j,k} \ge 1$, it suffices to consider the case that $\mathbb{E}_n[\epsilon_i^2] \le \mathbb{E}_n[x_{ij}^2\epsilon_i^2]$. Therefore we have that

$$(1+\Delta)\gamma_j \ge \sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}/\sqrt{\mathbb{E}_n[\epsilon_i^2]}$$

is implied by

$$\text{(C.2)} \qquad \Delta \ge 2(\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)x_{\infty j}/\{\sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]}\}.$$

The choice of $\Delta = \sqrt{1 + u_n} - 1$ is appropriate under the extra condition assumed in the theorem and by Theorem 1 to bound $\|\widehat{\beta} - \beta_0\|_{2,n}$. Thus, $\lambda/n \geq c\|\widehat{\Gamma}_k^{-1}\widetilde{S}\|_\infty$ for $k = 1$.

Next we establish an upper bound on $\widehat{\gamma}_{j,k}$.

$$
\begin{aligned}
\widehat{\gamma}_{j,k} \ &= 1 \vee \frac{\sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}}{\sqrt{\mathbb{E}_n[(y_i - x_i'\widehat{\beta})^2]}} \leq \frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} + \sqrt{\mathbb{E}_n[x_{ij}^2\{x_i'(\widehat{\beta} - \beta_0)\}^2]}/\sigma + \sqrt{\mathbb{E}_n[x_{ij}^2 r_i^2]}/\sigma}{\sqrt{\mathbb{E}_n[\epsilon_i^2]} - \|\widehat{\beta} - \beta_0\|_{2,n}/\sigma - \sqrt{\mathbb{E}_n[r_i^2]}/\sigma} \\
&\leq \frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} + x_{\infty j}(\|\widehat{\beta} - \beta_0\|_{2,n} + c_s)/\sigma}{\sqrt{\mathbb{E}_n[\epsilon_i^2]} - (\|\widehat{\beta} - \beta_0\|_{2,n} + c_s)/\sigma}.
\end{aligned}
$$

Under the conditions that $\max_{1 \leq i \leq n} \|x_i\|_\infty(\|\widehat{\beta} - \beta_0\|_{2,n} + c_s)/\sigma \leq u_n\sqrt{\mathbb{E}_n[\epsilon_i^2]}/2$, we have

$$
\widehat{\gamma}_{j,k} \leq 1 \vee \frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]} + u_n\sqrt{\mathbb{E}_n[\epsilon_i^2]}/2}{\sqrt{\mathbb{E}_n[\epsilon_i^2]} - u_n\sqrt{\mathbb{E}_n[\epsilon_i^2]}/2} \leq 1 \vee \frac{1 + u_n/2}{1 - u_n/2}\frac{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} \leq \frac{(1 + u_n/2)^2}{1 - u_n/2}\widehat{\gamma}_{j,0}.
$$

Let $\Gamma^* = \operatorname{diag}(\gamma_1^*, \ldots, \gamma_p^*)$ where $\gamma_j^* = 1 \vee \sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}/\sqrt{\mathbb{E}_n[\epsilon_i^2]}$, and recall that $(2 + u_n)/(2 - u_n) \leq 2$ since $u_n \leq 2/3$. We have that $\varrho_{\bar{c}}(\widehat{\Gamma}_k) \leq \varrho_{\bar{c}\|\widehat{\Gamma}_k\Gamma^{*-1}\|_\infty}(\Gamma^*) \leq \varrho_{2\bar{c}}(\Gamma^*)$.

Also, letting $\widetilde{\delta} = \Gamma^{*-1}\widehat{\Gamma}_k\delta$, note that

$$
\begin{aligned}
\bar{\kappa}(\widehat{\Gamma}_k) \ &= \min_{\|\widehat{\Gamma}_k\delta_{T^c}\|_1 < \|\widehat{\Gamma}_k\delta_T\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\widehat{\Gamma}_k\delta_T\|_1 - \|\widehat{\Gamma}_k\delta_{T^c}\|_1} \\
&= \min_{\|\Gamma^*\widetilde{\delta}_{T^c}\|_1 < \|\Gamma^*\widetilde{\delta}_T\|_1} \frac{\sqrt{s}\|\widehat{\Gamma}_k^{-1}\Gamma^*\widetilde{\delta}\|_{2,n}}{\|\Gamma^*\widetilde{\delta}_T\|_1 - \|\Gamma^*\widetilde{\delta}_{T^c}\|_1} \geq \bar{\kappa}(\Gamma^*)/\|(\widehat{\Gamma}_k^{-1}\Gamma^*)^{-1}\|_\infty.
\end{aligned}
$$

Thus by Theorem 1 we have that the estimator with $\widehat{\beta}$ based on $\widehat{\Gamma}_k$, $k = 1$, also satisfies (C.2) by the extra condition assumed in the theorem. Thus the same argument established $k > 1$. $\qquad \square$

APPENDIX D: PROOFS OF SECTION 4.6

PROOF OF THEOREM 9. Note that because $\sigma = 0$ and $c_s = 0$, we have $\sqrt{\widehat{Q}(\beta_0)} = 0$ and $\sqrt{\widehat{Q}(\widehat{\beta})} = \|\widehat{\beta} - \beta_0\|_{2,n}$. Thus, by optimality of $\widehat{\beta}$ we have

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

Therefore, $\|\Gamma\widehat{\beta}\|_1 \leq \|\Gamma\beta_0\|_1$ which implies that $\delta = \widehat{\beta} - \beta_0$ satisfies $\|\Gamma\delta_{T^c}\|_1 \leq \|\Gamma\delta_T\|_1$. In turn

$$
\|\delta\|_{2,n} \leq \frac{\lambda}{n}(\|\Gamma\widehat{\delta}_T\|_1 - \|\Gamma\widehat{\delta}_{T^c}\|_1) \leq \frac{\lambda\sqrt{s}}{n\bar{\kappa}}\|\delta\|_{2,n}.
$$

Since $\lambda\sqrt{s} < n\bar{\kappa}$ we have $\|\delta\|_{2,n} = 0$.

Next the relation $0 = \sqrt{s}\|\delta\|_{2,n} \geq \bar{\kappa}(\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1)$ implies $\|\Gamma\delta_T\|_1 = \|\Gamma\delta_{T^c}\|_1$ since $\bar{\kappa} > 0$ by our assumptions.

Also, if $\kappa_1 > 0$, $0 = \sqrt{s}\|\delta\|_{2,n} \geq \kappa_1\|\Gamma\delta_T\|_1 \geq \kappa_1\|\Gamma\delta\|_1/2$. Since $\Gamma > 0$, this shows that $\delta = 0$ and $\widehat{\beta} = \beta_0$.

$\square$

PROOF OF THEOREM 10. If $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Theorem 1, for $\bar{\rho} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$ we have

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}} + \bar{\rho}}{1 - \bar{\rho}^2},$$

and the stated bound on the prediction norm follows by $\sqrt{\widehat{Q}(\beta_0)} \leq c_s + \sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]}$.

Thus we need to show that the choice of λ and Γ is suitable for the desired probability on the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$. By the choice of u_n it suffices to show that

$$P\left(\max_{1 \leq j \leq p} \frac{\sqrt{n}|\mathbb{E}_n[(\sigma\epsilon_i + r_i)x_{ij}]|}{\max_{1 \leq i \leq n}|x_{ij}|\sqrt{\mathbb{E}_n[(\sigma\epsilon_i)^2]}} > 1 + \sqrt{2\log(2p/\alpha)}\right) \leq \alpha + o(1).$$

By Lemma 6 we have $\|\mathbb{E}_n[r_ix_i]\|_\infty \leq \sigma/\sqrt{n}$. Since $\max_{1 \leq i \leq n}|x_{ij}| \geq \mathbb{E}_n[x_{ij}^2] = 1$, and $P(\mathbb{E}_n[\epsilon_i^2] \leq 1) \leq \eta$ it suffices to establish that

$$P\left(\max_{1 \leq j \leq p} \frac{\sqrt{n}|\mathbb{E}_n[\epsilon_i x_{ij}]|}{\max_{1 \leq i \leq n}|x_{ij}|\sqrt{\mathbb{E}_n[\epsilon_i^2]}} > \sqrt{2\log(2p/\alpha)}\right) \leq \alpha.$$

This follows since

$$P\left(\max_{1 \leq j \leq p} \frac{\sqrt{n}|\mathbb{E}_n[\epsilon_i x_{ij}]|}{\max_{1 \leq i \leq n}|x_{ij}|\sqrt{\mathbb{E}_n[\epsilon_i^2]}} > \sqrt{2\log(2p/\alpha)}\right) \leq P\left(\max_{1 \leq j \leq p} \frac{\sqrt{n}|\mathbb{E}_n[\epsilon_i x_{ij}]|}{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}} > \sqrt{2\log(2p/\alpha)}\right)$$

$$\leq p\max_{1 \leq j \leq p} P\left(\frac{\sqrt{n}|\mathbb{E}_n[\epsilon_i x_{ij}]|}{\sqrt{\mathbb{E}_n[x_{ij}^2\epsilon_i^2]}} > \sqrt{2\log(2p/\alpha)}\right) \leq \alpha$$

where we used the union bound and Theorem 2.15 of [10] because ϵ_i's are independent and symmetric. $\square$

REFERENCES

[1] Stephen Becker, Emmanuel Candès, and Michael Grant. Templates for convex cone problems with applications to sparse signal recovery. *ArXiv*, 2010.

[2] A. Belloni and V. Chernozhukov. Post-ℓ_1-penalized estimators in high-dimensional linear regression models. *arXiv:[math.ST]*, 2009.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *accepted at the Annals of Statistics*, 2010.

[4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *arXiv:[math.ST]*, 2010.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

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

[7] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.

[8] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[9] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[10] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.

[11] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B*, 70(5):849–911, 2008.

[12] W. Feller. *An Introduction to Probability Theory and Its Applicaitons, Volume II*. Wiley Series in Probability and Mathematical Statistics, 1966.

[13] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[14] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.

[15] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.

[16] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.

[17] Z. Lu. Gradient based method for cone programming with application to large-scale compressed sensing. *Technical Report*, 2008.

[18] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.

[19] Y. Nesterov. Smooth minimization of non-smooth functions, mathematical programming. *Mathematical Programming*, 103(1):127–152, 2005.

[20] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[21] Y. Nesterov and A Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1993.

[22] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2001.

[23] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.

[24] H. P. Rosenthal. On the subspaces of l^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 9:273–303, 1970.

[25] A. D. Slastnikov. Limit theorems for moderate deviation probabilities. *Theory of*

Probability and its Applications, 23:322–340, 1979.

[26] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.

[27] R. H. Tütüncü, K. C. Toh, and M. J. Todd. SDPT3 — a MATLAB software package for semidefinite-quadratic-linear programming, version 3.0. Technical report, 2001. Available at http://www.math.nus.edu.sg/∼mattohkc/sdpt3.html.

[28] S. A. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.

[29] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

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

[31] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[32] Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \le r \le 2$. *Ann. Math. Statist*, 36:299–303, 1965.

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

[34] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

Supplementary Material for the paper "Pivotal Estimation of Nonparametric Functions via Square-root Lasso"

APPENDIX B: OMITTED PROOFS

PROOF OF LEMMA 4. In this step we show that $\widehat{\delta} = \widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$ under the prescribed penalty level. By definition of $\widehat{\beta}$

$$\text{(B.1)} \quad \sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq \frac{\lambda}{n}\|\Gamma\beta_0\|_1 - \frac{\lambda}{n}\|\Gamma\widehat{\beta}\|_1 \leq \frac{\lambda}{n}(\|\Gamma\widehat{\delta}_T\|_1 - \|\Gamma\widehat{\delta}_{T^c}\|_1),$$

where the last inequality holds because

$$\text{(B.2)} \quad \begin{aligned} \|\Gamma\beta_0\|_1 - \|\Gamma\widehat{\beta}\|_1 &= \|\Gamma\beta_{0T}\|_1 - \|\Gamma\widehat{\beta}_T\|_1 - \|\Gamma\widehat{\beta}_{T^c}\|_1 \\ &\leq \|\Gamma\widehat{\delta}_T\|_1 - \|\Gamma\widehat{\delta}_{T^c}\|_1. \end{aligned}$$

Note that using the convexity of $\sqrt{\widehat{Q}}$, $-\widetilde{S} \in \partial\sqrt{\widehat{Q}}(\beta_0)$, and if $\lambda/n \geq cn\|\Gamma^{-1}\widetilde{S}\|_\infty$, we have

$$\text{(B.3)} \quad \sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -\widetilde{S}'\widehat{\delta} \geq -\|\Gamma^{-1}\widetilde{S}\|_\infty\|\Gamma\widehat{\delta}\|_1$$

$$\text{(B.4)} \quad \geq -\frac{\lambda}{cn}(\|\Gamma\widehat{\delta}_T\|_1 + \|\Gamma\widehat{\delta}_{T^c}\|_1)$$

$$\text{(B.5)} \quad \geq -\frac{\lambda}{cn}(\|\Gamma\beta_0\|_1 + \|\Gamma\widehat{\beta}\|_1).$$

Combining (B.1) with (B.4) we obtain

$$\text{(B.6)} \quad -\frac{\lambda}{cn}(\|\Gamma\widehat{\delta}_T\|_1 + \|\Gamma\widehat{\delta}_{T^c}\|_1) \leq \frac{\lambda}{n}(\|\Gamma\widehat{\delta}_T\|_1 - \|\Gamma\widehat{\delta}_{T^c}\|_1),$$

that is

$$\text{(B.7)} \quad \|\Gamma\widehat{\delta}_{T^c}\|_1 \leqslant \frac{c+1}{c-1}\cdot\|\Gamma\widehat{\delta}_T\|_1 = \bar{c}\|\Gamma\widehat{\delta}_T\|_1, \text{ or } \widehat{\delta} \in \Delta_{\bar{c}}.$$

On the other hand, by (B.5) and (B.1) we have

$$\text{(B.8)} \quad -\frac{\lambda}{cn}(\|\Gamma\beta_0\|_1 + \|\Gamma\widehat{\beta}\|_1)+ \leq \frac{\lambda}{n}(\|\Gamma\beta_0\|_1 - \|\Gamma\widehat{\beta}\|_1).$$

which similarly leads to $\|\Gamma\widehat{\beta}\|_1 \leq \bar{c}\|\Gamma\beta_0\|_1$. $\qquad\square$

PROOF OF THEOREM 2. Let $\delta := \widehat{\beta} - \beta_0$. Under the condition on λ above, we have that $\delta \in \Delta_{\bar{c}}$. Thus, we have

$$\|\Gamma\delta\|_1 \leq (1+\bar{c})\|\Gamma\delta_T\|_1 \leq (1+\bar{c})\frac{\sqrt{s}\|\delta\|_{2,n}}{\kappa_{\bar{c}}},$$

by the restricted eigenvalue condition. The result follows by Theorem 1 to bound $\|\delta\|_{2,n}$. $\square$

PROOF OF THEOREM 3. Let $\delta := \widehat{\beta} - \beta_0$. We have that

$$\|\Gamma^{-1}\delta\|_\infty \leq \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' \delta]\|_\infty + \|\Gamma^{-1}(\mathbb{E}_n[x_i x_i' \delta] - \delta)\|_\infty.$$

Note that by the first-order optimality conditions of $\widehat{\beta}$ and the assumption on λ

$$
\begin{aligned}
\|\Gamma^{-1}\mathbb{E}_n[x_i x_i' \delta]\|_\infty &\leq \|\Gamma^{-1}\mathbb{E}_n[x_i(y_i - x_i'\widehat{\beta})]\|_\infty + \|\Gamma^{-1}\widetilde{S}\|_\infty\sqrt{\widehat{Q}(\beta_0)} \\
&\leq \frac{\lambda\sqrt{\widehat{Q}(\widehat{\beta})}}{n} + \frac{\lambda\sqrt{\widehat{Q}(\beta_0)}}{cn}
\end{aligned}
$$

by the first-order conditions and the condition on λ.

Next let e_j denote the jth-canonical direction.

$$
\begin{aligned}
\|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\delta\|_\infty &= \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\Gamma\delta\|_\infty \\
&\leq \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\|_\infty\|\Gamma\delta\|_1.
\end{aligned}
$$

Therefore, using the optimality of $\widehat{\beta}$ that implies $\sqrt{\widehat{Q}(\widehat{\beta})} \leq \sqrt{\widehat{Q}(\beta_0)} + (\lambda/n)(\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1) \leq \sqrt{\widehat{Q}(\beta_0)} + (\lambda\sqrt{s}/[n\bar{\kappa}])\|\delta\|_{2,n}$, we have

$$
\begin{aligned}
\|\Gamma^{-1}\delta\|_\infty &\leq \left(\sqrt{\widehat{Q}(\widehat{\beta})} + \frac{\sqrt{\widehat{Q}(\beta_0)}}{c}\right)\frac{\lambda}{n} + \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\|_\infty\|\Gamma\delta\|_1 \\
&\leq \left(1 + \frac{1}{c}\right)\frac{\lambda\sqrt{\widehat{Q}(\beta_0)}}{n} + \frac{\lambda^2\sqrt{s}}{n^2\bar{\kappa}}\|\delta\|_{2,n} + \|\Gamma^{-1}\mathbb{E}_n[x_i x_i' - I]\Gamma^{-1}\|_\infty\|\Gamma\delta\|_1.
\end{aligned}
$$

The result follows from Theorem 1 and 2. $\square$

PROOF OF LEMMA 5. Recall that $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$. First note that by strong duality

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

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

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

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

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

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

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

$$
\begin{array}{rcl}
\sqrt{\widehat{Q}(\widehat{\beta})}\sqrt{|\widehat{T}|}\lambda & = & \|\Gamma^{-1}(X'(Y - X\widehat{\beta}))_{\widehat{T}}\| \\
& \leq & \|\Gamma^{-1}(X'(Y - X\beta_0))_{\widehat{T}}\| + \|\Gamma^{-1}(X'X(\beta_0 - \widehat{\beta}))_{\widehat{T}}\| \\
& \leq & \sqrt{|\widehat{T}|}\,n\|\Gamma^{-1}\mathbb{E}_n\left[x_i(\sigma\epsilon_i + r_i)\right]\|_\infty + n\sqrt{\phi_{\max}(\widehat{m},\Gamma^{-1}\mathbb{E}_n[x_ix_i']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n} \\
& = & \sqrt{|\widehat{T}|}\,n\sqrt{\widehat{Q}(\beta_0)}\|\Gamma^{-1}\widetilde{S}\|_\infty + n\sqrt{\phi_{\max}(\widehat{m},\Gamma^{-1}\mathbb{E}_n[x_ix_i']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n},
\end{array}
$$

where we used

$$
\begin{array}{rcl}
\|\Gamma^{-1}(X'X(\widehat{\beta} - \beta_0))_{\widehat{T}}\| & \leq & \sup_{\|\alpha_{T^c}\|_0 \leq \widehat{m},\|\alpha\|\leq 1}|\alpha'\Gamma^{-1}X'X(\widehat{\beta} - \beta_0))| \\
& \leq & \sup_{\|\alpha_{T^c}\|_0 \leq \widehat{m},\|\alpha\|\leq 1}\|\alpha'\Gamma^{-1}X'\|\|X(\widehat{\beta} - \beta_0)\| \\
& = & n\sqrt{\phi_{\max}(\widehat{m},\Gamma^{-1}\mathbb{E}_n[x_ix_i']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}.
\end{array}
$$

$\square$

PROOF OF COROLLARY 4. We need to bound the probability of relevant events to establish the prediction norm bound by Theorem 10.

Applying Lemma 10(ii) with $a = 1/\log n$ we have $\eta = \frac{1}{n^{1/2}(1/6 - 1/\log n)^2} = \frac{36\log^2 n}{n^{1/2}(\log n - 6)^2}$.

Applying Lemma 10(iii) with $t_n = 4n/\tau$, $a = 1/\log n$, and $a_n = u_n/[1 + u_n]$, where we note the simplification that

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

we have

$$P\left(\sqrt{\mathbb{E}_n[\sigma^2\epsilon_i^2]} \leq (1 + u_n)\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}\right) \leq \gamma := \frac{2\log(4n/\tau)}{nu_n/[1 + u_n]} + \eta + \frac{\tau}{2}.$$

Thus, by Theorem 10, since $\bar{\rho} < 1$, with probability at least $1 - \alpha - \gamma - \eta$ we have

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \frac{2(1 + 1/c)}{1 - \bar{\rho}^2}\sqrt{\widehat{Q}(\beta_0)}(\varrho_{\bar{c}} + \bar{\rho}) \leq \frac{2(1 + 1/c)}{1 - \bar{\rho}^2}(c_s + \sigma\sqrt{\mathbb{E}_n[\epsilon_i^2]})(\varrho_{\bar{c}} + \bar{\rho}).$$

Finally, by Lemma 10(i) we have $\mathbb{E}_n[\epsilon_i^2] \leq 2\sqrt{2}/\tau + \log(4n/\tau)$ with probability at least $1 - \tau/2$.   □

APPENDIX C: TECHNICAL LEMMAS

LEMMA 6.   *Under Condition ASM we have*

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

PROOF. First note that for every $j = 1, \ldots, p$, we have $|\mathbb{E}_n[x_{ij}r_i]| \leq \sqrt{\mathbb{E}_n[x_{ij}^2]\mathbb{E}_n[r_i^2]} = c_s$. Next, by definition of β_0 in (2.2), for $j \in T$ we have $\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)] = \mathbb{E}_n[x_{ij}r_i] = 0$ since β_0 is a minimizer over the support of β_0. For $j \in T^c$ we have that for any $t \in \mathbb{R}$

$$\mathbb{E}_n[(f_i - x_i'\beta_0)^2] + \sigma^2\frac{s}{n} \leq \mathbb{E}_n[(f_i - x_i'\beta_0 - tx_{ij})^2] + \sigma^2\frac{s+1}{n}.$$

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

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

Taking the minimum over t in the right hand side at $t^* = \mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)]$ we obtain $-\sigma^2/n \leq -(\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)])^2$ or equivalently, $|\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)]| \leq \sigma/\sqrt{n}$.   □

LEMMA 7.   *Let $r_1, \ldots, r_n$ be fixed and assume ϵ_i are independent zero mean random variables such that $\bar{\mathbf{E}}[\epsilon_i^2] = 1$. Suppose that there is $q > 2$ such that $\bar{\mathbf{E}}[|\epsilon_i|^q] < \infty$. Then, for $u_n > 0$ we have*

$$P\left(\sqrt{\mathbb{E}_n[\sigma^2\epsilon_i^2]} > \sqrt{1 + u_n}\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}\right) \leq \min_{v \in (0,1)}\psi(v) + \frac{2(1 + u_n)\max_{1 \leq i \leq n}\mathrm{E}[\epsilon_i^2]}{u_n(1 - v)\,n},$$

where $\psi(v) := \frac{C_q\bar{\mathbf{E}}[|\epsilon_i|^{q \vee 4}]}{v^q n^{q/4}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon_i|^q]}{n^{1 \wedge (q/2 - 1)}v^{q/2}}$. Further we have $\max_{1 \leq i \leq n}\mathrm{E}[\epsilon_i^2] \leq n^{2/q}(\bar{\mathbf{E}}[|\epsilon_i|^q])^{2/q}$.

PROOF. Let $c_s = (\mathbb{E}_n[r_i^2])^{1/2}$ and $a_n = 1 - [1/(1 + u_n)] = u_n/(1 + u_n)$. We have that
(C.1)
$$P(\mathbb{E}_n[\sigma^2\epsilon_i^2] > (1 + u_n)\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]) = P(2\mathbb{E}_n[\epsilon_i r_i] < -c_s^2 - a_n\mathbb{E}_n[\sigma^2\epsilon_i^2]).$$

By Lemma 8 we have

$$\Pr(\sqrt{\mathbb{E}_n[\epsilon_i^2]} < 1 - v) \leq \Pr(|\mathbb{E}_n[\epsilon_i^2] - 1| > v) \leq \psi(v).$$

Thus,

$$P(\mathbb{E}_n[\sigma^2 \epsilon_i^2] > (1 + u_n)\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]) \leq \psi(v) + P(2\mathbb{E}_n[\sigma\epsilon_i r_i] < -c_s^2 - a_n\sigma^2(1 - v)).$$

Since ϵ_i's are independent of r_i's, we have

$$\mathrm{E}[(2\mathbb{E}_n[\sigma\epsilon_i r_i])^2] = 4\sigma^2 \bar{\mathbf{E}}[\epsilon_i^2 r_i^2]/n \leq \frac{4\sigma^2}{n} \left\{ \max_{1 \leq i \leq n} \mathrm{E}[\epsilon_i^2]c_s^2, \quad \max_{1 \leq i \leq n} r_i^2 \right\}.$$

By Chebyshev inequality we have

$$P\left(\sqrt{\mathbb{E}_n[\sigma^2 \epsilon_i^2]} > \sqrt{1 + u_n}\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]} \right) \begin{aligned} &\leq \psi(v) + \frac{4\sigma^2 c_s^2 \max\limits_{1 \leq i \leq n} \mathrm{E}[\epsilon_i^2]/n}{(c_s^2 + a_n\sigma^2(1 - v))^2} \\ &\leq \psi(v) + \frac{2(1 + u_n) \max\limits_{1 \leq i \leq n} \mathrm{E}[\epsilon_i^2]}{(1 - v)u_n n}. \end{aligned}$$

The result follows by minimizing over $v \in (0, 1)$.

Further, we have

$$\max_{1 \leq i \leq n} \mathrm{E}[\epsilon_i^2] \leq \mathrm{E}[\max_{1 \leq i \leq n} \epsilon_i^2] \leq (\mathrm{E}[\max_{1 \leq i \leq n} |\epsilon_i^q|])^{2/q} \leq n^{2/q}(\bar{\mathbf{E}}[|\epsilon_i^q|])^{2/q}.$$

$\square$

LEMMA 8. *Let* ϵ_i, $i = 1, \ldots, n$, *be independent random variables such that* $\bar{\mathbf{E}}[\epsilon_i^2] = 1$. *Assume that there is* $q > 2$ *such that* $\bar{\mathbf{E}}[|\epsilon_i|^q] < \infty$. *Then there is a constant* C_q, *that depends on* q *only, such that for* $v > 0$ *we have*

$$\Pr(|\mathbb{E}_n[\epsilon_i^2] - 1| > v) \leq \psi(v) := \frac{C_q \bar{\mathbf{E}}[|\epsilon_i|^{q \vee 4}]}{v^q n^{q/4}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon_i|^q]}{n^{1 \wedge (q/2 - 1)}v^{q/2}}.$$

PROOF. By the application of either Rosenthal's inequality [24] for the case of $q > 4$ or Vonbahr-Esseen's inequalities [32] for the case of $2 < q \leq 4$,

$$P(|\mathbb{E}_n[\epsilon_i^2] - 1| > v) \leq \psi(v) := \frac{C_q \bar{\mathbf{E}}[|\epsilon_i|^{q \vee 4}]}{v^q n^{q/4}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon_i|^q]}{n^{1 \wedge (q/2 - 1)}v^{q/2}}.$$

$\square$

LEMMA 9. *Let* ϵ_i, $i = 1, \ldots, n$, *be independent random variables such that* $\bar{\mathbf{E}}[|\epsilon_i|^q] < \infty$ *for* $q \geq 4$. *Conditional on* $x_1, \ldots, x_n \in \mathbb{R}^p$, *with probability* $1 - 4\tau_1 - 4\tau_2$

$$\max_{1 \leq j \leq p} |\mathbb{E}_n[x_{ij}^2(\epsilon_i^2 - \mathrm{E}[\epsilon_i^2])]| \leq 4\sqrt{\frac{2\log(2p/\tau_1)}{n}} \left(\frac{\bar{\mathbf{E}}[|\epsilon_i|^q]}{\tau_2}\right)^{\frac{2}{q}} \max_{1 \leq j \leq p} (\mathbb{E}_n[|x_{ij}|^{\frac{4q}{q-4}}])^{\frac{q-4}{2q}}$$

where in the case $q = 4$ *we have* $\{\mathbb{E}_n[|x_{ij}|^{4q/[q-4]}]\}^{[q-4]/q} = \max_{1 \leq i \leq n} |x_{ij}|$.

PROOF. For a random variable Z, let $\bar{q}(Z, 1 - \alpha) = (1 - \alpha)$-quantile of Z. Let

$$e_{1n} = 4\sqrt{\frac{2\log(2p/\tau_1)}{n}} \left(\frac{\bar{\mathbf{E}}[|\epsilon_i|^q]}{\tau_2}\right)^{2/q} \max_{1 \leq j \leq p} (\mathbb{E}_n[|x_{ij}|^{4q/[q-4]}])^{[q-4]/[2q]},$$

and note that

$$\begin{aligned}
\bar{q}(\max_{1 \leq j \leq p} \mathbb{E}_n[x_{ij}^4 \epsilon_i^4], 1 - \tau) &\leq \max_{1 \leq j \leq p} \{\mathbb{E}_n[|x_{ij}|^{4q/[q-4]}]\}^{[q-4]/q} \{\bar{q}(\mathbb{E}_n[|\epsilon_i|^q], 1 - \tau)\}^{4/q} \\
&\leq \max_{1 \leq j \leq p} \{\mathbb{E}_n[|x_{ij}|^{4q/[q-4]}]\}^{[q-4]/q} \bar{\mathbf{E}}[|\epsilon_i|^q]/\tau_2)^{4/q}
\end{aligned}$$

since $\bar{\mathbf{E}}[|\epsilon_i|^q] < \infty$, $\bar{q}(\mathbb{E}_n[|\epsilon_i|^q], 1 - \tau_2) \leq \bar{\mathbf{E}}[|\epsilon_i|^q]/\tau_2$, and

$$\max_{1 \leq j \leq p} \bar{q}(\mathbb{G}_n(x_{ij}^2 \epsilon_i^2), \frac{1}{2}) \leq \sqrt{2\bar{\mathbf{E}}[x_{ij}^4 \epsilon_i^4]} \leq \sqrt{2} \max_{1 \leq j \leq p} (\mathbb{E}_n[|x_{ij}|^{\frac{4q}{q-4}}])^{\frac{q-4}{2q}} \left(\frac{\bar{\mathbf{E}}[|\epsilon_i|^q]}{\tau_2}\right)^{\frac{2}{q}}.$$

Therefore, by Lemma 18, we have

$$P\left(\max_{1 \leq j \leq p} |\mathbb{E}_n[x_{ij}^2(\epsilon_i^2 - \mathrm{E}[\epsilon_i^2])]| > e_{1n}\right) \leq 4\tau_1 + 4\tau_2.$$

□

LEMMA 10. *Consider* $\epsilon_i \sim t(2)$. *Then, for* $\tau \in (0, 1)$ *we have that:*
(i) $P(\mathbb{E}_n[\epsilon_i^2] \geq 2\sqrt{2}/\tau + \log(4n/\tau)) \leq \tau/2$.
(ii) For $0 < a < 1/6$, *we have* $P(\mathbb{E}_n[\epsilon_i^2] \leq a\log n) \leq \frac{1}{n^{1/2}(1/6 - a)^2}$.
(iii) For $u_n \geq 0$ *and* $0 < a < 1/6$, *we have*

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

PROOF OF LEMMA 10. 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

(C.2)
$$\begin{aligned}
\mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}] &= 2\int_0^{\sqrt{2}} \frac{x^2 dx}{(2+x^2)^{3/2}} + 2\int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^2 dx}{(2+x^2)^{3/2}} \\
&\leq 2\int_0^{\sqrt{2}} \frac{x^2 dx}{2^{3/2}} + 2\int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^2 dx}{x^3} \\
&\leq \log t_n. \\
\mathrm{E}[\epsilon_i^4 1\{\epsilon_i^2 \leq t_n\}] &\leq 2\int_0^{\sqrt{2}} \frac{x^4 dx}{2^{3/2}} + 2\int_{\sqrt{2}}^{\sqrt{t_n}} \frac{x^4 dx}{x^3} \leq t_n. \\
\mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}] &\geq 2\int_0^1 \frac{x^2 dx}{3^{3/2}} + 2\int_1^{\sqrt{2}} \frac{x^2 dx}{4^{3/2}} + \frac{2}{2\sqrt{2}}\int_{\sqrt{2}}^{\sqrt{t_n}} \frac{dx}{x} \\
&\geq \frac{\log t_n}{2\sqrt{2}}.
\end{aligned}$$

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

(C.3)
$$P(|\epsilon_i|^2 > t_n) = \left(1 - \sqrt{\frac{t_n}{2+t_n}}\right) \leq 2/(2+t_n).$$

Thus, by setting $t_n = 4n/\tau$ and $t = 2\sqrt{2}/\tau$ we have [12], relation (7.5),

(C.4)
$$\begin{aligned}
P(|\mathbb{E}_n[\epsilon_i^2] - \mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq t_n\}]| \geq t) &\leq \frac{\mathrm{E}[\epsilon_i^4 1\{\epsilon_i^2 \leq t_n\}]}{nt^2} + nP(|\epsilon_i^2| > t_n) \\
&\leq \frac{t_n}{nt^2} + \frac{2n}{2+t_n} \leq \tau/2.
\end{aligned}$$

Thus, (i) is established.

To show (ii), for $0 < a < 1/6$, we have

(C.5)
$$\begin{aligned}
P(\mathbb{E}_n[\epsilon_i^2] \leq a \log n) &\leq P(\mathbb{E}_n[\epsilon_i^2 1\{\epsilon_i^2 \leq n^{1/2}\}] \leq a \log n) \\
&\leq P(|\mathbb{E}_n[\epsilon_i^2 1\{\epsilon_i^2 \leq n^{1/2}\}] - \mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq n^{1/2}\}]| \geq (\tfrac{1}{6} - a)\log n) \\
&\leq \frac{1}{n^{1/2}(1/6-a)^2}
\end{aligned}$$

by Chebyshev inequality and since $\mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \leq n^{1/2}\}] \geq (1/6)\log n$.

To show (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 (C.2), (C.4), and (C.5) we have

$$\begin{aligned}
P\left(\sqrt{\mathbb{E}_n[\sigma^2 \epsilon_i^2]} > (1+u_n)\sqrt{\mathbb{E}_n[(\sigma\epsilon_i + r_i)^2]}\right) &= P(2\sigma\mathbb{E}_n[\epsilon_i r_i] > c_s^2 + a_n \mathbb{E}_n[\sigma^2 \epsilon_i^2]) \\
&\leq P(2\sigma\mathbb{E}_n[\epsilon_i r_i 1\{\epsilon_i^2 \leq t_n\}] > c_s^2 + a_n \sigma^2 a \log n) + P(\mathbb{E}_n[\epsilon_i^2] \leq a \log n) + nP(\epsilon_i^2 \leq t_n) \\
&\leq \frac{4\sigma^2 c_s^2 \log t_n}{n(c_s^2 + a_n\sigma^2 a \log n)^2} + \frac{1}{n^{1/2}(1/6-a)^2} + \tau/2.
\end{aligned}$$

$\square$

APPENDIX D: LEMMAS FOR PROJECTION ESTIMATORS

LEMMA 11. *Consider the oracle approximation error and the optimal cardinality as*

$$c_k^2 = \min_{\|\beta\|_0 \leq k} \mathbb{E}_n[(f_i - x_i'\beta)^2] \quad and \quad s \in \arg\min_k\{c_k^2 + \sigma^2 k/n \;:\; k \geq 0, k \text{ integer}\}.$$

If the approximation error satisfies $c_k^2 \leq Ck^{-2\alpha}$ for every $k \geq 0$, then

$$s \leq 1 + n^{1/[1+2\alpha]}\frac{1}{\sigma^2}\left(\frac{2\alpha C}{\sigma^2}\right)^{-2\alpha/[1+2\alpha]}.$$

PROOF. Let $\tilde{s} = \arg\min_{k\geq 0} Ck^{-2\alpha}+\sigma^2 k/n$, so that $\tilde{s} = \left(\frac{2\alpha C}{\sigma^2}\right)^{\frac{1}{1+2\alpha}} n^{\frac{1}{1+2\alpha}}$ which might not be an integer.

By definition of s we have

$$c_s^2 + \sigma^2 s/n \leq c_{\lceil\tilde{s}\rceil}^2 + \sigma^2\lceil\tilde{s}\rceil/n \leq C\tilde{s}^{-2\alpha} + \sigma^2\tilde{s}/n + \sigma^2/n.$$

Therefore, we have

$$\begin{aligned} s &\leq 1 + (n/\sigma^2)(2\alpha C/\sigma^2)^{-2\alpha/[1+2\alpha]}n^{-2\alpha/[1+2\alpha]} \\ &= 1 + n^{1/[1+2\alpha]}(1/\sigma^2)(2\alpha C/\sigma^2)^{-2\alpha/[1+2\alpha]}. \end{aligned}$$

$\square$

LEMMA 12.   Consider the nonparametric model (2.1) where $f : [0,1] \to \mathbb{R}$ belongs to the Sobolev class $W(\alpha, L)$, and $z_i \sim \text{Uniform}(0,1), i = 1,\ldots,n$. Given a bounded orthonormal basis $\{P_j(\cdot)\}_{j=1}^\infty$, the coefficients of the projection estimator satisfy for any $k \geq 1$

$$\mathrm{E}[\|\widehat{\theta}^{(k)} - \theta\|_2|z_1,\ldots,z_n] \lesssim_{P_z} k^{-\alpha} + \sqrt{\frac{k}{n}}$$

where $\widehat{\theta}^{(k)} = (\widehat{\theta}_1,\ldots,\widehat{\theta}_k, 0, 0, 0,\ldots)$.

PROOF. Let $Z = [z_1,\ldots,z_n]$ and recall that $y_i = f(z_i) + \sigma\epsilon_i$, $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{E}[\epsilon_i^2] = 1$. Essentially by Proposition 1.16 of [26] we have $\mathrm{E}[\widehat{\theta}_j|Z] = \theta_j + \gamma_j$, where $\gamma_j = \mathbb{E}_n[f(z_i)P_j(z_i)] - \theta_j$, and $\mathrm{E}[(\widehat{\theta}_j - \theta_j)^2|Z] = \mathbb{E}_n[P_j(z_i)^2]\sigma^2/n + \gamma_j^2$.

Since $f(z) = \sum_{m\geq 1}\theta_m P_m(z)$ for any $z \in [0,1]$, we have for $1 \leq j \leq k \leq \bar{k}$

$$\begin{aligned} \gamma_j &= \sum_{m=1}^\infty \theta_m \mathbb{E}_n[P_m(z_i)P_j(z_i)] - \theta_j \\ &= \theta_j(\mathbb{E}_n[P_j^2(z_i)] - 1) + \sum_{m=1,m\neq j}^{\bar{k}} \theta_m\mathbb{E}_n[P_m(z_i)P_j(z_i)] + \\ &\quad + \sum_{m\geq\bar{k}+1}\theta_m\mathbb{E}_n[P_m(z_i)P_j(z_i)]. \end{aligned}$$

Next, note that θ satisfies $\sum_{m=1}^\infty a_m^2\theta_m^2 \leq L$, we have

(D.1) $\sum_{m=1}^{\bar{k}}|\theta_m| \leq (\sum_{m=1}^{\bar{k}} a_m^2\theta_m^2)^{1/2}(\sum_{m=1}^{\bar{k}} a_m^{-2})^{1/2} \leq C_\beta L^{1/2},$
          $\sum_{m=\bar{k}}^\infty|\theta_m| \leq (\sum_{m=1}^\infty a_m^2\theta_m^2)^{1/2}(\sum_{m=\bar{k}}^\infty a_m^{-2})^{1/2} \leq C_\beta L^{1/2}\bar{k}^{-\alpha+1/2}.$

For convenience define $M = \{1, \ldots, \bar{k}\}$ so that

$$\sum_{j=1}^{k} \gamma_j^2 \lesssim \sum_{j=1}^{k} (\mathbb{E}_n[\theta_M' P_M(z_i) P_j(z_i)] - \theta_j)^2 + \sum_{j=1}^{k} \sum_{m \geq \bar{k}+1} |\theta_m| \lesssim_{P_z} \frac{k}{n} + k\bar{k}^{-\alpha+1/2}.$$

Indeed, note that since the basis is bounded and (D.1) holds, we have $|\theta_M' P_M(z_i) P_j(z_i) - \theta_j| \lesssim \|\theta_M\|_1 \lesssim 1$, and thus $Z_j := \mathbb{E}_n[\theta_M' P_M(z_i) P_j(z_i) - \theta_j]$ satisfies $E_Z[Z_j] = 0$ and $E_Z[Z_j^2] \lesssim 1/n$. Hence, by Markov inequality we have

$$\sum_{j=1}^{k} Z_j^2 \lesssim_{P_z} \frac{k}{n}.$$

For some constant $V > 0$, setting $k = \lfloor Vn^{1/[2\alpha+1]} \rfloor$, $\bar{k} = n$, we have

$$\sum_{j=1}^{k} \mathbb{E}_n[P_j(z_i)^2] \frac{\sigma^2}{n} \quad \lesssim \max_{1 \leq j \leq k} \mathbb{E}_n[P_j(z_i)^2] \frac{\sigma^2 k}{n} \lesssim \sigma^2 n^{-1+1/[2\alpha+1]} \lesssim n^{-2\alpha/[2\alpha+1]},$$

$$\sum_{m=k+1}^{\infty} \theta_j^2 \quad \lesssim k^{-2\alpha} \lesssim n^{-2\alpha/[2\alpha+1]},$$
$$\sum_{m=1}^{k} \gamma_m^2 \quad \lesssim_{P_z} \frac{k}{n} + kn^{-2\alpha+1} \lesssim n^{-2\alpha/[2\alpha+1]}$$

where we used the fact that the basis is bounded, $\max_{1 \leq j \leq k} \mathbb{E}_n[P_j(z_i)^2] \lesssim 1$, and $kn^{-2\alpha+1} \leq k/n$ for $\alpha \geq 1$. Finally,

$$\mathrm{E}[\|\widehat{\theta}^{(k)} - \theta\|^2 | Z] \quad \lesssim \sum_{j=1}^{k} \mathbb{E}_n[P_j(z_i)^2] \frac{\sigma^2}{n} + \sum_{m=1}^{k} \gamma_m^2 + \sum_{m \geq k+1} \theta_m^2$$
$$\lesssim_{P_z} \frac{k}{n} + k^{-2\alpha}$$

by the relations above. The result follows by Jensen's inequality. $\qquad\square$

APPENDIX E: EMPIRICAL PERFORMANCE OF $\sqrt{\text{LASSO}}$

E.1. Estimation performance of $\sqrt{\text{lasso}}$, homoskedastic. In this section we use Monte carlo experiments to assess the finite-sample performance of the following estimators:

- the (infeasible) lasso, which knows σ (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 σ, and
- ols post $\sqrt{\text{lasso}}$, which applies ols to the model selected by $\sqrt{\text{lasso}}$.

In the homoskedastic case there is no need to estimate the loadings so we set $\widehat{\gamma}_j = 1$ for all $j = 1, \ldots, p$. We set the penalty level for lasso as the standard choice in the literature, $\lambda = c2\sigma\sqrt{n}\Phi^{-1}(1 - \alpha/2p)$, and $\sqrt{\text{lasso}}$ according to $\lambda = c\sqrt{n}\Phi^{-1}(1 - \alpha/2p)$, both with $1 - \alpha = .95$ and $c = 1.1$ to both estimators.

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

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

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

(E.1) $\qquad f(x_i) = x_i'\beta_0^*, \quad \text{where} \quad \beta_{0j}^* = 1/j^{3/2}, \quad j = 1, \ldots, p.$

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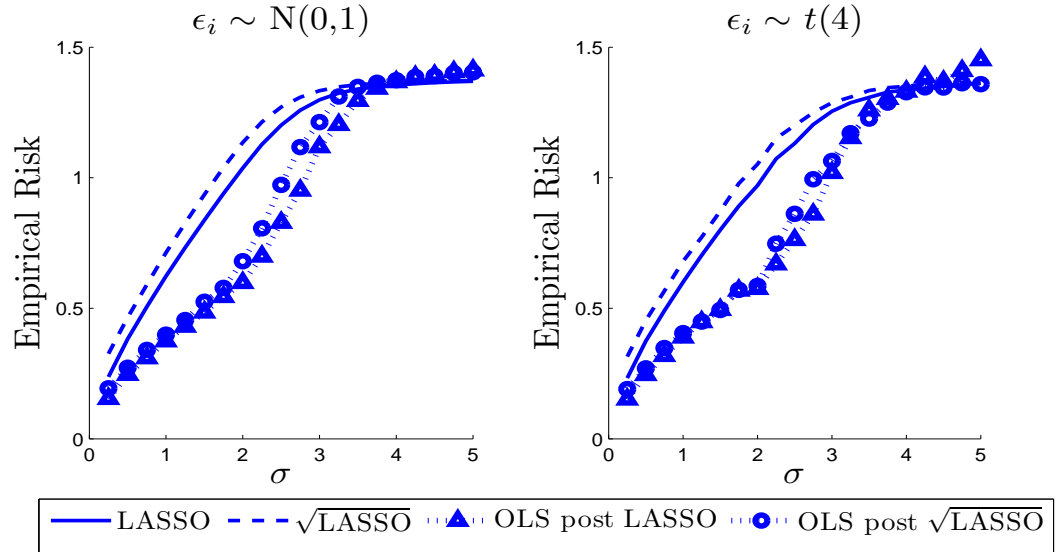

FIG 1. *The average empirical risk of the estimators as a function of the scaling parameter* σ.

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
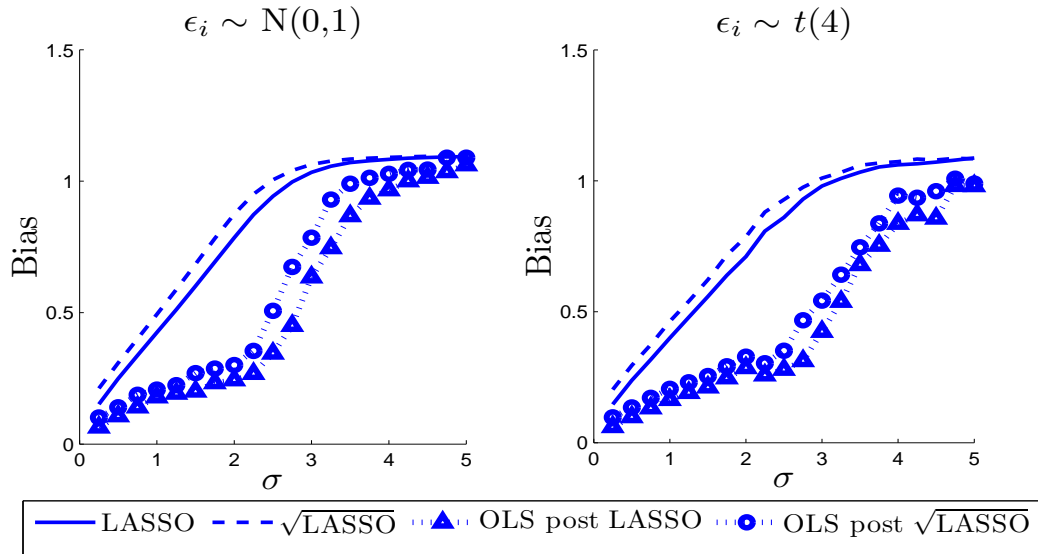
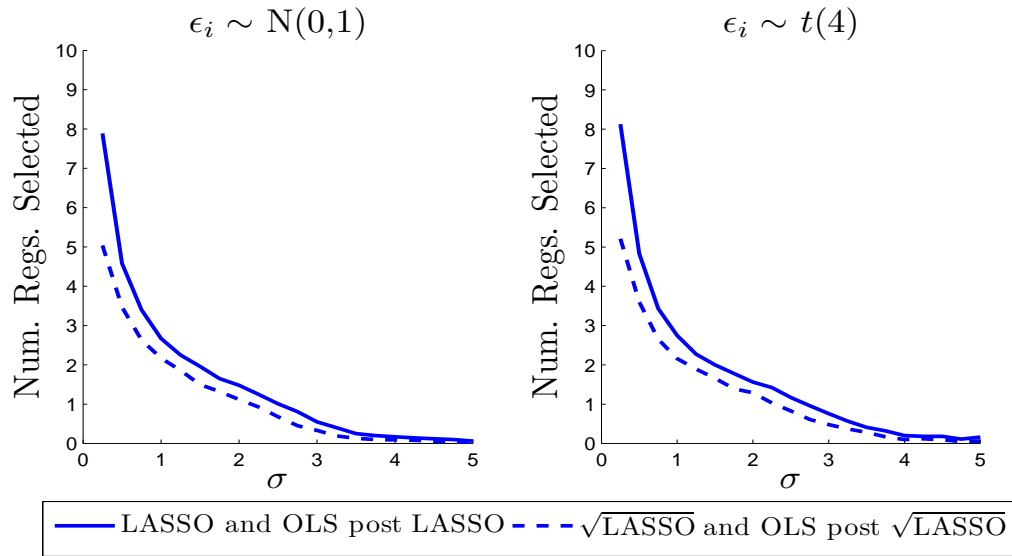FIG 2. *The norm of the bias of the estimators as a function of the scaling parameter σ.*

FIG 3. *The average number of regressors selected as a function of the scaling parameter σ.*

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 scaling parameter σ for each estimator $\widetilde{\beta}$:

- Figure 1 – the average empirical risk, $\mathrm{E}[\|\widetilde{\beta} - \beta_0\|_{2,n}]$,
- Figure 2 – the norm of the bias, $\|\mathrm{E}[\widetilde{\beta} - \beta_0]\|$, and
- Figure 3 – the average number of regressors selected, $\mathrm{E}[|\mathrm{support}(\widetilde{\beta})|]$.

Figure 1, left panel, shows the empirical risk for the Gaussian case. We see that, for a wide range of the scaling parameter σ, lasso and $\sqrt{\mathrm{lasso}}$ perform similarly in terms of empirical risk, although standard lasso outperforms somewhat $\sqrt{\mathrm{lasso}}$. At the same time, ols post lasso outperforms slightly ols post $\sqrt{\mathrm{lasso}}$ for larger signal strengths. This is expected since $\sqrt{\mathrm{lasso}}$ over regularize to simultaneously estimate σ when compared to lasso (since it essentially uses $\sqrt{\widehat{Q}(\widehat{\beta})}$ as an estimate of σ). 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 $\sqrt{\mathrm{lasso}}$ and ols post $\sqrt{\mathrm{lasso}}$ achieve its goal. Despite not knowing σ, $\sqrt{\mathrm{lasso}}$ 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 $\sqrt{\mathrm{lasso}}$ 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 $\sqrt{\mathrm{lasso}}$ arise primarily due to $\sqrt{\mathrm{lasso}}$ having a larger bias than standard lasso. This bias arises because $\sqrt{\mathrm{lasso}}$ uses an effectively heavier penalty. Figure 3 shows that such heavier penalty translates into $\sqrt{\mathrm{lasso}}$ 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 $\sqrt{\mathrm{lasso}}$ under $t(4)$ errors nearly coincides with their performance under Gaussian errors. This is exactly what is predicted by our theoretical results.

E.2. Estimation performance of $\sqrt{\mathrm{lasso}}$, heteroskedastic.

In this section we use Monte carlo experiments to assess the finite-sample performance under heteroskedastic errors of the following estimators:

- the (infeasible) oracle estimator,
- heteroskedastic $\sqrt{\mathrm{lasso}}$ (as Algorithm 1),
- ols post heteroskedastic $\sqrt{\mathrm{lasso}}$, which applies ols to the model selected by heteroskedastic $\sqrt{\mathrm{lasso}}$.

- the (infeasible) ideal heteroskedastic $\sqrt{\text{lasso}}$ (which uses exact loadings),
- ols post ideal heteroskedastic $\sqrt{\text{lasso}}$, which applies ols to the model selected by ideal heteroskedastic $\sqrt{\text{lasso}}$.

We use the linear regression model stated in the introduction as a data-generating process. We set the regression function as

(E.2) $\qquad f(x_i) = x_i'\beta_0^*, \quad \text{where} \quad \beta_{0j}^* = 1/j^2, \quad j = 1, \ldots, p.$

The error term ϵ_i is normal with zero mean and variance given by:

$$\sigma_i^2 = \sigma^2 \frac{|1 + x_i'\beta_0^*|^2}{\mathbb{E}_n[\{1 + x_i'\beta_0^*\}^2]}$$

where the scaling parameter σ vary between 0.1 and 1. For the fixed design, as the scaling parameter σ increases, the number of non-zero components in the oracle vector s decreases. The number of regressors $p = 200$, the sample size $n = 200$, and we used 500 simulations for each design. We generate regressors as $x_i \sim N(0, \Sigma)$ with the Toeplitz correlation matrix $\Sigma_{jk} = (1/2)^{|j-k|}$. We set the penalty level $\sqrt{\text{lasso}}$ according to the recommended parameters of Algorithm 1.

Figure 4 displays the average sparsity achieve by each estimator and the average empirical risk. The heteroskedastic $\sqrt{\text{lasso}}$ exhibits a stronger degree of regularization. This is reflected by the smaller number of components selected and the substantially larger empirical risk. Nonetheless, the selected support seems to achieve good approximation performance since the ols post heteroskedastic $\sqrt{\text{lasso}}$ performs very close to its ideal counterpart and to the oracle.

APPENDIX F: COMPARING COMPUTATIONAL METHODS FOR LASSO AND $\sqrt{\text{LASSO}}$

Next we proceed to evaluate the computational burden of $\sqrt{\text{lasso}}$ relative to lasso, from computational and theoretical perspective.

F.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

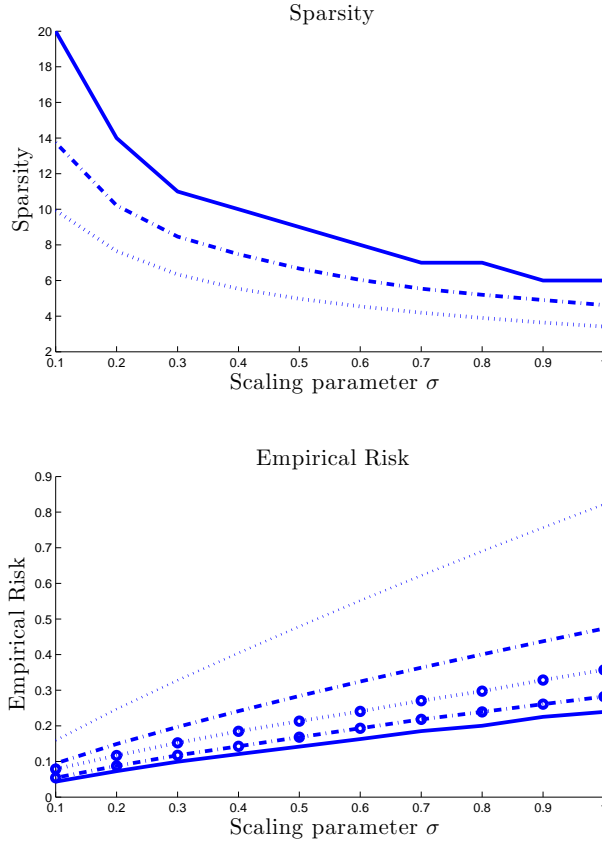$$\begin{aligned} \min_x \quad & c(x) \\ & A(x) = b \\ & x \in K \end{aligned}$$

Fig 4. *For each estimator the top figure displays the corresponding sparsity and the bottom figure displays the empirical risk as a function of the scaling parameter σ. The solid line corresponds to the oracle estimator, the dotted line corresponds to the heteroskedastic $\sqrt{lasso}$, the dashed-dot line corresponds to the ideal heteroskedastic $\sqrt{lasso}$. The dotted line with circles corresponds to ols post heteroskedastic $\sqrt{lasso}$ and the dashed-dotted line with circles corresponds to ols post ideal heteroskedastic $\sqrt{lasso}$.*

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 in the case where K is also convex. Convex conic programming problems have greatly extended the scope of applications of linear programming problems[3]

---

[3]The relevant cone in linear programs is the non-negative orthant, $\min_w \{c'w : Aw = b, w \in \mathbb{R}_+^k\}$.

in several fields including optimal control, learning theory, eigenvalue optimization, combinatorial optimization and others. Under mild regularities conditions, duality theory for conic programs has 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 [21, 22], and efficient software implementations [27] made large instances of these problems computational tractable. More recently, first-order methods have also been propose to approximately solve even larger instances of structured conic problem [19, 20, 17].

It follows that (2.3) 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 $\mathbb{R}^{n+1}$, we can recast (2.3) as the following conic program:

$$(\text{F.1}) \quad \min_{t,v,\beta^+,\beta^-} \quad \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \sum_{i=1}^{p} \left(\gamma_j \beta_j^+ + \gamma_j \beta_j^- \right)$$
$$v_i = y_i - x_i'\beta^+ + x_i'\beta^-, \ i = 1, \dots, n$$
$$(t, v) \in Q^{n+1}, \ \beta^+ \geq 0, \ \beta^- \geq 0.$$

Conic duality immediately yields the following dual problem

$$(\text{F.2}) \quad \max_{a \in \mathbb{R}^n} \quad \mathbb{E}_n \left[y_i a_i \right]$$
$$|\mathbb{E}_n \left[x_{ij} a_i \right]| \leq \lambda \gamma_j / n, \ j = 1, \dots, p$$
$$\|a\| \leq \sqrt{n}.$$

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 $\widehat{\beta}_j$ 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 compute the $\sqrt{\text{lasso}}$ estimator. In this section we compare three

$n = 100,\ p = 500$	Componentwise	First-order	Interior-point
lasso	0.2173	10.99	2.545
$\sqrt{\text{lasso}}$	0.3268	7.345	1.645

$n = 200,\ p = 1000$	Componentwise	First-order	Interior-point
lasso	0.6115	19.84	14.20
$\sqrt{\text{lasso}}$	0.6448	19.96	8.291

$n = 400,\ p = 2000$	Componentwise	First-order	Interior-point
lasso	2.625	84.12	108.9
$\sqrt{\text{lasso}}$	2.687	77.65	62.86

TABLE 1

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

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 F.1 illustrates that the average computational time to solve lasso and $\sqrt{\text{lasso}}$ optimization problems are comparable. Table F.1 also reinforces typical behavior of these methods. As the size increases, the running time for interior-point method grows faster than other first-order method. Simple componentwise method is particular effective when the solution is highly sparse. This is the case of the parametric design considered in these experiments. We emphasize the performance of each method depends on the particular design and choice of λ.

F.2. discussion of Implementation Details. 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. In what follows we were given data $\{Y = [y_1, \ldots, y_n]', X = [x_1, \ldots, x_n]'\}$ and penalty $\{\lambda, \Gamma = \text{diag}(\gamma_1, \ldots, \gamma_p)\}$.

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

$$\text{(F.3)} \qquad \min_{w} c'w \ : \ Aw = b, w \in K.$$

The main difficulty of the problem arises because the conic constraint will be biding 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, a IPM creates a sequence of solutions that converges to the solution of the original problem (F.3).

In order to formulate the optimization problem associated with the lasso estimator as a conic programming problem (F.3), 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

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

Thus, we have that lasso optimization problem can be cast

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

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

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

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

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

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

$$\beta_j(\beta^k) = \text{sign}\left(\beta_j^+\right)\max\left\{\left|\beta_j^+\right| - \lambda\gamma_j/[n\mu],\ 0\right\}$$

where $\beta_j^+ = \beta_j^k + 2\mathbb{E}_n[x_{ij}(y_i - x_i'\beta^k)]/\mu$.

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 $h(w) = f(\beta,t) = t/\sqrt{n} + (\lambda/n)\|\Gamma\beta\|_1$ we have that

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

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

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

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

$$\beta_j(\beta^k, z^k) = \text{sign}\left(\beta_j^k + (X'z^k/\mu)_j\right)\max\left\{\left|\beta_j^k + (X'z^k/\mu)_j\right| - \lambda\gamma_j/[n\mu],\ 0\right\}.$$

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

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

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 particulary 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 $\mathbb{E}_n[x_{ij}^2] = 1$ for $j = 1,\ldots,p$. Below we describe the rule to set optimally the value of β_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 β, 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 β_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 β_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 $\lambda \geq cn\|S\|_{\infty}$.

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 $\mathbb{E}_n[x_{ij}^2] = 1$ for $j = 1, \ldots, p$. Below we describe the rule to set optimally the value of β_j given fixed the values of the remaining variables.

- If $\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] > (\lambda/n)\gamma_j\sqrt{\widehat{Q}(\beta_{-j})}$, we have

$$\beta_j = -\frac{\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]}{\mathbb{E}_n[x_{ij}^2]} + \frac{\lambda\gamma_j}{\mathbb{E}_n[x_{ij}^2]}\frac{\sqrt{\widehat{Q}(\beta_{-j}) - (\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})])^2/\mathbb{E}_n[x_{ij}^2])}}{\sqrt{n^2 - (\lambda^2\gamma_j^2/\mathbb{E}_n[x_{ij}^2])}};$$

- if $\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})] < -(\lambda/n)\gamma_j\sqrt{\widehat{Q}(\beta_{-j})}$, we have

$$\beta_j = -\frac{\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})]}{\mathbb{E}_n[x_{ij}^2]} - \frac{\lambda\gamma_j}{\mathbb{E}_n[x_{ij}^2]}\frac{\sqrt{\widehat{Q}(\beta_{-j}) - (\mathbb{E}_n[x_{ij}(y_i - x_i'\beta_{-j})])^2/\mathbb{E}_n[x_{ij}^2])}}{\sqrt{n^2 - (\lambda^2\gamma_j^2/\mathbb{E}_n[x_{ij}^2])}};$$

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

APPENDIX G: PROBABILITY INEQUALITIES

G.1. Moment Inequalities. We begin with Rosenthal and Von Bahr-Esseen Inequalities.

LEMMA 13 (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 5 (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 $\ell_n > 0$*

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

where $C(r)$ is a constant depend only on r.

Remark. To verify the corollary, note that by Rosenthal's inequality we have $E\left[|\sum_{i=1}^n X_i|^r\right] \leq Cn^{r/2}$. By Markov inequality,

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

so the corollary follows. We refer [24] for complete proofs.

LEMMA 14 (Vonbahr-Esseen inequality). *Let $X_1, \ldots, X_n$ be independent zero-mean random variables. Then for $1 \le r \le 2$*

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

We refer to [32] for proofs.

COROLLARY 6 (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 $\ell_n > 0$*

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

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

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

which implies the corollary.

G.2. Moderate Deviations for Sums of Independent Random Variables. Next we consider Slastnikov-Rubin-Sethuraman Moderate Deviation Theorem.

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

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

LEMMA 15 (Slastnikov, Theorem 1.1). *If for sufficiently large n and some positive constant c,*

$$\sum_{i=1}^{k_n} E[|X_{ni}|^{2+c^2}]\rho(|X_{ni}|)\log^{-(1+c^2)/2}(3 + |X_{ni}|) \le g(B_n)B_n^2,$$

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 $0 \le x \le c\sqrt{\log B_n^2}$

COROLLARY 7 (Slastnikov, Rubin-Sethuraman).   *If $q > c^2 + 2$ and*

$$\sum_{i=1}^{k_n} E[|X_{ni}|^q] \leq K B_n^2,$$

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

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

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 [25] for proofs.

G.3. Moderate Deviations for Self-Normalized Sums.   We shall be using the following result – Theorem 7.4 in [10].

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

$$B_n^2 = \sum_{i=1}^{n} \mathrm{E} X_i^2, \quad L_{n,\delta} = \sum_{i=1}^{n} \mathrm{E}|X_i|^{2+\delta}, \quad d_{n,\delta} = B_n / L_{n,\delta}^{1/(2+\delta)}.$$

Then for uniformly in $0 \leq x \leq d_{n,\delta}$,

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

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

where the terms $O(1)$ are bounded in absolute value by a universal constant A, and $\bar{\Phi} := 1 - \Phi$.

Application of this result gives the following lemma:

LEMMA 16 (Moderate deviations for self-normalized sums). *Let* $X_{1,n}, \ldots, X_{n,n}$ *be a triangular array of i.n.i.d, zero-mean random variables. Suppose that*

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

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

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

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 ℓ_n and M_n. □

G.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 $\|f\|_{\mathbb{P}_n,2} = \sqrt{\mathbb{E}_n[f(Z_i)^2]}$, $\mathbb{G}_n(f) = (1/\sqrt{n}) \sum_{i=1}^n \{f(Z_i) - E[f(Z_i)]\}$, and $\mathbb{G}_n^o(f) = (1/\sqrt{n}) \sum_{i=1}^n \varepsilon_i f(Z_i)$ where ε_i are independent Rademacher random variables.

In order to prove a bound on tail probabilities of a 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 $x : \mathbb{R}^n \mapsto \mathbb{R}$ 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 2.3.7 of [31]) to the case of a random threshold x that is sub-exchangeable. The proof of Lemma 17 can be found in [3].

LEMMA 17 (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_τ denote the τ 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 $\inf_{f \in \mathcal{F}} P\left(|\sum_{i=1}^n Z_i(f)| \leq \frac{x_0}{2}\right) \geq 1 - \frac{\bar{p}_\tau}{2}$.

THEOREM 11 (Maximal Inequality for Empirical Processes).   *Let*

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

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

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

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 17 we obtain

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

Thus a union bound yields

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

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 $\mathbb{G}_n^o$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$ norm,

namely, for $f \in \mathcal{F}$, $\mathbb{P}_\varepsilon\{|\mathbb{G}_n^o(f)| > x\} \leq 2\exp(-x^2/[2\|f\|_{\mathbb{P}_n,2}^2])$. Hence, we can bound

$$
\begin{aligned}
\mathbb{P}_\varepsilon\left\{|\mathbb{G}_n^o(f)| \geq Ce_n(\mathcal{F}, \mathbb{P}_n)|Z_1, \ldots, Z_n\right\} &\leq 2\exp(-C^2 e_n(\mathcal{F}, \mathbb{P}_n)^2/[2\|f\|_{\mathbb{P}_n,2}^2]) \\
&\leq 2\exp(-C^2 \log|\mathcal{F}|).
\end{aligned}
$$

Taking the expectation over $Z_1, \ldots, Z_n$ does not affect the right hand side bound. Plugging in this bound into relation (G.1) yields the result.

Step 2. (Auxiliary calculations.) To establish that $\sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n,2}$ 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

$$
\begin{aligned}
\sqrt{2}(\sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(\tilde{Z}),2} \vee \sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(\tilde{Y}),2}) &\geq (\sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(\tilde{Z}),2}^2 + \sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(\tilde{Y}),2}^2)^{1/2} \\
&\geq (\sup_{f \in \mathcal{F}}\mathbb{E}_n[f(\tilde{Z}_i)^2] + \mathbb{E}_n[f(\tilde{Y}_i)^2])^{1/2} = (\sup_{f \in \mathcal{F}}\mathbb{E}_n[f(Z_i)^2] + \mathbb{E}_n[f(Y_i)^2])^{1/2} \\
&\geq (\sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(Z),2}^2 \vee \sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(Y),2}^2)^{1/2} = \sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(Z),2} \vee \sup_{f \in \mathcal{F}}\|f\|_{\mathbb{P}_n(Y),2}.
\end{aligned}
$$

$\square$

COROLLARY 8 (Data-dependent probability inequality). *Let ϵ_i be i.i.d random variables such that $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{E}[\epsilon_i^2] = \sigma^2$ 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 $1 - 1/[9C^2 \log p]$,*

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

PROOF OF COROLLARY 8. Consider the class of separable empirical process induced by $\|\mathbb{E}_n[x_i\epsilon_i]\|_\infty$, i.e. the class of functions $f \in \mathcal{F} = \{\epsilon_i x_{ij} : j \in \{1, \ldots, p\}\}$ so that $\sqrt{n}\|\mathbb{E}_n[x_i\epsilon_i]\|_\infty = \sup_{f \in \mathcal{F}}|\mathbb{G}_n(f)|$. Define the data dependent quantity

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

Then, by Theorem 11, for any constant $C \geq 1$

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

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

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

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

$$\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| \leq \left(6 \frac{C}{3} \sqrt{\log p} \max_{j=1,\ldots,p} \sqrt{\mathrm{E}[\epsilon_i^2 x_{ij}^2]} \right) \vee \left(24 \frac{C}{3} \sqrt{\log p} \max_{j=1,\ldots,p} \sqrt{\mathbb{E}_n[\epsilon_i^2 x_{ij}^2]} \right).$$

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

G.5. Bounds via Symmetrization. Next we proceed to use symmetrization arguments to bound the empirical process. Let

$$\|f\|_{\mathbb{P}_n,2} = \sqrt{\mathbb{E}_n[f(X_i)^2]}, \ \mathbb{G}_n(f) = \sqrt{n}\mathbb{E}_n[f(X_i) - \mathrm{E}[f(X_i)]],$$

and for a random variable Z let $q(Z, 1 - \tau)$ denote its $(1 - \tau)$-quantile.

LEMMA 18 (Maximal inequality via symmetrization). *Let $Z_1, \ldots, Z_n$ be arbitrary independent stochastic processes and $\mathcal{F}$ a finite set of measurable functions. For any $\tau \in (0, 1/2)$, and $\delta \in (0, 1)$ we have that with probability at least $1 - 4\tau - 4\delta$*

$$\max_{f \in \mathcal{F}} |\mathbb{G}_n(f(Z_i))| \quad \leq \max \left\{ 4\sqrt{2 \log(2|\mathcal{F}|/\delta)} \ q \left(\max_{f \in \mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]}, 1 - \tau \right), \right.$$
$$\left. 2 \max_{f \in \mathcal{F}} q \left(|\mathbb{G}_n(f(Z_i))|, \tfrac{1}{2} \right) \right\}.$$

REFERENCES

[1] Stephen Becker, Emmanuel Candès, and Michael Grant. Templates for convex cone problems with applications to sparse signal recovery. *ArXiv*, 2010.

[2] A. Belloni and V. Chernozhukov. Post-ℓ_1-penalized estimators in high-dimensional linear regression models. *arXiv:[math.ST]*, 2009.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *accepted at the Annals of Statistics*, 2010.

[4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *arXiv:[math.ST]*, 2010.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

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

[7] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.

[8] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[9] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[10] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.

[11] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B*, 70(5):849–911, 2008.

[12] W. Feller. *An Introduction to Probability Theory and Its Applicaitons, Volume II.* Wiley Series in Probability and Mathematical Statistics, 1966.

[13] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized cramr-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[14] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.

[15] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.

[16] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.

[17] Z. Lu. Gradient based method for cone programming with application to large-scale compressed sensing. *Technical Report*, 2008.

[18] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.

[19] Y. Nesterov. Smooth minimization of non-smooth functions, mathematical programming. *Mathematical Programming*, 103(1):127–152, 2005.

[20] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[21] Y. Nesterov and A Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1993.

[22] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2001.

[23] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.

[24] H. P. Rosenthal. On the subspaces of l^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 9:273–303, 1970.

[25] A. D. Slastnikov. Limit theorems for moderate deviation probabilities. *Theory of Probability and its Applications*, 23:322–340, 1979.

[26] A. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2008.

[27] R. H. Tütüncü, K. C. Toh, and M. J. Todd. SDPT3 — a MATLAB software package for semidefinite-quadratic-linear programming, version 3.0. Technical report, 2001. Available at http://www.math.nus.edu.sg/∼mattohkc/sdpt3.html.

[28] S. A. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.

[29] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

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

[31] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer Series in Statistics, 1996.

[32] Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \le r \le 2$. *Ann. Math. Statist*, 36:299–303, 1965.

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

[34] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.