Belloni, Alexandre; Chernozhukov, Victor; Wang, Lie

Working Paper

Pivotal estimation via square-root lasso in nonparametric regression

cemmap working paper, No. CWP62/13

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Belloni, Alexandre; Chernozhukov, Victor; Wang, Lie (2013) : Pivotal estimation via square-root lasso in nonparametric regression, cemmap working paper, No. CWP62/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.6213

This Version is available at:
https://hdl.handle.net/10419/97392

cemmap
centre for microdata methods and practice

Pivotal estimation via square-root lasso in nonparametric regression

Alexandre Belloni
Victor Chernozhukov
Lie Wang

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP62/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Submitted to the Annals of Statistics

PIVOTAL ESTIMATION VIA SQUARE-ROOT LASSO IN NONPARAMETRIC REGRESSION*

By Alexandre Belloni, Victor Chernozhukov and Lie Wang

We propose a self-tuning $\sqrt{\text{Lasso}}$ method that simultaneously resolves three important practical problems in high-dimensional regression analysis, namely it handles the unknown scale, heteroscedasticity, and (drastic) non-Gaussianity of the noise. In addition, our analysis allows for badly behaved designs, for example perfectly collinear regressors, and generates sharp bounds even in extreme cases, such as the infinite variance case and the noiseless case, in contrast to Lasso. We establish various non-asymptotic bounds for $\sqrt{\text{Lasso}}$ including prediction norm rate and sharp sparsity bound. Our analysis is based on new impact factors that are tailored to establish prediction rates. In order to cover heteroscedastic 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. As an application, we consider the use of $\sqrt{\text{Lasso}}$ and post $\sqrt{\text{Lasso}}$ as estimators of nuisance parameters in a generic semi-parametric problem (nonlinear instrumental/moment condition or Z-estimation problem).

1. Introduction.

We consider a nonparametric regression model:

$$y_i = f(z_i) + \sigma\epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

where y_i's are the outcomes, z_i's are vectors of fixed basic covariates, ϵ_i's are independent noise, f is the regression function, and σ is an unknown scaling parameter. The goal is to recover the values $(f_i)_{i=1}^n = (f(z_i))_{i=1}^n$ of the regression function f at z_i's. To achieve this goal, we use linear combinations of technical regressors $x_i = P(z_i)$ to approximate f, where $P(z_i)$ is a dictionary of p-vector of transformations of z_i. We are interested in the high dimension low sample size case, where we potentially have $p > n$, to attain a flexible functional form. In particular, we are interested in a sparse model over the technical regressors x_i to describe the regression function.

*First arXiv version: 7 May 2011; current version: December 9, 2013.

AMS 2000 subject classifications: Primary 62G05, 62G08; secondary 62G35

Keywords and phrases: pivotal, square-root Lasso, model selection, non-Gaussian heteroscedastic, generic semi-parametric problem, nonlinear instrumental variable, Z-estimation problem, $\sqrt{n}$-consistency and asymptotic normality after model selection

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 defined as a solution of an optimization problem to compute the oracle risk, which balances bias and variance (see Section 2) . 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, namely that $s \ll n$, makes it possible to estimate these models effectively by searching for approximately the right set of the regressors. In particular, ℓ_1-penalization have played a central role [14, 15, 20, 34, 40, 53, 61, 59]. 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 [14, 15, 17, 18, 16, 25, 37, 38, 46, 53] for a more detailed review of the existing literature which has been focusing on the homoscedastic case.

In this paper, we attack the problem of nonparametric regression under non-Gaussian, heteroscedastic errors ϵ_i, having an unknown scale σ. We propose to use a self-tuning $\sqrt{\text{Lasso}}$ which is pivotal with respect to the scaling parameter σ, and which handles non-Gaussianity and heteroscedasticity in the errors. The resulting rates and performance guarantees are very similar to the Gaussian case, thanks due to the use of self-normalized moderate deviation theory. Such results and properties,[1] particularly the pivotality with respect to the scale, are in contrast to the previous results and methods on others ℓ_1-regularized methods, for example Lasso and Dantzig selector that use penalty levels that depend linearly on the unknown scaling parameter σ.

There is now a growing literature on high-dimensional linear models[2] allowing for unknown scale σ. [48] propose a ℓ_1-penalized maximum likelihood estimator for parametric Gaussian regression models. [12] considers $\sqrt{\text{Lasso}}$ for a parametric homoscedastic model with both Gaussian and non-Gaussian errors and establish that the choice of the penalty parameter in $\sqrt{\text{Lasso}}$ becomes pivotal with respect to σ. [49] considers an equivalent formulation

[1] Earlier literature, e.g. in bounded designs, [15] provide bounds using refinements of Nemirovski's inequality, see [36]. These provide rates as good as in the Gaussian case. However, when the design is unbounded (e.g. regressors generated as realizations of Gaussian), the rates of convergence provided by these techniques are no longer sharp. The use of self-normalized moderate deviations in the present context allows to handle the latter cases, with sharp rates.

[2] There is also a literature on penalized median regression, which can be used in the case of symmetric errors, since some methods are independent of the unknown σ, cf. [4, 60].

of the (homoscedastic) $\sqrt{\text{Lasso}}$ to establish finite sample results and derives primitive results in the parametric homoscedastic Gaussian setting. [22] consider scaled Fused Dantzig selector to allow for different sparsity patterns and provide primitive results under homoscedastic Gaussian errors. [6] studies Lasso with a plug-in estimator based on Lasso iterations in a parametric homoscedastic setting. [23] studies plug-in estimators and a trade-off penalty choice between fit and penalty in the parametric case with homoscedastic Gaussian errors under random support assumption (similarly to [19]) using coherence condition. In a trace regression model for recovery of a matrix, [33] proposed and analysed a version of the square-root lasso under homoscedasticity. A comprehensive review on this literature is given in [29]. All these works rely essentially on the restricted eigenvalue condition [14] and homoscedasticity and do not differentiate penalty levels across components.

In order to address the nonparametric, heteroscedastic, and non-Gaussian cases, we develop covariate-specific penalty loadings. To derive a practical and theoretically justified choice of penalty level and loadings, we need to account for the impact of the approximation error. We rely on moderate deviation theory for self-normalized sums of [32] 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 [12]. (In the context of standard Lasso, the self-normalized moderated deviation theory was first employed in [3].)

Our first contribution is the proposal of new design and noise impact factors, in order to allow for more general designs. Unlike previous conditions, these factors are tailored for establishing performance bounds with respect to the prediction norm, which is appealing in nonparametric problems. In particular, collinear designs motivate our new condition. In studying their properties we further exploit the oracle based definition of the approximating function. (For instance, our results for rates in prediction norm remain unaffected if repeated regressors are added.) The analysis based on these impact factors complements the analysis based on restricted eigenvalue proposed in [14] and compatibility condition in [54], which are more suitable for establishing rates for ℓ_k-norms.

The second contribution is a set of finite sample upper bounds and lower bounds for estimation errors under prediction norm, and upper bounds on the sparsity of the $\sqrt{\text{Lasso}}$ estimator. These results are "geometric," in that they hold conditional on the design and errors provided some key events occur. We further develop primitive sufficient conditions that allow for these results to be applied to heteroscedastic non-Gaussian errors. We also give results for other norms in the Supplementary Material (SM).

The third contribution develops properties of the estimator that applies ordinary least squares (ols) to the model selected by $\sqrt{\text{Lasso}}$. Our focus is on the case that $\sqrt{\text{Lasso}}$ fails to achieve perfect model selection, including cases where the oracle model is not completely selected by $\sqrt{\text{Lasso}}$. This is usually the case in a nonparametric setting. This estimator intends to remove the potentially significant bias towards zero introduced by the ℓ_1-norm regularization employed in the $\sqrt{\text{Lasso}}$ estimator.

The fourth contribution is to study two extreme cases: (i) parametric noiseless case and (ii) nonparametric infinite variance case. $\sqrt{\text{Lasso}}$ does have interesting theoretical guarantees for these two extreme cases. For case (i) $\sqrt{\text{Lasso}}$ achieves exact recovery in sharp contrast to Lasso, under some conditions. For case (ii), $\sqrt{\text{Lasso}}$ estimator can still be consistent with penalty choice that does not depend on the scale of the noise. We develop the necessary modifications of the penalty loadings and derive finite-sample bounds for the case of symmetric noise. We provide specific bounds to the case of Student's t-distribution with 2 degrees of freedom where Gaussian-noise rates up to a factor of $\log^{1/2} n$.

The final contribution is to provide an application of $\sqrt{\text{Lasso}}$ methods to a generic semi-parametric problem, where some low-dimensional parameters are of interest and these methods are used to estimate nonparametric nuisance parameters. These results extend the $\sqrt{n}$ consistency and asymptotic normality results of [8, 3] on a rather specific linear model to a generic nonlinear problem, which covers covers smooth frameworks in statistics and in econometrics, where the main parameters of interest are defined via nonlinear instrumental variable/moment conditions or z-conditions containing unknown nuisance functions (as in [21]). This and all the above results illustrate the wide applicability of the proposed estimation procedure.

Notation. To make asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we allow for $s = s_n \to \infty$. In what follows, all parameters are indexed by the sample size n, but we omit the index whenever it does not cause confusion. We work with i.n.i.d, independent but not necessarily identically distributed data $(w_i)_{i=1}^n$, with k-dimensional real vectors w_i containing $y_i \in \mathbb{R}$ and $z_i \in \mathbb{R}^{p_z}$, the latter taking values in a set $\mathcal{Z}$. 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. The transpose of a matrix A is denoted by A'. 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. For a measurable function $f : \mathbb{R}^k \to \mathbb{R}$, the symbol $\mathrm{E}[f(w_i)]$ denotes the expected value of $f(w_i)$; $\mathbb{E}_n[f(w)]$ denotes the average $n^{-1} \sum_{i=1}^n f(w_i)$;

$\bar{\mathbf{E}}[f(w)]$ denotes the average expectation $n^{-1}\sum_{i=1}^n \mathrm{E}[f(w_i)]$; and $\mathbb{G}_n(f(w))$ denotes $n^{-1/2}\sum_{i=1}^n (f(w_i) - \mathrm{E}[f(w_i)])$. We will work with regressor values $(x_i)_{i=1}^n$ generated via $x_i = P(z_i)$, where $P(\cdot) : \mathcal{Z} \mapsto \mathbb{R}^p$ is a measurable dictionary of transformations, where p is potentially larger than n. We define the prediction norm of a vector $\delta \in \mathbb{R}^p$ as $\|\delta\|_{2,n} = \{\mathbb{E}_n[(x'\delta)^2]\}^{1/2}$, and given values $y_1, \ldots, y_n$ we define $\widehat{Q}(\beta) = \mathbb{E}_n[(y - x'\beta)^2]$. 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 $a = O_P(b)$. Φ denotes the cumulative distribution of a standard Gaussian distribution and Φ^{-1} its inverse function.

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

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

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 $\mathbb{E}_n[x_j^2] = 1$ $(j = 1, \ldots, p)$.

The goal is to estimate the value of the nonparametric regression function f at the design points, namely the values $(f_i)_{i=1}^n := (f(z_i))_{i=1}^n$. In many applications of interest, especially in the nonparametric settings, there is no exact sparse model or, due to noise. However, there might be a sparse model $x_i'\beta_0$ that yields a good approximation to the true regression function f in equation (2.1). One way to find such approximating model is to let β_0 be a solution of the following risk minimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(f - x'\beta)^2] + \frac{\sigma^2\|\beta\|_0}{n}. \tag{2.2}$$

The problem (2.2) yields the so called oracle risk – an upper bound on the risk of the best k-sparse least squares estimator in the case of homoscedastic 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. The solution β_0 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).

In what follows, we call β_0 the target parameter value, $T := \mathrm{supp}(\beta_0)$ the oracle model, $s := |T| = \|\beta_0\|_0$ the dimension of the oracle model, and $x_i'\beta_0$ the oracle or the best sparse approximation to f_i. We note that T is generally unknown. We summarize the preceding discussion as follows.

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 basic covariates, the regressors* $x_i := P(z_i)$ *are transformations of* z_i*, and* ϵ_i *are i.n.i.d. errors. The vector* β_0 *is defined by (2.2) where the regressors* x_i *are normalized so that* $\mathbb{E}_n[x_j^2] = 1$*,* $j = 1, \ldots, p$*. We let*

$$T := \operatorname{supp}(\beta_0), \quad s := |T|, \quad r_i := f_i - x_i'\beta_0, \quad \text{and} \quad c_s^2 := \mathbb{E}_n[r^2]. \tag{2.3}$$

REMARK 1 (Targeting $x_i'\beta_0$ is the same as targeting f_i's.). We focus on estimating the oracle model $x_i'\beta_0$ using estimators of the form $x_i'\widehat{\beta}$, and we seek to bound estimation errors with respect to the prediction norm $\|\widehat{\beta} - \beta_0\|_{2,n} := \{\mathbb{E}_n[(x'\beta_0 - x'\widehat{\beta})^2]\}^{1/2}$. The bounds on estimation error for the ultimate target f_i then follow from the triangle inequality, namely

$$\sqrt{\mathbb{E}_n[(f - x'\widehat{\beta})^2]} \leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s. \tag{2.4}$$

REMARK 2 (Bounds on the Approximation error). The approximation errors typically satisfy $c_s \leq K\sigma\sqrt{(s \vee 1)/n}$ for some fixed constant K, since the optimization problem (2.2) balances the (squared) norm of the approximation error (the norm of the bias) and the variance, see [50, 5, 6] . In particular, this condition holds for wide classes of functions, see Example S of Section 4 dealing with Sobolev classes and SM's Section C.

2.1. *Heteroscedastic* $\sqrt{Lasso}$*.* In this section we formally define the estimators which are tailored to deal with heteroscedasticity.

We propose to define the $\sqrt{\text{Lasso}}$ estimator as

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} + \frac{\lambda}{n}\|\Gamma\beta\|_1, \tag{2.5}$$

where $\widehat{Q}(\beta) = \mathbb{E}_n[(y - x'\beta)^2]$, $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_p)$ is a diagonal matrix of penalty loadings. The scaled ℓ_1-penalty allows component specific adjustments to more efficiently deal with heteroscedasticity.[3] Throughout we assume $\gamma_j \geq 1$ for $j = 1, \ldots, p$.

In order to reduce the shrinkage bias of $\sqrt{\text{Lasso}}$, we consider the post model selection estimator that applies ordinary least squares (ols) to a model $\widehat{T}$ that contains the model selected by $\sqrt{\text{Lasso}}$. Formally, let $\widehat{T}$ be such that

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

[3]In the traditional case of homoscedastic errors every penalty loading can be taken equal to 1. In the heteroscedastic case, if λ and Γ are appropriate choices, then $\lambda\|\Gamma\|_\infty$ and I_p are also an appropriate choice but potentially conservative, i.e. leading to over penalization and worse finite sample performance.

and define the ols post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$ associated with $\widehat{T}$

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} \quad : \quad \beta_j = 0 \ \text{ if } \ j \notin \widehat{T}. \tag{2.6}$$

A sensible choice for $\widehat{T}$ is simply to set $\widehat{T} = \text{supp}(\widehat{\beta})$. We allow for additional components (potentially selected through an arbitrary data-dependent procedure) to be added, which is relevant for practice.

2.2. *Typical Conditions on the Gram Matrix.* The Gram matrix $\mathbb{E}_n[xx']$ plays an important role in the analysis of estimators in this setup. When $p > n$, the smallest eigenvalue of the Gram matrix is 0, which creates identification problems. Thus, to restore identification, one needs to restrict the type of deviation vectors δ corresponding to the potential deviations of the estimator from the target value β_0. Because of the ℓ_1-norm regularization, the following restricted set is important:

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

The restricted eigenvalue $\kappa_{\bar{c}}$ of the Gram matrix $\mathbb{E}_n[xx']$ is defined as

$$\kappa_{\bar{c}} := \min_{\delta \in \Delta_{\bar{c}}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma \delta_T\|_1}. \tag{2.7}$$

The restricted eigenvalues can depend on n, T, and Γ, but we suppress the dependence in our notations. The restricted eigenvalues (2.7) are variants of the restricted eigenvalue introduced in [14] and of compatibility condition in [54] that accommodate the penalty loadings Γ. They proved to be suitable for many design of interest specially to establish ℓ_k-norm rates. Below we discuss new variants of restricted eigenvalues and compatibility conditions in [52] and [54] that are tailored for deriving prediction error rates.

The minimal and maximal m-sparse eigenvalues of a matrix M,

$$\phi_{\min}(m, M) := \min_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}, \quad \phi_{\max}(m, M) := \max_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}. \tag{2.8}$$

Typically we consider $M = \mathbb{E}_n[xx']$ or $M = \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}$. When M is not specified we mean $M = \mathbb{E}_n[xx']$, i.e. $\phi_{\min}(m) = \phi_{\min}(m, \mathbb{E}_n[xx'])$ and $\phi_{\max}(m) = \phi_{\max}(m, \mathbb{E}_n[xx'])$. 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, see [14].

3. 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 impact factors that are invariant to the introduction of repeated regressors and well-behaved if the restricted eigenvalue (2.7) is well-behaved.

The following event plays a central role in the analysis

$$\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty, \quad \text{where} \quad \widetilde{S} := \mathbb{E}_n[x(\sigma\epsilon + r)]/\sqrt{\mathbb{E}_n[(\sigma\epsilon + r)^2]} \tag{3.1}$$

is the score of $\widehat{Q}^{1/2}$ at β_0. Throughout the section we assume such event holds. Later we provide choices of λ and Γ based on primitive conditions such that the event in (3.1) holds with high probability (see Lemma 7 for a detailed finite sample analysis).

3.1. *Noise and Design Impact Factors.* In this section we propose new impact factors that are tailored to establish prediction error rates which will allow for more general designs than previous conditions proposed in the literature [15]. We define the following *noise* and *design* impact factors:

$$\varrho_{\bar{c}} := \sup_{\substack{\delta \in \Delta_{\bar{c}},\ \|\delta\|_{2,n} > 0 \\ \|\Gamma(\delta+\beta_0)\|_1 \leq \bar{c}\|\Gamma\beta_0\|_1}} \frac{|\widetilde{S}'\delta|}{\|\delta\|_{2,n}}, \tag{3.2}$$

$$\bar{\kappa} := \inf_{\|\Gamma\delta_{T^c}\|_1 < \|\Gamma\delta_T\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1}. \tag{3.3}$$

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

An analysis based on the quantities $\varrho_{\bar{c}}$ and $\bar{\kappa}$ will be more general than the one relying only on restricted eigenvalue condition (2.7). This follows because (2.7) yields one possible way to bound both $\bar{\kappa}$ and $\varrho_{\bar{c}}$, namely,

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

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

Moreover, we stress that the quantities $\bar{\kappa}$ and $\varrho_{\bar{c}}$ can be well-behaved even in the presence of repeated regressors while restricted eigenvalues and compatibility constants proposed in the literature would be trivially zero in that case.

The design impact factor $\bar{\kappa}$ in (3.3) strictly generalizes the original restricted eigenvalue (2.7) conditions proposed in [14] and the compatibility condition defined in [54]. It also generalizes the compatibility condition in

[52].[4] Thus (3.3) is an interesting condition since it was shown in [14] and [54] that the restricted eigenvalue and the compatibility assumptions are relatively weak conditions.

The noise impact factor $\varrho_{\bar{c}}$ also plays a critical role in our analysis. It depends not only on the design but also on the error and approximation terms, and can be controlled via empirical process techniques. Finally, the deviation from β_0 of the $\sqrt{\text{Lasso}}$ estimator, $\delta = \widehat{\beta} - \beta_0$, satisfies the two constraints in the definition of $\varrho_{\bar{c}}$ provided the penalty level λ is set appropriately. The lemmas below summarize the above discussion.

LEMMA 1 (Bounds on and Invariance of Design Impact Factor). *Under Condition ASM we have $\bar{\kappa} \geq \kappa_1 \geq \kappa_{\bar{c}}$. 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, $\bar{\kappa}$ does not change.*

LEMMA 2 (Bounds on and Invariance of Noise Impact Factor). *Under Condition ASM we have $\varrho_{\bar{c}} \leq (1+\bar{c})\sqrt{s}\|\Gamma^{-1}\widetilde{S}\|_\infty/\kappa_{\bar{c}}$. Moreover, if copies of regressors are included with the same corresponding penalty loadings, $\varrho_{\bar{c}}$ does not change.*

LEMMA 3 (Estimators belong to Restricted Sets). *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*

$$\|\Gamma\widehat{\beta}_{T^c}\|_1 \leq \bar{c}\|\Gamma(\widehat{\beta}_T - \beta_0)\|_1 \quad \text{and} \quad \|\Gamma\widehat{\beta}\|_1 \leq \bar{c}\|\Gamma\beta_0\|_1. \tag{3.4}$$

3.2. *Finite-sample bounds on $\sqrt{Lasso}$.* In this section we derive finite-sample bounds for the prediction norm of the $\sqrt{\text{Lasso}}$ estimator. These bounds are established under heteroscedasticity, without knowledge of the scaling parameter σ, and using the impact factors proposed in Section 3.1. For $c > 1$, let $\bar{c} = (c+1)/(c-1)$ and consider the event

$$\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty \quad \text{and} \quad \bar{\zeta} := \lambda\sqrt{s}/(n\bar{\kappa}) < 1. \tag{3.5}$$

THEOREM 1 (Finite Sample Bounds on Estimation Error). *Under Condition ASM and (3.5) we have*

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

[4]The compatibility condition defined in [52] would be stated in the current notation as $\exists\nu(T) > 0$ such that

$$\inf_{\|\Gamma\delta_{T^c}\|_1 < 3\|\Gamma\delta_T\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{(1+\nu(T))\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} > 0.$$

By using $\nu(T) = 0$ and Δ_1 which weakens the conditions $\nu(T) > 0$ and Δ_3 required in [52]. Allowing for $\nu(T) = 0$ is necessary to cover designs with repeated regressors.

We recall that the choice of λ does not depend on the scaling parameter σ. The impact of σ in the bound of Theorem 1 comes through the factor $\widehat{Q}^{1/2}(\beta_0) \leq \sigma\sqrt{\mathbb{E}_n[\epsilon^2]} + c_s$ where c_s is the size of the approximation error defined in Condition ASM. Moreover, under the typical condition that imply $\kappa_{\bar{c}}$ is bounded away from zero, for example under Condition P of Section 4 and standard choice of penalty, we have with a high probability

$$\frac{\varrho_{\bar{c}} + \bar{\zeta}}{1 - \bar{\zeta}^2} \lesssim \sqrt{\frac{s \log(p \vee n)}{n}} \implies \|\widehat{\beta} - \beta_0\|_{2,n} \lesssim \sqrt{\frac{s \log(p \vee n)}{n}}.$$

Thus, Theorem 1 generally leads to the same rate of convergence as in the case of the Lasso estimator that knows σ since $\mathbb{E}_n[\epsilon^2]$ concentrates around one under (2.1) and a law of large numbers holds. We derive performance bounds for other norms of interest in the Supplementary Material (SM).

The next result deals with $\widehat{Q}(\widehat{\beta})$ as an estimator for $\widehat{Q}(\beta_0)$ and σ^2.

THEOREM 2 (Estimation of σ). *Under Condition ASM and (3.5)*

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

Under only Condition ASM we have

$$\left|\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma\right| \leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s + \sigma|\mathbb{E}_n[\epsilon^2] - 1|.$$

We note that further bounds on $|\mathbb{E}_n[\epsilon^2] - 1|$ are implied by Vonbahr-Esssen's and Markov's inequalities, or by self-normalized moderate deviation (SNMD) theory as in Lemma 4. As a result, the theorem implies consistency $|\widehat{Q}^{1/2}(\widehat{\beta}) - \sigma| = o_P(1)$ under mild moment conditions; see Section 4. The stated bounds on $\widehat{Q}^{1/2}(\widehat{\beta})$ are also useful for establishing sparsity properties, which is what we deal with in the next result.

THEOREM 3 (Sparsity bound for $\sqrt{\text{Lasso}}$). *Under Condition ASM and (3.5), let* $\widehat{m} := |\text{supp}(\widehat{\beta}) \setminus T|$ *and* $\widehat{Q}(\beta_0) > 0$. *If* $\frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\zeta})}{1 - \bar{\zeta}^2} \leq 1/\bar{c}$, *we have*

$$|\text{supp}(\widehat{\beta})| \leq s \cdot 4\bar{c}^2 \left(\frac{1 + \varrho_{\bar{c}}/\bar{\zeta}}{\bar{\kappa}(1 - \bar{\zeta}^2)}\right)^2 \min_{m \in \mathcal{M}} \phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > s\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}) \cdot 8\bar{c}^2 \left(\frac{1 + \varrho_{\bar{c}}/\bar{\zeta}}{\bar{\kappa}(1 - \bar{\zeta}^2)}\right)^2\}$. *Moreover, if* $\kappa_{\bar{c}} > 0$ *and* $\bar{\zeta} < 1/\sqrt{2}$ *we have*

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

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

REMARK 3 (On the Sparsity Bound). Section 4 will show that under minimal and maximal sparse eigenvalues of order $s \log n$ bounded away from zero and from above, the bound derived above implies that with a high probability

$$|\text{supp}(\widehat{\beta})| \lesssim s := |\text{supp}(\beta_0)|,$$

that is the selected model's size will be of the same order as the size of the oracle model. We note however that the former condition is merely a sufficient condition. The bound $|\text{supp}(\widehat{\beta})| \lesssim s$ will apply for other designs of interest. This can be the case even if $\kappa_{\bar{c}} = 0$. For instance, this would be the case for the aforemention design, if we change it by adding a single repeated regressor. ∎

REMARK 4 (Maximum Sparse Eigenvalue and Sparsity). Consider the case of $f(z) = z$ with p repeated regressors $x_i = (z_i, \ldots, z_i)'$ where $|z| \leq B$. In this case one could set $\Gamma = I \cdot B$. In this setting, there is a sparse solution for $\sqrt{\text{Lasso}}$, but there is also a solution which has p nonzero regressors. Nonetheless, the bound for the prediction error rate will be well-behaved since $\bar{\kappa}$ and $\bar{\zeta}$ are invariant under repeated regressors and satisfy:

$$\bar{\kappa} \geq 1/B \text{ and } \varrho_{\bar{c}} = |\mathbb{E}_n[\epsilon z]|/\{\mathbb{E}_n[\epsilon^2]\mathbb{E}_n[z^2]\}^{1/2} \lesssim_P 1/\sqrt{n}.$$

Thus, the sparsity bound above will become trivial because of the maximum sparse eigenvalue. Indeed, in this case $\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}) = (m+1)\mathbb{E}_n[z^2]/B^2$ and the set $\mathcal{M}$ becomes empty leading to the trivial bound $\widehat{m} \leq p$. Note that in this case it is possible to use the goodness-of-fit based thresholding procedure of [6] to get the model of the right sparsity. ∎

3.3. *Finite-sample bounds on ols post $\sqrt{Lasso}$.* Next we consider the ols estimator applied to the models $\widehat{T}$ that was selected by $\sqrt{\text{Lasso}}$ or includes such model (plus other components that the data analyst may wish to include), namely $\text{supp}(\widehat{\beta}) \subseteq \widehat{T}$. We are interested in the case when model selection does not work perfectly, as occurs in applications.

The following result establishes performance bounds for the ols post $\sqrt{\text{Lasso}}$ estimator. Following [6], the analysis accounts for the data-driven choice of components and for the possibly misspecified selected model (i.e. $T \not\subseteq \widehat{T}$).

THEOREM 4 (Performance of ols post $\sqrt{\text{Lasso}}$). *Under Condition ASM and (3.5), let* $\text{supp}(\widehat{\beta}) \subseteq \widehat{T}$, *and* $\widehat{m} = |\widehat{T} \setminus T|$. *Then we have that the ols post $\sqrt{Lasso}$ estimator based on $\widehat{T}$ satisfies*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq \frac{\sigma\sqrt{s + \widehat{m}}\|\mathbb{E}_n[x\epsilon]\|_\infty}{\sqrt{\phi_{\min}(\widehat{m})}} + 2c_s + 2\sqrt{\widehat{Q}(\beta_0)}\frac{(\varrho_{\bar{c}} + \bar{\zeta})}{1 - \bar{\zeta}^2}.$$

The result is derived from the sparsity of the model $\widehat{T}$ and from its approximating ability. Note the presence of the new term– $\|\mathbb{E}_n[x\epsilon]\|_\infty$. Bounds on $\|\mathbb{E}_n[x\epsilon]\|_\infty$ can be derived using the same tools used to justify the penalty level λ, via self-normalized moderate deviation theory [32] or using empirical process inequalities as derived in [4]. Under mild conditions we have $\|\mathbb{E}_n[x\epsilon]\|_\infty \leq C\sqrt{\log(pn)/n}$ with probability $1 - o(1)$.

3.4. *Two extreme cases. Case (i): Parametric noiseless case.* Consider the case that $\sigma = 0$ and $c_s = 0$. Therefore the regression function is exactly sparse, $f(z_i) = x_i'\beta_0$. In this case $\sqrt{\text{Lasso}}$ can exactly recover the regression function under weak conditions.

THEOREM 5 (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 $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$. Then we have $\|\widehat{\beta} - \beta_0\|_{2,n} = 0$. Moreover, if $\kappa_1 > 0$, we have $\widehat{\beta} = \beta_0$.*

Thus, a sufficient condition for exact recovery of the regression function in the noiseless case depends on the design impact factor $\bar{\kappa}$. If, further, the restricted eigenvalue κ_1 is bounded away from zero, the $\sqrt{\text{Lasso}}$ estimator will perfectly recover β_0 under a wide range of penalty levels.

REMARK 5 (Perfect Recovery and Lasso). 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, $\sqrt{\text{Lasso}}$ can automatically adapt to the noiseless case. ∎

Case (ii): Nonparametric infinite variance. We conclude this section with the infinite variance case. The finite sample theory does not rely on $\bar{\mathbf{E}}[\epsilon^2] = 1$. Instead it relies on the choice of penalty level and penalty loadings to satisfy $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$. Under symmetric errors we exploit the self-normalized theory to develop a choice of penalty level and loadings,

$$\lambda = (1 + u_n)c\sqrt{n}(1 + \sqrt{2\log(2p/\alpha)}) \quad \text{and} \quad \gamma_j = \max_{1\leq i\leq n} |x_{ij}|. \tag{3.6}$$

The sequence u_n is defined below and typically we can set $u_n = o(1)$.

THEOREM 6 ($\sqrt{\text{Lasso}}$ prediction norm for symmetric errors). *Consider a nonparametric regression model with data $(y_i, z_i)_{i=1}^n$, $y_i = f(z_i) + \epsilon_i$, $x_i = P(z_i)$ such that $\mathbb{E}_n[x_j^2] = 1$ $(j = 1, \ldots, p)$, 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 (3.6) where u_n is such that $P(\mathbb{E}_n[\sigma\epsilon^2] > (1+u_n)\mathbb{E}_n[(\sigma\epsilon + r)^2]) \leq \eta_1$. Moreover let $P(\mathbb{E}_n[\epsilon^2] \leq \{1+u_n\}^{-1}) \leq \eta_2$. If $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, then with probability at least $1-\alpha-\eta_1-\eta_2$ we have $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ and

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

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

COROLLARY 1 ($\sqrt{\text{Lasso}}$ prediction norm for $\epsilon_i \sim t(2)$). *Under the setting of Theorem 6, suppose that $\epsilon_i \sim t(2)$ are i.i.d. noise. 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 $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ and*

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

Asymptotically, provided that regressors are uniformly bounded and satisfy the sparse eigenvalues condition (4.3), we have that the restricted eigenvalue $\kappa_{\bar{c}}$ is bounded away from zero for the choice of Γ. Because Corollary 1 ensures $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ with the stated probability, by Lemmas 1 and 2 we have $\varrho_{\bar{c}} + \bar{\zeta} \lesssim \lambda\sqrt{s}/\{n\kappa_{\bar{c}}\} \lesssim \sqrt{s\log(p \vee n)/n}$. Therefore, under these design conditions, if $\tau = 1/\log n$, $1/\alpha = o(\log n)$ and $s\log(p/\alpha) = o(n)$, Corollary 1 yields that the $\sqrt{\text{Lasso}}$ estimator satisfies

$$\|\widehat{\beta} - \beta_0\|_{2,n} \lesssim (c_s + \sigma\sqrt{\log n})\sqrt{\frac{s\log(p \vee n)}{n}}, \tag{3.7}$$

with probability $1 - \alpha(1 + o(1))$, where the scale $\sigma < \infty$ is fixed. Despite the infinite variance of the noise in the $t(2)$ case, the bound (3.7) differs from the Gaussian noise case by a $\sqrt{\log n}$ factor.

4. Asymptotics Analysis under Primitive Conditions. In this section we formally state an algorithm to compute the estimators and we provide rates of convergence results under simple primitive conditions.

We propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}(1 - \alpha/2p), \tag{4.1}$$

where α controls the confidence level, and $c > 1$ is a slack constant similar to [14], and the penalty loadings according to the following iterative algorithm.

ALGORITHM 1 (Estimation of Square-root Lasso Loadings). Choose $\alpha \in (1/n, 1)$, and a constant $K \geq 1$ as an upper bound on the number of iterations. (0) Set $k = 0$, λ as defined in (4.1), and set $\widehat{\gamma}_{j,0} = \max_{1 \leq i \leq n} |x_{ij}|$ for each $j = 1, \ldots, p$. (1) Compute the $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$ based on the current penalty loadings $\{\widehat{\gamma}_{j,k}, j = 1, \ldots, p\}$. (2) Set

$$\widehat{\gamma}_{j,k+1} := 1 \vee \sqrt{\mathbb{E}_n[x_j^2(y - x'\widehat{\beta})^2]} / \sqrt{\mathbb{E}_n[(y - x'\widehat{\beta})^2]}.$$

(3) If $k > K$, stop; otherwise set $k \leftarrow k + 1$ and go to step 1.

REMARK 6 (Parameters of the Algorithm). 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 as in [14]; we recommend $c = 1.01$. To invoke self-normalized moderate deviations, we just need to be able to bound with a high probability:

$$\sqrt{\mathbb{E}_n[x_j^2 \epsilon^2]} / \sqrt{\mathbb{E}_n[\epsilon^2]} \leq \gamma_{j,0}. \tag{4.2}$$

The choice of $\widehat{\gamma}_{j,0} = \max_{1 \leq i \leq n} |x_{ij}|$ automatically achieves (4.2). Nonetheless, we recommend iterating the procedure to avoid unnecessary overpenalization since at each iteration more precise estimates of the penalty loadings are 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 (see SM's Section G). ∎

REMARK 7 (Alternative Estimation of Loadings). Algorithm 1 relies on the $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$. Another possibility is to use the post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$. This leads to similar theoretical and practical results. Moreover, we can define the initial penalty loading as $\widehat{\gamma}_{j,0} = W\{\mathbb{E}_n[x_j^4]\}^{1/4}$ where the kurtosis parameter $W > \{\bar{\mathbf{E}}[\epsilon^4]\}^{1/4} / \{\bar{\mathbf{E}}[\epsilon^2]\}^{1/2}$ is pivotal with respect to the scaling parameter σ, but we need to assume an upper bound for this quantity. The purpose of this parameter is to bound the kurtosis of the marginal distribution of errors, namely that of $\bar{F}_\epsilon(v) = n^{-1} \sum_{i=1}^n P(\epsilon_i \leq v)$. We recommend $W = 2$, which permits a wide class of marginal distributions of errors, in particular it allows $\bar{F}_\epsilon$ to have tails as heavy as v^{-a} with $a > 5$. Either option is a reasonable way of achieving (4.2) and we analyze both options in the SM's Section C.1 under weaker conditions than Condition P below. ∎

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

CONDITION P. *The noise and the covariates obey* $\sup_{n\geq 1}\bar{\mathbf{E}}[|\epsilon|^q] < \infty$, $q > 4$, $\inf_{n\geq 1}\min_{1\leq j\leq p}\mathbb{E}_n[x_j^2\mathrm{E}[\epsilon^2]] > 0$, $\sup_{n\geq 1}\max_{1\leq j\leq p}\mathbb{E}_n[x_j^3\mathrm{E}[\epsilon^3]] < \infty$ *and*

$$\sup_{n\geq 1}\phi_{\max}(s\log n, \mathbb{E}_n[x_i x_i'])/\phi_{\min}(s\log n, \mathbb{E}_n[x_i x_i']) < \infty. \tag{4.3}$$

Moreover, we have that $\log p \leq C(n/\log n)^{1/3}$, $\max_{i\leq n}\|x_i\|_\infty^4 s\log(p\vee n) \leq Cn/\log n$, $s \geq 1$, *and* $c_s^2 \leq C\sigma^2(s\log(p\vee n)/n)$.

Condition P collects moment conditions that allow us to use results of the self-normalized moderate deviation theory, weak requirements on (s, p, n), well behaved sparse eigenvalue as a sufficient condition on the design to bound the impact factors, and a mild condition on the approximation errors (see Remark 2 for a discussion and references).

The proofs in this section rely on the following result due to [32].

LEMMA 4 (Moderate deviations for self-normalized sums). *Let* $X_1, \ldots, X_n$ *be independent, zero-mean random variables and* $\delta \in (0, 1]$. *Let* $S_{n,n} = n\mathbb{E}_n[X]$, $V_{n,n}^2 = n\mathbb{E}_n[X^2]$ *and* $M_n = \{\bar{\mathbf{E}}[X^2]\}^{1/2}/\{\bar{\mathbf{E}}|X|^{2+\delta}]\}^{1/\{2+\delta\}} > 0$. *Suppose that for some* $\ell_n \to \infty$ *such that* $n^{\frac{\delta}{2(2+\delta)}}M_n/\ell_n \geq 1$. *Then for some absolute constant* A, *uniformly on* $0 \leq x \leq n^{\frac{\delta}{2(2+\delta)}}M_n/\ell_n - 1$, *we have*

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

The following theorem summarizes the asymptotic performance of $\sqrt{\text{Lasso}}$, based upon Algorithm 1, for commonly used designs.

THEOREM 7 (Performance of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ under Condition P). *Suppose Conditions ASM and P hold. Let* $\alpha \in (1/n, 1/\log n)$, $c > 1.01$, *the penalty level* λ *be set as in (4.1) and the penalty loadings as in Algorithm 1. Then for all* $n \geq n_0$, *with probability at least* $1 - \alpha\{1 + \bar{C}/\log n\} - \bar{C}\{n^{-1/2}\log n\}$ *we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \sigma\bar{C}\sqrt{\tfrac{s\log(n\vee(p/\alpha))}{n}}, \quad \sqrt{\mathbb{E}_n[(f - x'\widehat{\beta})^2]} \leq \sigma\bar{C}\sqrt{\tfrac{s\log(n\vee(p/\alpha))}{n}}$$
$$\|\widehat{\beta} - \beta_0\|_1 \leq \sigma\bar{C}\sqrt{\tfrac{s^2\log(n\vee(p/\alpha))}{n}} \quad \text{and} \quad |\mathrm{supp}(\widehat{\beta})| \leq \bar{C}s,$$

where n_0 *and* $\bar{C}$ *depend only on the constants in Condition P. Moreover, the ols post* $\sqrt{\text{Lasso}}$ *estimator satisfies with the same probability for all* $n \geq n_0$,

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq \sigma\bar{C}\sqrt{\tfrac{s\log(n\vee(p/\alpha))}{n}}, \quad \sqrt{\mathbb{E}_n[(f - x'\widetilde{\beta})^2]} \leq \sigma\bar{C}\sqrt{\tfrac{s\log(n\vee(p/\alpha))}{n}}$$
$$\text{and } \|\widehat{\beta} - \beta_0\|_1 \leq \sigma\bar{C}\sqrt{\tfrac{s^2\log(n\vee(p/\alpha))}{n}}$$

Theorem 7 yields bounds on the estimation errors that are "Gaussian-like," namely the factor $\sqrt{\log p/\alpha}$ and other constants in the performance bound are the same as if errors were Gaussian, but the probabilistic guarantee is not $1-\alpha$ but rather $1-\alpha(1+o(1))$, which together with mildly more restrictive growth conditions is the cost of non-Gaussianity here. Lemma 7 in the SM Section C.1 derives a more precise calculation of the probability of success under weaker conditions.

The results above establish that $\sqrt{\text{Lasso}}$ achieves the same near oracle rate of convergence of Lasso despite not knowing the scaling parameter σ. It allows for heteroscedastic errors with mild restrictions on its moments. Moreover, it allows for any number of iterations K in Algorithm 1. The result also establishes that the upper bounds on the rates of convergence of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ coincide. This is confirmed also by Monte-Carlo experiments reported in the SM, with ols post $\sqrt{\text{Lasso}}$ performing no worse and often outperforming $\sqrt{\text{Lasso}}$ due to having a much smaller bias. Notably this theoretical and practical performance occurs despite the fact that $\sqrt{\text{Lasso}}$ may in general fail to correctly select the oracle model T as a subset and potentially select variables not in T.

EXAMPLE S. (Performance for Sobolev Balls and p-Rearranged Sobolev Balls.) In this example we show how our results apply to an important class of Sobolev functions, and illustrates how modern selection drastically reduces the dependency on knowing the order of importance of the basis functions.

Following [50], for an orthonormal bounded basis $\{P_j(\cdot)\}_{j=1}^{\infty}$ in $L^2[0,1]$, consider functions $f(z)=\sum_{j=1}^{\infty}\theta_j P_j(z)$ in a Sobolev space $\mathcal{S}(\alpha,L)$ for some $\alpha\geq 1$ and $L>0$. This space consist of functions whose Fourier coefficients θ satisfy $\sum_{j=1}^{\infty}|\theta_j|<\infty$ and

$$\theta\in\Theta(\alpha,L)=\left\{\theta\in\ell^2(\mathbf{N}):\ \textstyle\sum_{j=1}^{\infty}j^{2\alpha}\theta_j^2\leq L^2\ \right\}.$$

We also consider functions in a p-Rearranged Sobolev space $\mathcal{RS}(\alpha,p,L)$. These functions take the form $f(z)=\sum_{j=1}^{\infty}\theta_j P_j(z)$ such that $\sum_{j=1}^{\infty}|\theta_j|<\infty$ and $\theta\in\Theta^R(\alpha,p,L)$, where

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

Note that $\mathcal{S}(\alpha,L)\subset\mathcal{RS}(\alpha,p,L)$.

In the SM, we show that the rate-optimal choice for the size of the support of the oracle model β_0 is $s\lesssim n^{1/[2\alpha+1]}$, with the $\text{supp}(\beta_0)$ consisting of indices j that correspond to the s largest coefficients $|\theta_j|$. The oracle projection estimator $\widehat{\beta}^{\text{or}}$ that uses these "ideal" s components with the largest

coefficients achieves optimal prediction error rate under a (sequence of) regression functions $f \in \mathcal{S}(\alpha, L)$ or $f \in \mathcal{RS}(\alpha, p, L)$:

$$\sqrt{\mathbb{E}_n[\{f - \sum_{j=1}^{\infty} \widehat{\beta}_j^{\text{or}} P_j(z)\}^2]} \lesssim_P n^{-\alpha/[2\alpha+1]}.$$

Under mild regularity conditions, as in Theorem 7, $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$ that uses $x_i = (P_1(z_i), ..., P_p(z))'$ achieves

$$\sqrt{\mathbb{E}_n[(f - x'\widehat{\beta})^2]} \lesssim_P n^{-\alpha/[2\alpha+1]} \sqrt{\log(n \vee p)},$$

without knowing the "ideal" s components amongst x_i. The same statement also holds for the post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$.

Therefore the $\sqrt{\text{Lasso}}$ or post $\sqrt{\text{Lasso}}$ estimators achieves near oracle rates uniformly over $f \in \mathcal{S}(\alpha, L)$, provided conditions of preceding theorem hold for any sequence of regression functions $f \in \mathcal{S}(\alpha, L)$ and the corresponding sequence of p. In the case function in a p-Rearranged Sobolev ball, the adaptivity of the $\sqrt{\text{Lasso}}$ estimator allows it to achieve the same uniform rates over $\mathcal{RS}(\alpha, p, L)$. Finally, consider the "naive oracle" series projection estimator that consider the first s components of the basis, assuming that the parameter space is $\mathcal{S}(\alpha, L)$. This estimator achieves the optimal rate for the Sobolev space $\mathcal{S}(\alpha, L)$, but fails to be uniformly consistent over p-Rearranged Sobolev space $\mathcal{RS}(\alpha, p, L)$, since we can select a model $f \in \mathcal{RS}(\alpha, p, L)$ such that its first s Fourier coefficients are zero, and the remaining coefficients are non-zero, therefore the "naive oracle" fit will be 0 plus some centered noise, and the estimator will be inconsistent for this f. ∎

We proceed to state a result on estimation of σ^2 under the asymptotic framework.

COROLLARY 2 (Estimation of σ^2 under Asymptotics). *Suppose Conditions ASM and P hold. Let $\alpha \in (1/n, 1/\log n)$, $c > 1.01$, the penalty level λ be set as in (4.1) and the penalty loadings as in Algorithm 1. Then for all $n \geq n_0$, with probability at least $1 - \alpha\{1 + \bar{C}/\log n\} - \bar{C}\{n^{-1/2}\log n\} - 2\delta$ we have*

$$\left|\widehat{Q}(\widehat{\beta}) - \sigma^2\right| \leq \frac{\sigma^2 \bar{C} s \log(n \vee (p/\alpha))}{n} + \frac{\sigma^2 \bar{C}\sqrt{s \log(p \vee n)}}{\sqrt{\delta} n^{1-1/q}} + \frac{\sigma^2 \bar{C}}{\sqrt{\delta n}}.$$

Moreover, provided further that $s^2 \log^2(p \vee n) \leq Cn/\log n$, we have that $\{\sigma^2 \xi_n\}^{-1} n^{1/2} \left(\widehat{Q}(\widehat{\beta}) - \sigma^2\right) \Rightarrow N(0, 1)$ where $\xi_n^2 = \bar{\mathbf{E}}[\{\epsilon^2 - \mathrm{E}[\epsilon^2]\}^2]$.

This result extends [6, 49] to the heteroscedastic, non-Gaussian cases.

5. An application to a generic semi-parametric problem. In this section we present a generic application of the methods of this paper to semi-parametric problems, where some lower-dimensional structural parameter is of interest and the $\sqrt{\text{Lasso}}$ or ols post $\sqrt{\text{Lasso}}$ are used to estimate the high-dimensional nuisance function. We denote the true value of the target parameter by $\theta_0 \in \Theta \subset \mathbb{R}^d$, and assume that it satisfies the following moment condition:

$$\mathrm{E}[\psi(w_i, \theta_0, h_0(z_i))] = 0, \quad i = 1, ..., n, \tag{5.1}$$

where w_i is a random vector taking values in $\mathcal{W}$, containing vector z_i taking values in $\mathcal{Z}$ as a subcomponent; the function $(w, \theta, t) \mapsto \psi(w, \theta, t) = (\psi_j(w, \theta, t))_{j=1}^d$ is a measurable map from an open neighborhood of $\mathcal{W} \times \Theta \times T$, a subset of Euclidian space, to $\mathbb{R}^d$, and $z \mapsto h_0(z) = (h_{m0}(z))_{m=1}^M$ is a vector of measurable nuisance functions mapping $\mathcal{Z}$ to $T \subset \mathbb{R}^M$. We note that M and d are fixed and do not depend on n in what follows.

Perhaps the simplest, that is linear, example of this kind arises in the instrumental variable (IV) regression problem in [3, 8], where $\psi(w_i, \theta_0, h_0(z_i)) = (u_i - \theta_0 d_i)h_0(z_i)$, where u_i is the response variable, d_i is the endogenous variable, z_i is the instrumental variable, $h_0(z_i) = \mathrm{E}[d_i \mid z_i]$ is the optimal instrument, and $\mathrm{E}[(u_i - \theta_0 d_i) \mid z_i] = 0$. Other examples include partially linear models, heterogeneous treatment effect models, nonlinear instrumental variable, Z-estimation problems as well as many others (see, e.g., [1, 31, 30, 21, 3, 10, 62, 9, 28, 55, 11, 44, 13, 26, 7]), which all give rise to bi-linear and nonlinear moment conditions with respect to the nuisance functions.

We assume that the nuisance functions h_0 arise as conditional expectations of some variables that can be modelled and estimated in the approximately sparse framework, as formally described below. For instance, in the example mentioned above, the function h_0 is indeed a conditional expectation of the endogenous variable given the instrumental variable. We let $\widehat{h} = (\widehat{h}_m)_{m=1}^M$ denote the estimator of h_0, which obeys conditions stated below. The estimator $\widehat{\theta}$ of θ_0 is constructed as any approximate ϵ_n-solution in Θ to a sample analog of the estimating equation above:

$$\|\mathbb{E}_n[\psi(w, \widehat{\theta}, \widehat{h}(z))]\| \leq \epsilon_n, \text{ where } \epsilon_n = o(n^{-1/2}). \tag{5.2}$$

The key condition needed for regular estimation of θ_0 is the orthogonality condition:

$$\mathrm{E}[\partial_t \psi(w_i, \theta_0, h_0(z_i))|z_i] = 0, \quad i = 1, ..., n, \tag{5.3}$$

where here and below we use the symbol ∂_t to abbreviate $\frac{\partial}{\partial t}$. For instance in the IV example this condition holds, since $\partial_t \psi(w_i, \theta_0, h_0(z_i)) = (u_i - \theta_0 d_i)$

and $\mathrm{E}[(u_i - \theta_0 d_i)|z_i] = 0$ by assumption. In other examples, it is important to construct the scores that have this orthogonality property. Generally, if we have a score, which identifies the target parameter but does not have the orthogonality property, we can construct the score that has the required property by projecting the original score onto the orthocomplement of the tangent space for the nuisance parameter; see, e.g., [57, 56, 35] for detailed discussion.

The orthogonality condition reduces sensitivity to "crude" estimation of the nuisance function h_0. Indeed, under appropriate sparsity assumptions stated below, the estimation errors for h_0, arising as sampling, approximation, and model selection errors, will be of order $o_P(n^{-1/4})$. The orthogonality condition together with other conditions will guarantee that these estimation errors do not impact the first order asymptotic behavior of the estimating equations, so that

$$\sqrt{n}\mathbb{E}_n[\psi(w, \widehat{\theta}, \widehat{h}(z))] = \sqrt{n}\mathbb{E}_n[\psi(w, \widehat{\theta}, h_0(z))] + o_P(1). \tag{5.4}$$

This leads us to a regular estimation problem, despite $\widehat{h}$ being highly nonregular.

In what follows, we shall denote by c and C some positive constants, and by L_n a sequence of positive constants that may grow to infinity as $n \to \infty$.

CONDITION SP. *For each n, we observe the independent data vectors $(w_i)_{i=1}^n$ with law determined by the probability measure $\mathrm{P} = \mathrm{P}_n$. Uniformly for all n the following conditions hold. (i) The true parameter values θ_0 obeys (5.1) and is interior relative to Θ, namely there is a ball of fixed positive radius centered at θ_0 contained in Θ, where Θ is a fixed compact subset of $\mathbb{R}^d$. (ii) The map $\nu \mapsto \psi(w, \nu)$ is twice continuously differentiable with respect to $\nu = (\nu_k)_{k=1}^K = (\theta, t)$ for all $\nu \in \Theta \times T$, with derivatives of the second order bounded in absolute value by L_n, uniformly for all $w \in \mathcal{W}$. The conditional second moments of the first derivatives are bounded as follows: P-a.s. $\mathrm{E}\left(\sup_{\nu \in \Theta \times T} |\partial_{\nu_k} \psi_j(w_i, \nu)|^2 \mid z_i\right) \leq C$ for each k, j, and i. (iii) The orthogonality condition (5.3) holds. (iv) The following identifiability condition holds: for all $\theta \in \Theta$, $\|\bar{\mathbf{E}}[\psi(w, \theta, h_0(z))]\| \geq 2^{-1}(\|J_n(\theta - \theta_0)\| \vee c)$, where $J_n = \bar{\mathbf{E}}[\partial_\theta \psi(w, \theta_0, h_0(z))]$ has eigenvalues bounded away from zero and above. (v) $\bar{\mathbf{E}}[\|\psi(w, \theta_0, h_0(z))\|^3]$ is bounded from above.*

In addition to the previous conditions, Condition SP imposes standard identifiability and certain smoothness on the problem, requiring second derivatives to be bounded by L_n, which is allowed to grow with n subject to restrictions specified below. It is possible to allow for non-differentiable ψ

at the cost of a more complicated argument; see [11]. In what follows, let $\delta_n \searrow 0$ be a sequence of constants approaching zero from above.

CONDITION AS. *The following conditions hold for each n. (i) The function $h_0 = (h_{0m})_{m=1}^M : \mathcal{Z} \mapsto T$ is approximately sparse, namely, for each m, $h_{0m}(z) = \sum_{l=1}^p P_l(z)\beta_{0ml} + r_m(z)$, where $P_l : \mathcal{Z} \mapsto T$ are measurable functions, $\beta_{0m} = (\beta_{0ml})_{l=1}^p$ obeys $|\mathrm{supp}(\beta_{0m})| \leq s$, $s \geq 1$, and the approximation errors obey $\bar{\mathbf{E}}[r_m^2(z)] \leq Cs\log(p \vee n)/n$. There is an estimator $\widehat{h}_m(\cdot) = \sum_{l=1}^p P_l(\cdot)\widehat{\beta}_{ml}$ of h_{0m} such that $\widehat{h} = (\widehat{h}_m)_{m=1}^M$ maps $\mathcal{Z}$ into T, and with probability $1-\delta_n$, $\widehat{\beta}_m = (\widehat{\beta}_{ml})_{l=1}^p$ satisfies $\|\widehat{\beta}_m - \beta_{0m}\|_1 \leq C\sqrt{s^2\log(p \vee n)/n}$ and $\mathbb{E}_n(\widehat{h}_m(z) - h_{0m}(z))^2 \leq Cs\log(p \vee n)/n$. (ii) The scalar variables $\dot{\psi}_{mjl}(w_i) := \partial_{t_m}\psi_j(w_i, \theta_0, h_0(z))P_l(z_i)$ obey $\max_{m,j,l}\mathbb{E}_n[|\dot{\psi}_{mjl}(w)|^2] \leq L_n^2$ with probability $1-\delta_n$ and $\max_{m,j,l}(\bar{\mathbf{E}}[|\dot{\psi}_{mjl}(w)|^3])^{1/3}/(\bar{\mathbf{E}}[|\dot{\psi}_{mjl}(w)|^2])^{1/2} \leq B_n$. (iii) Finally, the following growth restrictions hold as $n \to \infty$:*

$$L_n^2 s^2 \log^2(p \vee n)/n \to 0 \text{ and } \log(p \vee n)n^{-1/3}B_n^2 \to 0. \tag{5.5}$$

The assumption records a formal sense in which approximate sparsity is used, as well as requires reasonable behavior of the estimator $\widehat{h}$. In the previous sections, we established primitive conditions under which this behavior occurs in problems where h_0 arise as conditional expectation functions. By virtue of (5.5) the assumption implies that $\{\mathbb{E}_n(\widehat{h}_m(z) - h_{0m}(z))^2\}^{1/2} = o_P(n^{-1/4})$. It is standard that the square of this term multiplied by $\sqrt{n}$ shows up as a linerization error for $\sqrt{n}(\widehat{\theta} - \theta_0)$, and therefore this term does not affect its first order behavior. Moreover, the assumption implies by virtue of (5.5) that $\|\widehat{\beta}_m - \beta_{0m}\|_1 = o_P(L_n^{-1}(\log(p \vee n))^{-1})$, which is used to control another key term in the linearization as follows:

$$\sqrt{n}\max_{j,m,l}|\mathbb{E}_n[\dot{\psi}_{mjl}(w)]|\|\widehat{\beta}_m - \beta_{0m}\|_1 \lesssim_P L_n\sqrt{\log(p \vee n)}\|\widehat{\beta}_m - \beta_{0m}\|_1 = o_P(1),$$

where the bound follows from an application of the moderate deviation inequalities for self-normalized sums. The idea for this type of control is borrowed from [3], who used in the context of the IV example mentioned above.

THEOREM 8. *Under Conditions SP and AS holding for each n, the estimator $\widehat{\theta}$ that obeys equation (5.2) and $\widehat{\theta} \in \Theta$ with probability approaching 1, satisfies $\sqrt{n}(\widehat{\theta} - \theta_0) = -J_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(w_i, \theta_0, h_0(z_i)) + o_P(1)$. Furthermore, provided $\Omega_n = \bar{\mathbf{E}}[\psi(w, \theta_0, h_0(z))\psi(w, \theta_0, h_0(z))']$ has eigenvalues bounded away from zero, we have that*

$$\Omega_n^{-1/2}J_n\sqrt{n}(\widehat{\theta} - \theta_0) \Rightarrow N(0, I).$$

This theorem extends the analogous result from [8, 3] for a specific linear to a generic non-linear setting, and could be of independent interest in many problems cited above.

APPENDIX A: PROOFS OF SECTION 3

PROOF OF LEMMA 1. 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 $\mathbb{E}_n[x_j^2] = 1$, $\|v\|_{2,n} \leq \|v\|_1$ for any $v \in \mathbb{R}^p$. For any δ such that $\|\Gamma\delta_{T^c}\|_1 < \|\Gamma\delta_T\|_1$ we have that

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

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 (that is, since T is selected by the oracle (2.2), T will not contain 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 $\tilde{\delta}^2 = 0$. ∎

PROOF OF LEMMA 2. The first part follows from Hölder's inequality and the definition of $\kappa_{\bar{c}}$. To show the second part 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 $\delta = (\delta_T^{1\prime}, \delta_{T^c}^{1\prime}, \delta_T^{2\prime}, \delta_{T^c}^{2\prime})'$. It follows that $|\widetilde{S}'\delta|/\|\delta\|_{2,n} = |\widetilde{S}'\tilde{\delta}|/\|\tilde{\delta}\|_{2,n}$ where $\tilde{\delta}_T = (\delta_T^{1\prime} + \delta_T^{2\prime}, \delta_{T^c}^{1\prime} + \delta_{T^c}^{2\prime}, 0', 0')'$, and $\tilde{\delta}_{T^c} = \delta - \tilde{\delta}_T$. This transformation yields $\|\Gamma(\tilde{\delta} + \beta_0)\|_1 \leq \|\Gamma(\delta + \beta_0)\|_1$ and that $\delta \in \Delta_{\bar{c}}$ implies $\tilde{\delta} \in \Delta_{\bar{c}}$. 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. ∎

PROOF OF LEMMA 3. See SM. ∎

PROOF OF THEOREM 1. First note that by Lemma 3 we have $\widehat{\delta} := \widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$. By optimality of $\widehat{\beta}$ and definition of $\bar{\kappa}$, $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}]$ we have

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

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

$$\|\widehat{\delta}\|_{2,n}^2 \leq 2\mathbb{E}_n[(\sigma\epsilon + r)x'\widehat{\delta}] + \left(\sqrt{\widehat{Q}(\widehat{\beta})} + \sqrt{\widehat{Q}(\beta_0)}\right)\bar{\zeta}\|\widehat{\delta}\|_{2,n}. \tag{A.2}$$

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

$$\|\widehat{\delta}\|_{2,n}^2 \quad \leq \quad 2\mathbb{E}_n[(\sigma\epsilon + r)x'\widehat{\delta}] + 2\sqrt{\widehat{Q}(\beta_0)}\bar{\zeta}\|\widehat{\delta}\|_{2,n} + \bar{\zeta}^2\|\widehat{\delta}\|_{2,n}^2.$$

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

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

and the result follows provided $\bar{\zeta} < 1$. ∎

PROOF OF THEOREM 2. 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 3.

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 $\bar{\zeta} := \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we obtain

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

Moreover, we have

$$\left|\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma\right| \quad \leq \quad \left|\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma\{\mathbb{E}_n[\epsilon^2]\}^{1/2}\right| + \sigma\left|\{\mathbb{E}_n[\epsilon^2]\}^{1/2} - 1\right|$$

and the right side is bounded by $\|\widehat{\beta} - \beta_0\|_{2,n} + c_s + \sigma|\mathbb{E}_n[\epsilon^2] - 1|$. ∎

PROOF OF THEOREM 3. For notational convenience we denote $\phi_n(m) = \phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})$. We shall rely on the following lemma, whose proof is given after the proof of this theorem:

LEMMA 5 (Relating Sparsity and Prediction Norm). *Under Condition ASM, let $G \subseteq \operatorname{supp}(\widehat{\beta})$. For any $\lambda > 0$ we have*

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

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

$$\left(\sqrt{\frac{\widehat{Q}(\widehat{\beta})}{\widehat{Q}(\beta_0)}} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\text{supp}(\widehat{\beta})|} \leq \sqrt{\phi_n(\widehat{m})}\|\widehat{\beta} - \beta_0\|_{2,n}. \tag{A.3}$$

Under the condition $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have by Theorems 1 and 2 that

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

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

$$\sqrt{|\text{supp}(\widehat{\beta})|} \leq 2\bar{c}\sqrt{\phi_n(\widehat{m})}\ \frac{n}{\lambda}\frac{\varrho_{\bar{c}} + \bar{\zeta}}{1-\bar{\zeta}^2} = \sqrt{s}\sqrt{\phi_n(\widehat{m})}\ 2\bar{c}\ \frac{1 + \varrho_{\bar{c}}/\bar{\zeta}}{\bar{\kappa}(1-\bar{\zeta}^2)}$$

where the last equality follows from $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}]$.

Let $L := 2\bar{c}\ \{1 + \varrho_{\bar{c}}/\bar{\zeta}\}/\{\bar{\kappa}(1-\bar{\zeta}^2)\}$. Consider any $m \in \mathcal{M}$, and suppose $\widehat{m} > m$. Therefore by the well-known sublinearity of sparse eigenvalues, $\phi_n(\ell m) \leq \lceil \ell \rceil\, \phi_n(m)$ for $\ell \geq 1$, and $\widehat{m} \leq |\text{supp}(\widehat{\beta})|$ we have $\widehat{m} \leq s \cdot \left\lceil \frac{\widehat{m}}{m} \right\rceil \phi_n(m)L^2$. Thus, since $\lceil k \rceil < 2k$ for any $k \geq 1$ we have $m < s \cdot 2\phi_n(m)L^2$ which violates the condition of $m \in \mathcal{M}$ and s. Therefore, we must have $\widehat{m} \leq m$. In turn, applying (A.4) once more with $\widehat{m} \leq m$ we obtain $\widehat{m} \leq s \cdot \phi_n(m)L^2$. The result follows by minimizing the bound over $m \in \mathcal{M}$.

To show the second part, by Lemma 2 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 1 yields $\bar{\kappa} \geq \kappa_{\bar{c}}$ so that $\bar{\zeta} \leq \lambda\sqrt{s}/(n\kappa_{\bar{c}})$. Therefore

$$\frac{n}{\lambda}\frac{\varrho_{\bar{c}} + \bar{\zeta}}{1-\bar{\zeta}^2} \leq \frac{\sqrt{s}}{\kappa_{\bar{c}}(1-\bar{\zeta}^2)}\left\{\frac{1+\bar{c}}{c} + 1\right\} = \frac{\bar{c}\sqrt{s}}{\kappa_{\bar{c}}(1-\bar{\zeta}^2)}.$$

Thus, under the condition $\bar{\zeta} \leq 1/\sqrt{2}$,

$$|\text{supp}(\widehat{\beta})| \leq s\ \phi_n(\widehat{m})\left(\frac{4\bar{c}^2}{\kappa_{\bar{c}}}\right)^2. \tag{A.4}$$

The same argument used before with $L = 4\bar{c}^2/\kappa_{\bar{c}}$ yields the second result. ∎

Proof of Lemma 5. Recall that $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_p)$. First note that by strong duality

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

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

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

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

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 $j \in \mathrm{supp}(\widehat{\beta})$ we have $\mathbb{E}_n[x_j\widehat{a}] = \mathrm{sign}(\widehat{\beta}_j)\lambda\gamma_j/n$. Therefore, for any subset $G \subseteq \mathrm{supp}(\widehat{\beta})$ we have

$$\begin{aligned}
&\sqrt{\widehat{Q}(\widehat{\beta})}\sqrt{|G|}\lambda = \|\Gamma^{-1}(X'(Y - X\widehat{\beta}))_G\| \\
&\leq \|\Gamma^{-1}(X'(Y - X\beta_0))_G\| + \|\Gamma^{-1}(X'X(\beta_0 - \widehat{\beta}))_G\| \\
&\leq \sqrt{|G|}\; n\|\Gamma^{-1}\mathbb{E}_n[x(\sigma\epsilon + r)]\|_\infty + n\sqrt{\phi_{\max}(|G|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n} \\
&= \sqrt{|G|}\; n\sqrt{\widehat{Q}(\beta_0)}\|\Gamma^{-1}\widetilde{S}\|_\infty + n\sqrt{\phi_{\max}(|G \setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n},
\end{aligned}$$

where we used

$$\begin{aligned}
&\|\Gamma^{-1}(X'X(\widehat{\beta} - \beta_0))_G\| \leq \sup_{\|\alpha_{Tc}\|_0 \leq |G\setminus T|, \|\alpha\| \leq 1} |\alpha'\Gamma^{-1}X'X(\widehat{\beta} - \beta_0))| \\
&\leq \sup_{\|\alpha_{Tc}\|_0 \leq |G\setminus T|, \|\alpha\| \leq 1} \|\alpha'\Gamma^{-1}X'\|\|X(\widehat{\beta} - \beta_0)\| \\
&\leq n\sqrt{\phi_{\max}(|G \setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}.
\end{aligned}$$

■

Proof of Theorem 4. 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 $f - X\widetilde{\beta} = (I - \mathcal{P}_{\widehat{T}})f - \sigma\mathcal{P}_{\widehat{T}}\epsilon$ where I is the identity operator. Therefore we have

$$\begin{aligned}
&\sqrt{n}\|\beta_0 - \widetilde{\beta}\|_{2,n} = \|X\beta_0 - X\widetilde{\beta}\| = \|f - X\widetilde{\beta} - R\| \\
&= \|(I - \mathcal{P}_{\widehat{T}})f - \sigma\mathcal{P}_{\widehat{T}}\epsilon - R\| \leq \|(I - \mathcal{P}_{\widehat{T}})f\| + \sigma\|\mathcal{P}_{\widehat{T}}\epsilon\| + \|R\|
\end{aligned} \tag{A.5}$$

where we have $\|R\| \leq \sqrt{n}c_s$. Since for $\widehat{m} = |\widehat{T} \setminus T|$, we have

$$\|X[\widehat{T}](X[\widehat{T}]'X[\widehat{T}])^{-1}\|_{op} \leq \sqrt{1/\phi_{\min}(\widehat{m}, \mathbb{E}_n[xx'])} = \sqrt{1/\phi_{\min}(\widehat{m})},$$

the term $\|\mathcal{P}_{\widehat{T}}\epsilon\|$ in (A.5) satisfies

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

Therefore, we have

$$\|\widetilde{\beta}-\beta_0\|_{2,n} \leq \frac{\sigma\sqrt{s+\widehat{m}}\|\mathbb{E}_n[x\epsilon]\|_\infty}{\sqrt{\phi_{\min}(\widehat{m})}} + c_s + c_{\widehat{T}}$$

where $c_{\widehat{T}} = \min_{\beta\in\mathbb{R}^p}\sqrt{\mathbb{E}_n[(f-x'\beta_{\widehat{T}})^2]}$. Since $\mathrm{supp}(\widehat{\beta}) \subseteq \widehat{T}$ and (3.5) holds,

$$\begin{aligned} c_{\widehat{T}} &= \min_{\beta\in\mathbb{R}^p}\{\mathbb{E}_n[(f-x'\beta_{\widehat{T}})^2]\}^{1/2} \leq \{\mathbb{E}_n[(f-x'\widehat{\beta})^2]\}^{1/2} \\ &\leq c_s + \|\beta_0-\widehat{\beta}\|_{2,n} \leq c_s + 2\sqrt{\widehat{Q}(\beta_0)}\tfrac{(\varrho_{\bar{c}}+\bar{\zeta})}{1-\bar{\zeta}^2}. \end{aligned}$$

where we have used Theorem 1. ∎

Proof of Theorem 5. 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 $\widehat{\beta} = \beta_0$. ∎

Proof of Theorem 6. If $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Theorem 1, for $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$ we have $\|\widehat{\beta}-\beta_0\|_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}\frac{\varrho_{\bar{c}}+\bar{\zeta}}{1-\bar{\zeta}^2}$, and the bound on the prediction norm follows by $\sqrt{\widehat{Q}(\beta_0)} \leq c_s + \sigma\sqrt{\mathbb{E}_n[\epsilon^2]}$.

Thus we need to show that the choice of λ and Γ is suitable for the desired probability on the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$. Since $\gamma_j = \max_{1\leq i\leq n}|x_{ij}| \geq \mathbb{E}_n[x_j^2] = 1$, by the choice of u_n we have

$$\begin{aligned} &P\left(c\|\Gamma^{-1}\widetilde{S}\|_\infty > \tfrac{\lambda}{n}\right) \leq P\left(c\max_{1\leq j\leq p}\tfrac{|\mathbb{E}_n[(\sigma\epsilon+r)x_j]|}{\gamma_j\sqrt{\mathbb{E}_n[(\sigma\epsilon)^2]}} > \tfrac{\lambda}{n(1+u_n)^{1/2}}\right) + \eta_1 \\ &\leq P\left(\max_{1\leq j\leq p}\tfrac{|\mathbb{E}_n[\epsilon x_j]|}{\gamma_j\sqrt{\mathbb{E}_n[\epsilon^2]}} > \tfrac{\sqrt{2\log(2p/\alpha)}}{\sqrt{n}}\right) + P\left(\tfrac{\|\mathbb{E}_n[rx]\|_\infty}{\sqrt{\mathbb{E}_n[(\sigma\epsilon)^2]}} > \tfrac{(1+u_n)^{1/2}}{\sqrt{n}}\right) + \eta_1. \end{aligned}$$

We invoke the following lemma, proven in SM:

Lemma 6. *Under Condition ASM we have* $\|\mathbb{E}_n[xr]\|_\infty \leq \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\}$.

By Lemma 6 $\|\mathbb{E}_n[rx]\|_\infty \leq \sigma/\sqrt{n}$ and $P(\mathbb{E}_n[\epsilon^2] \leq \{1+u_n\}^{-1}) \leq \eta_2$ we have $P\left(\frac{\|\mathbb{E}_n[rx]\|_\infty}{\sqrt{\mathbb{E}_n[(\sigma\epsilon)^2]}} > \frac{(1+u_n)^{1/2}}{\sqrt{n}}\right) \leq P\left(\sqrt{\mathbb{E}_n[(\sigma\epsilon)^2]} \leq \{1+u_n\}^{-1/2}\right) \leq \eta_2$.

To bound the last term we have

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

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

where we used the union bound and Theorem 2.15 of [24] because ϵ_i's are independent and symmetric. ∎

Proof of Lemma 6. First note that for every $j=1,\dots,p$, we have $|\mathbb{E}_n[x_jr]|\le\sqrt{\mathbb{E}_n[x_j^2]\mathbb{E}_n[r^2]}=c_s$. Next, by definition of β_0 in (2.2), for $j\in T$ we have $\mathbb{E}_n[x_j(f-x'\beta_0)]=\mathbb{E}_n[x_jr]=0$ since β_0 is a minimizer over the support of β_0. For $j\in T^c$ we have that for any $t\in\mathbb{R}$

$$\mathbb{E}_n[(f-x'\beta_0)^2]+\sigma^2\frac{s}{n}\le\mathbb{E}_n[(f-x'\beta_0-tx_j)^2]+\sigma^2\frac{s+1}{n}.$$

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

$$-\sigma^2/n\le\mathbb{E}_n[(f-x'\beta_0-tx_j)^2]-\mathbb{E}_n[(f-x'\beta_0)^2]=-2t\mathbb{E}_n[x_j(f-x'\beta_0)]+t^2\mathbb{E}_n[x_j^2].$$

Taking the minimum over t in the right hand side at $t^*=\mathbb{E}_n[x_j(f-x'\beta_0)]$ we obtain $-\sigma^2/n\le-(\mathbb{E}_n[x_j(f-x'\beta_0)])^2$ or equivalently, $|\mathbb{E}_n[x_j(f-x'\beta_0)]|\le\sigma/\sqrt{n}$. ∎

Proof of Corollary 1. See SM. ∎

APPENDIX B: PROOFS OF SECTION 4

Proof of Theorem 7. Conditions ASM, P and the choice of penalty level with $\alpha\in(1/n,1/\log n)$ imply the assumptions of Lemma 7 with $q=4$, $u_n=1/\log\log n$, the slack constant as $\sqrt{c}>1$, $\ell_n=\log^{1/3}n$, and $\eta=0$ (given the choice of $\widehat{\gamma}_{j,0}$) for $n\ge n_0$ for sufficiently large n_0. Thus by Lemma 7 we have $\lambda/n\ge\sqrt{c}\|\widehat{\Gamma}^{-1}\widetilde{S}\|_\infty$ with probability at least $1-\alpha\{1+\bar{C}/\log n\}-\bar{C}\{n^{-1/2}\log n\}$.

Moreover, since $\bar{c}=(\bar{c}+1)/(\bar{c}-1)$, $\kappa_{\bar{c}}(\widehat{\Gamma}_0)\ge\kappa_{\bar{c}}(I)/\max_{i\le n}\|x_i\|_\infty$, $\kappa_{\bar{c}}(I)$ is bounded away from zero for large enough n_0 by the sparse eigenvalue conditions,

$$\lambda\lesssim\sqrt{n\log(p/\alpha)}\lesssim\sqrt{n\log(p\vee n)},$$

and the condition $\max_{i\le n}\|x_i\|_\infty^4 s\log(p\vee n)\le Cn/\log n$, the side condition in Lemma 7. The results then follow from Theorems 1, 3 and 4 because $\bar{\kappa}\ge\kappa_{\bar{c}}(\widehat{\Gamma}_0)$ by Lemma 1, $\bar{\varrho}_{\bar{c}}\le C\lambda\sqrt{s}/n\kappa_{\bar{c}}(\widehat{\Gamma}_0)$ by Lemma 2, and $\bar{\zeta}\le C\lambda\sqrt{s}/n$. The result for the ℓ_1-rate follows from $\|\check{\beta}-\beta_0\|_1\le\sqrt{\bar{C}'s}\|\check{\beta}-\beta_0\|_{2,n}/\sqrt{\phi_{\min}(C's)}\le C''\sqrt{s}\|\check{\beta}-\beta_0\|_{2,n}$ for $\check{\beta}=\widehat{\beta}$ and $\check{\beta}=\widetilde{\beta}$. ∎

Proof of Corollary 2. See SM. ∎

APPENDIX C: PROOFS FOR SECTION 5

Proof of Theorem 8. Throughout the proof we use the notation

$$B(w) := \max_{j,k} \sup_{\nu \in \Theta \times T} |\partial_{\nu_k} \psi_j(w, \nu)|, \quad \tau_n := \sqrt{s \log(p \vee n)/n}.$$

Step 1. (A Preliminary Rate Result). In this step we claim that $\|\widehat{\theta} - \theta_0\| \lesssim_P \tau_n$. By definition $\|\mathbb{E}_n \psi(w, \widehat{\theta}, \widehat{h}(z))\| \leq \epsilon_n$ and $\widehat{\theta} \in \Theta$, which implies via triangle inequality that:

$$\left\| [\bar{\mathbf{E}}[\psi(w, \theta, h_0(z))]|_{\theta = \widehat{\theta}} \right\| \leq \epsilon_n + I_1 + I_2 \lesssim_P \tau_n,$$

where I_1 and I_2 are defined in Step 2 below, and the last bound also follows from Step 2 below and from the numerical tolerance obeying $\epsilon_n = o(n^{-1/2})$ by assumption. Since by condition SP(iv), $2^{-1}(\|J_n(\widehat{\theta} - \theta_0)\| \vee c)$ is weakly smaller than the left side of the display, we conclude that $\|\widehat{\theta} - \theta_0\| \lesssim_P (\text{mineig}(J_n))^{-1} \tau_n$, which gives the stated claim since $\text{mineig}(J_n)$ is bounded away from zero uniformly in n by condition SP (v).

Step 2 (Define and bound I_1 and I_2.) We claim that:

$$\begin{aligned} I_1 &:= \sup_{\theta \in \Theta} \left\| \mathbb{E}_n \psi(w, \theta, \widehat{h}(z)) - \mathbb{E}_n \psi(w, \theta, h_0(z)) \right\| \lesssim_P \tau_n, \\ I_2 &:= \sup_{\theta \in \Theta} \left\| \mathbb{E}_n \psi(w, \theta, h_0(z)) - \bar{\mathbf{E}} \psi(w, \theta, h_0(z)) \right\| \lesssim_P n^{-1/2}. \end{aligned}$$

Using Taylor's expansion, for $\tilde{h}(z; \theta, j)$ denoting a point on a line connecting vectors $h_0(z)$ and $h(z)$, which can depend on θ and j,

$$\begin{aligned} I_1 &\leq \sum_{j=1}^{d} \sum_{m=1}^{M} \sup_{\theta \in \Theta} \left\| \mathbb{E}_n [\partial_{t_m} \psi_j(w, \theta, \tilde{h}(z; \theta, j))(\widehat{h}_m(z) - h_{0m}(z))] \right\| \\ &\leq dM\sqrt{d} \{\mathbb{E}_n B^2(w)\}^{1/2} \max_m \{\mathbb{E}_n (\widehat{h}_m(z) - h_{0m}(z))^2\}^{1/2}, \end{aligned}$$

where the last inequality holds by definition of $B(w)$ given earlier and the Holder's inequality. Since $\bar{\mathbf{E}} B^2(w) \leq C$ by condition SP(ii), $\mathbb{E}_n B^2(w) \lesssim_P 1$ by Markov's inequality. By this, by condition AS(ii), by d and M fixed, conclude that $I_1 \lesssim_P \tau_n$.

Using Jain-Marcus theorem, as stated in Example 2.11.13 in [57], we conclude that $\sqrt{n} I_2 \lesssim_P 1$. Indeed the hypotheses of that example follow from the assumption that Θ is a fixed compact subset of $\mathbb{R}^d$, and from the Lipschitz property, $\|\psi(w, \theta, h_0(z)) - \psi(w, \tilde{\theta}, h_0(z))\| \lesssim B(w) \|\tilde{\theta} - \theta\|$ holding uniformly for all θ and $\tilde{\theta}$ in Θ, with $\bar{\mathbf{E}} B^2(w) \leq C$ by condition SP(ii).

Step 3. (Main Step) We have that $\sqrt{n}\|\mathbb{E}_n\psi(w,\widehat{\theta},\widehat{h}(z))\| \leq \epsilon_n\sqrt{n}$. Application of Taylor's theorem and the triangle inequality gives

$$\left\|\sqrt{n}\mathbb{E}_n\psi(w,\theta_0,h_0(z)) + J_n\sqrt{n}(\theta-\theta_0)\right\| \leq \epsilon\sqrt{n} + \|II_1\| + \|II_2\| + \|II_3\| = o_P(1),$$

where the terms II_1, II_2, and II_3 are defined and bounded below in Step 4; the $o_P(1)$ bound follows from Step 4 and from $\epsilon_n\sqrt{n} = o(1)$ holding by assumption. Conclude using condition SP(iv) that

$$\left\|J_n^{-1}\sqrt{n}\mathbb{E}_n\psi(w,\theta_0,h_0(z)) + \sqrt{n}(\theta-\theta_0)\right\| \leq o_P(1)(\text{mineg}(J_n))^{-1} = o_P(1),$$

which verifies the first claim of the theorem. Application of Liapunov's central limit theorem in conjunction with condition SP(v) and the conditions on Ω_n imposed by the theorem imply the second claim.

Step 4. (Define and Bound II_1, II_2, and II_3). Let $II_1 := (II_{1j})_{j=1}^d$ and $II_2 = (II_{2j})_{j=1}^d$, where

$$\begin{aligned}
II_{1j} &:= \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\left[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))(\widehat{h}_m(z) - h_{0m}(z))\right], \\
II_{2j} &:= \sum_{r,k=1}^K \sqrt{n}\mathbb{E}_n\left[\partial_{\nu_k}\partial_{\nu_r}\psi_j(w,\tilde{\nu}(w;j))\{\widehat{\nu}_r(w) - \nu_{0r}(w)\}\{\widehat{\nu}_k(w) - \nu_{0k}(w)\}\right], \\
II_3 &:= \sqrt{n}\left(\mathbb{E}_n\partial_\theta\psi(w,\theta_0,h_0(z)) - J_n\right)(\theta-\theta_0),
\end{aligned}$$

where $\nu_0(w) := (\nu_{0k})_{k=1}^K := (\theta_0', h_0(z)')'$; $K = d + M$; $\widehat{\nu}(w) := (\widehat{\nu}_k(w))_{k=1}^K := (\widehat{\theta}', \widehat{h}(z)')'$, and $\tilde{\nu}(w;j)$ is a vector on the line connecting $\nu_0(w)$ and $\widehat{\nu}(w)$ that may depend on j; and $J_n = \bar{\mathbf{E}}\partial_\theta\psi(w,\theta_0,h_0(z))$. We show in this step that $II_1 + II_2 + II_3 \lesssim_P \tau_n$,

The key portion of the proof is bounding II_{1j}, which is very similar to the argument first given in [3] (pp. 2421-2423). We repeat it here for completeness. In order to bound II_{1j} we split it $II_{1j} = III_{1j} + III_{2j}$, where

$$\begin{aligned}
III_{1j} &:= \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\left[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))\sum_{l=1}^p P_l(z)(\widehat{\beta}_{ml} - \beta_{0ml})\right], \\
III_{2j} &:= \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\left[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))r_m(z)\right].
\end{aligned}$$

Using Holder inequality, $\max_j |III_{1j}| \leq M\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{jml}(w)|\|\widehat{\beta}_m - \beta_{0m}\|_1$. Now $\max_m\|\widehat{\beta}_m - \beta_{0m}\|_1 \leq C\sqrt{s}\tau_n$ with probability at least $1-\delta_n$ by condition AS(i). Moreover, using they key property that $\mathrm{E}\dot{\psi}_{jml}(w_i) = 0$ which holds by the orthogonality property and that $\max_{j,m,l}\mathbb{E}_n|\psi_{jml}(w)|^2 \leq L_n^2$ with probability at least $1-\delta_n$ by condition AS(ii), we can apply the moderate deviation inequality for self-normalized sum, following the idea in

[3], to conclude that $\max_j |III_{1j}| \leq \sqrt{2\log(pn)}L_n$ with probability $1-o(1)$. Note that this application requires the side condition $\sqrt{2\log(pn)}B_n n^{-1/6} = o(1)$ be satisfied for B_n defined in condition AS(ii), which indeed holds by condition AS(iii). We now recall the details of this calculation:

$$
\begin{aligned}
&P\left(\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{jml}(w)| > \sqrt{2\log(pn)}L_n\right)\\
&\leq P\left(\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{jml}(w)|/\sqrt{\mathbb{E}_n|\dot{\psi}_{jml}(w)|^2} > \sqrt{2\log(pn)}\right) + \delta_n\\
&\leq dMp\max_{j,m,l}P\left(|\sqrt{n}\mathbb{E}_n\dot{\psi}_{jml}(w)|/\sqrt{\mathbb{E}_n|\dot{\psi}_{jml}(w)|^2} > \sqrt{2\log(pn)}\right) + \delta_n\\
&\leq dMp\Phi(\sqrt{2\log(pn)})(1+o(1)) + \delta_n \leq dMp\frac{1}{pn}(1+o(1)) + \delta_n = o(1),
\end{aligned}
$$

where the penultimate inequality occurs due to the application of the moderate deviation theorems for self-normalized sums. Putting bounds together we conclude that $III_1 \lesssim_P L_n\sqrt{\log(p\vee n)}\sqrt{s}\tau_n = o(1)$, where $o(1)$ holds by the growth restrictions imposed in condition AS(iii).

The bound on III_2 also follows similarly to [3]. III_{2j} is a $n^{-1/2}$ times the sums of M terms each having mean zero and variance of order $s\log(p\vee n)/n = o(1)$. Indeed, the mean zero occurs because

$$
n^{-1}\sum_{i=1}^n \mathrm{E}\left[\partial_{t_m}\psi_j(w_i,\theta_0,h_0(z_i))r_m(z_i)\right] = n^{-1}\sum_{i=1}^n \mathrm{E}[0\cdot r_m(z_i)] = 0,
$$

for each m-th term, which holds by $\mathrm{E}[\partial_{t_m}\psi_j(w_i,\theta_0,h_0(z_i))|z_i] = 0$, i.e. the orthogonality property, and the law of iterated expectations. To derive the variance bound, note that for each m-th term,

$$
n^{-1}\sum_{i=1}^n \mathrm{E}[\{\partial_{t_m}\psi_j(w_i,\theta_0,h_0(z_i))\}^2 r_m^2(z_i)] \leq C\bar{\mathbf{E}}[r_m^2(z)] \leq C^2 s\log(p\vee n)/n,
$$

which holds by $\mathrm{E}[\{\partial_{t_m}\psi_j(w_i,\theta_0,h_0(z_i))\}^2|z_i] \leq \mathrm{E}[B^2(w)|z_i] \leq C$ a.s. by virtue of condition SP(iii), and the law iterated expectations; the last bound in the display holds by AS(i). Hence $\mathrm{var}(III_{2j}) \leq nn^{-1}M^2C^2 s\log(p\vee n)/n \lesssim s\log(p\vee n)/n = o(1)$. Therefore, $\|III_2\| \leq \sum_{j=1}^d |III_{2j}| \lesssim_P s\log(p\vee n)/n$ by Chebyshev's inequality.

To deduce that $\|II_2\| = o_P(1)$, we use condition AS(i)-(iii), the claim of Step 1, and Holder inequalities, concluding that

$$
\max_j |II_{2j}| \leq \sqrt{n}K^2 L_n \max_k \mathbb{E}_n\{\widehat{\nu}_k(w) - \nu_{0k}(w)\}^2 \lesssim_P \sqrt{n}L_n\tau_n^2 = o(1).
$$

Finally, since $\|II_3\| \leq \sqrt{n}\| \left(\mathbb{E}_n \partial_\theta \psi(w, \theta_0, h_0(z)) - J_n\right) \|_{op} \|\widehat{\theta} - \theta_0\|$ and since $\|\mathbb{E}_n \partial_\theta \psi(w, \theta_0, h_0(z)) - J_n\|_{op} \lesssim_P n^{-1/2}$ by Cbebyshev's inequality, using that $\bar{\mathbf{E}} B^2(w) \leq C$ by condition AS(ii), and $\|\widehat{\theta} - \theta_0\| \lesssim_P \tau_n$ by Step 1, conclude that $\|II_3\| \lesssim_P \tau_n$. ∎

ACKNOWLEDGEMENTS

We are grateful to the editors and two referees for thoughtful comments and suggestions, which helped improve the paper substantially. We also thank seminar participants at the Joint Statistical Meetings, INFORMS, Duke and MIT for many useful suggestions. We gratefully acknowledge research support from the NSF.

SUPPLEMENTARY MATERIAL

Supplementary Material for the paper "Pivotal Estimation via Square-root Lasso for Nonparametric Regression": Additional Results and Simulations
(http://arxiv.org/abs/1105.1475). The supplementary material contains deferred proofs, additional theoretical results on convergence rates in ℓ_2, ℓ_1, and ℓ_∞, lower bound on the prediction rate, Monte Carlo simulations, and auxiliary probability inequalities.

REFERENCES

[1] Takeshi Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pages 955–968, 1977.

[2] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *ArXiv*, 2010.

[3] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012. Arxiv, 2010.

[4] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[5] A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. *Inverse problems and high dimensional estimation - Stats in the Château summer school in econometrics and statistics, 2009, Springer Lecture Notes in Statistics - Proceedings*, pages 121–156, 2011.

[6] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. Arxiv, 2009.

[7] A. Belloni, V. Chernozhukov, I. Fernandez-Val, and C. Hansen. Program evaluation with high-dimensional data. *arXiv:1311.2645*, 2013.

[8] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. ArXiv, 2010.

[9] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011. forthcoming, The Review of Economic Studies.

[10] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2013. ArXiv, 2011.

[11] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for lad regression models. *ArXiv*, 2013.

[12] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. Arxiv, 2010.

[13] A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *ArXiv:1304.3969*, 2013.

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

[15] P. Bülhmann and S. van de Geer. *Statistics for High-Dimensional Data*. Statistics (New York). Springer-Verlag, Berlin, 2011. Methods, Theory and Applications.

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

[17] 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.

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

[19] E. Candès and Y. Plan. Near-ideal model selection by l_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.

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

[21] Gary Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596, 1992.

[22] Y. Chen and A. S. Dalalyan. Fused sparsity and robust estimation for linear models with unknown variance. *Advances in Neural Information Processing Systems*, 25:1268–1276, 2012.

[23] S. Chrétien and S. Darses. Sparse recovery with unknown variance: a LASSO-type approach. *arXiv:1101.04334*, 2012.

[24] V. H. de la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.

[25] 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.

[26] M. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *University of Michigan Department of Economics Working Paper*, 2013.

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

[28] E. Gautier and A. Tsybakov. High-dimensional instrumental variables rergession and confidence sets. arXiv:1105.2454v2 [math.ST], 2011.

[29] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *arXiv:1109.5587v2*, 2012.

[30] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

[31] Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

[32] B.-Y. Jing, Q.-M. Shao, and Q. Wang. Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[33] O. Klopp. High dimensional matrix estimation with unknown variance of the noise.

arXiv, (arXiv:1112.3055), 2011.
[34] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.
[35] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.
[36] M. C. Veraar L. Duembgen, S. A. van de Geer and J. A. Wellner. Nemirovski's inequalities revisited. *American Mathematical Monthly*, 117:138–160, 2010.
[37] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
[38] 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.
[39] Z. Lu. Gradient based method for cone programming with application to large-scale compressed sensing. *Technical Report*, 2008.
[40] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
[41] Y. Nesterov. Smooth minimization of non-smooth functions, mathematical programming. *Mathematical Programming*, 103(1):127–152, 2005.
[42] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.
[43] Y. Nesterov and A Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1993.
[44] Zhao R., T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.
[45] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2001.
[46] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
[47] 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.
[48] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19:209–285, 2010.
[49] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.
[50] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.
[51] 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.
[52] S. A. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.
[53] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
[54] 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.
[55] S. A. van de Geer, P. Bühlmann, and Y. Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *ArXiv*, 2013.
[56] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
[57] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.
[58] B. von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist*, 36:299–303, 1965.

[59] 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.

[60] L. Wang. L1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135151, September 2013.

[61] 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.

[62] C.-H. Zhang and S. S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.

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

APPENDIX A: ADDITIONAL RESULTS FOR SECTION 3

A.1. Bounds in Other Norms. This section contains bounds for ℓ_1 and ℓ_∞-rates and lower bounds on prediction norm Rates.

THEOREM 9 (Lower Bound on Prediction Error). *Under Condition ASM and (3.5), let* $\widehat{m} = |\text{supp}(\widehat{\beta}) \setminus T|$. *We have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \geq \frac{\lambda\sqrt{|\text{supp}(\widehat{\beta})|}\,\sqrt{\widehat{Q}(\beta_0)}}{n\sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}}\left(1 - \frac{1}{c} - \frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\zeta})}{1 - \bar{\zeta}^2}\right).$$

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 [38], the lower bound does not have the term $\widehat{Q}^{1/2}(\beta_0)$ 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.

THEOREM 10 (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 $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, *we have*

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

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

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

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

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

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 $\|\Gamma^{-1}\mathbb{E}_n[xx' - I]\Gamma^{-1}\|_\infty$ is small. In fact, near orthogonality also allows to bound the restricted eigenvalue $\kappa_{\bar{c}}$ from below. In the homoscedastic case for Lasso (which corresponds to $\Gamma = I$) [14] and [37] established that if for some $u \geq 1$ we have $\|\mathbb{E}_n[xx'] - I\|_\infty \leq 1/(u(1+\bar{c})s)$ then $\kappa_{\bar{c}} \geq \sqrt{1 - 1/u}$. In that case, the first term determines the rate of convergence in the ℓ_∞-norm.

APPENDIX B: DEFERRED PROOFS FOR RESULTS AND ADDITIONAL RESULTS FOR SECTION 3

B.1. Deferred Proofs for Results in Section 3.

PROOF OF COROLLARY 1. We will bound the probability of relevant events to establish $\lambda/n \geq c\|\Gamma^{-1}\widehat{S}\|_\infty$ and consequentially the prediction norm bound by Theorem 6.

Applying Lemma 13(ii) with $a = 1/\log n$ we have $\eta_2 = \frac{1}{n^{1/2}(1/6 - 1/\log n)^2} = \frac{36\log^2 n}{n^{1/2}(\log n - 6)^2}$.

Applying Lemma 13(iii) with $t_n = 4n/\tau$, $a = 1/\log n$, and $a_n = u_n/[1+u_n]$, where we note the simplification that

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

we have

$$P\left(\sqrt{\mathbb{E}_n[\sigma^2\epsilon^2]} \leq (1+u_n)\sqrt{\mathbb{E}_n[(\sigma\epsilon + r)^2]}\right) \leq \eta_1 := \frac{2\log(4n/\tau)}{nu_n/[1+u_n]} + \eta_2 + \frac{\tau}{2}.$$

Thus, by Theorem 6, since $\bar{\zeta} < 1$, with probability at least $1 - \alpha - \eta_1 - \eta_2$ we have

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

Finally, by Lemma 13(i) we have $\mathbb{E}_n[\epsilon^2] \leq 2\sqrt{2}/\tau + \log(4n/\tau)$ with probability at least $1 - \tau/2$. ∎

PROOF OF LEMMA 3. In this step we show that $\widehat{\delta} = \widehat{\beta} - \beta_0 \in \Delta_{\bar{c}}$ under the prescribed penalty level. By definition of $\widehat{\beta}$

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

where the last inequality holds because

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

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

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

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

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

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

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

that is

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

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

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

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

B.2. Proofs of Additional Results for Section 3.

PROOF OF THEOREM 9. We can assume that $\widehat{Q}(\beta_0) > 0$ otherwise the result is trivially true. In the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Lemma 5 with $G = \text{supp}(\widehat{\beta})$ we have

$$\left(\sqrt{\frac{\widehat{Q}(\widehat{\beta})}{\widehat{Q}(\beta_0)}} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|G|} \leq \sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}. \tag{B.9}$$

Under the condition $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have by the lower bound in Theorem 2

$$\left(1 - \frac{1}{c} - \frac{2\varrho_{\bar{c}}(\varrho_{\bar{c}} + \bar{\zeta})}{1 - \bar{\zeta}^2}\right) \frac{\lambda\sqrt{\widehat{Q}(\beta_0)}\sqrt{|G|}}{n\sqrt{\phi_{\max}(\widehat{m}, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}} \leq \|\widehat{\beta} - \beta_0\|_{2,n}.$$

■

PROOF OF THEOREM 10. 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 $\|\delta\|_{2,n}$.

■

PROOF OF THEOREM 11. Let $\delta := \widehat{\beta} - \beta_0$. We have that

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

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

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

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

Next let e_j denote the jth-canonical direction.

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

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

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

The result follows from Theorem 1 and 10. ■

APPENDIX C: ADDITIONAL RESULTS FOR SECTION 4

C.1. An Analysis of the Penalty Level and Loadings for $\sqrt{\text{Lasso}}$. 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

$$\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty, \quad \text{where} \quad c > 1, \tag{C.1}$$

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 [14]'s choice of penalty level for Lasso under homoscedasticity and known σ. Here, in order to account for heteroscedasticity the penalty level λ/n needs to majorate the score rescaled by the penalty loadings.

REMARK 8 (Pivotality in the Parametric Case). In the parametric case, $r_i = 0$, $i = 1, \ldots, n$, the score does not depend on σ nor β_0. Under the homoscedastic 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 quantile of $\|\Gamma^{-1}\widetilde{S}\|_\infty$, see [12].

In order to achieve Gaussian-like behavior under heteroscedastic non-Gaussian noise 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. That is, under these conditions the penalty level λ can be set to $\sqrt{n}\Phi^{-1}(1-\alpha/[2p])$ as in the Gaussian noise case despite the non-Gaussian noise.

Let the penalty level λ be set to

$$\lambda = (1+u_n)c\sqrt{n}\{\Phi^{-1}(1-\alpha/[2p]) + 1 + u_n\} \tag{C.2}$$

and the initial penalty loadings $\widehat{\gamma}_{j,0}$ are given by

$$\widehat{\gamma}_{j,0}^{(I)} = \max_{1\leq i\leq n} |x_{ij}| \quad \text{or} \quad \widehat{\gamma}_{j,0}^{(II)} = W\{\mathbb{E}_n[x_j^4]\}^{1/4}$$

where $W > \{\bar{\mathbf{E}}[\epsilon^4]\}^{1/4}/\{\bar{\mathbf{E}}[\epsilon^2]\}^{1/2}$ is the kurtosis parameter discussed in Remark 7. In this section we focus on the following set of regularities conditions.

CONDITION D. *There exist a finite constant $q \geq 4$ such that* $\sup_{n\geq 1}\bar{\mathbf{E}}[|\epsilon|^q] < \infty$, *and* $\sup_{n\geq 1}\max_{1\leq j\leq p}\mathbb{E}_n[|x_j|^q] < \infty$.

CONDITION R. *Let* $w_n = \left(\alpha^{-1}\log nC_q\bar{\mathbf{E}}[|\epsilon|^{q\vee 4}]\right)^{1/q}/n^{1/4} < 1/2$, *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*

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

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 $\gamma_j^* = 1 \vee \sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}/\sqrt{\mathbb{E}_n[\epsilon^2]}$, $j = 1,\ldots,p$.

LEMMA 7. *Suppose that Conditions ASM, D and R hold. Consider the choice of penalty level* λ *in (C.2) and penalty loadings* Γ_k, $k \geq 0$, *in Algorithm 1. For* $k = 0$ *we have that*

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

where $\eta = 0$ *if* $\widehat{\gamma}_{j,0}^{(I)}$ *is used, and*

$$\eta = \frac{C_q\bar{\mathbf{E}}[|\epsilon|^{q\vee 8}]}{(W^4-\bar{\mathbf{E}}[\epsilon^4])^{q/4}n^{q/8}} \wedge \frac{2\bar{\mathbf{E}}[|\epsilon|^q]}{n^{1\wedge(q/4-1)}(W^4-\bar{\mathbf{E}}[\epsilon^4])^{q/4}}$$

if $\widehat{\gamma}_{j,0}^{(II)}$ *is used. Moreover, conditioned on* $\lambda/n \geq c\|\widehat{\Gamma}_0^{-1}\widetilde{S}\|_\infty$, *provided*

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

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, [32] and [24]. 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 than the condition $\log p \leq (q-2)\log n$ required in [12]. However, condition D is more stringent than some conditions in [12] thus neither set of condition dominates the other.

C.2. Performance in Sobolev classes. Here were provide additional comparisons to projection estimators under orthonormal random design for cases where the regression function belongs to the Sobolev space or rearranged versions of the Sobolev space. We will also provide bounds on the upper bounds on the sparsity s of the solution β_0 of the oracle problem and on the corresponding mean square of the approximation error c_s.

Throughout this section we consider the nonparametric model (2.1) where f is a function from $[0,1]$ to $\mathbb{R}$, $\epsilon_i \sim N(0,1)$ and $z_i \sim \text{Uniform}(0,1)$, which are independent across $i = 1,\ldots,n$. We will assume that the given basis $\{P_j(\cdot), j \geq 1\}$ is bounded and orthonormal.

The projection estimator with k terms is defined as

$$\widehat{f}^{(k)}(z) = \sum_{j=1}^{k} \widehat{\theta}_j P_j(z) \quad \text{where } \widehat{\theta}_j = \mathbb{E}_n[yP_j(z)]. \tag{C.3}$$

Projection estimators are particularly appealing in orthonormal designs considered here but are applicable to other designs as well.

EXAMPLE 1 **Series Approximations in Sobolev Balls**. Similarly to [50], for an orthonormal bounded basis $\{P_j(\cdot)\}_{j=1}^{\infty}$ in $L^2[0,1]$, consider functions $f(z) = \sum_{j=1}^{\infty} \theta_j P_j(z)$ in a Sobolev space $\mathcal{S}(\alpha, L)$ for some where $\alpha \geq 1$ and $L > 0$. This space consist of functions whose Fourier coefficients θ satisfy $\sum_{j=1}^{\infty} |\theta_j| < \infty$ and

$$\theta \in \Theta(\alpha, L) = \left\{\theta \in \ell^2(\mathbf{N}) : \ \textstyle\sum_{j=1}^{\infty} j^{2\alpha}\theta_j^2 \leq L^2 \ \right\}.$$

Among the projection estimators defined in (C.3) the oracle projection estimator picks k^* to minimize the mean squared error. The following characterize the performance of the projection estimators (C.3).

LEMMA 8. *Consider the nonparametric model (2.1) where $f : [0,1] \to \mathbb{R}$ belongs to the Sobolev space $\mathcal{S}(\alpha, L)$ with $\alpha \geq 1$ and $L > 0$, and $z_i \sim$ Uniform$(0,1)$, independent across $i = 1,\ldots,n$. Given a bounded orthonormal basis $\{P_j(\cdot)\}_{j=1}^{\infty}$, the coefficients of the projection estimator satisfy for any $k \leq \sqrt{n/\log n}$*

$$\mathbb{E}_n[(f(z) - \widehat{f}^{(k)}(z))^2] \lesssim_P k^{-2\alpha} + \frac{k}{n}.$$

By selecting the optimal value of k^* which balances $(k^*)^{-2\alpha}$ and k^*/n, the rate-optimal choice of the number of series terms k^* satisfies $k^* \leq \lfloor V n^{\frac{1}{2\alpha+1}} \rfloor$

(which is smaller than $\sqrt{n/\log n}$ for n sufficiently large). In turn this implies an upper bound on the oracle projection estimator risk given by

$$(k^*)^{-2\alpha} + \sigma^2 \frac{k^*}{n} \lesssim \sigma^2 n^{-\frac{2\alpha}{2\alpha+1}}$$

which achieves minimax bounds, see [50]. ∎

EXAMPLE 2 **Series Approximations in Rearranged Sobolev Balls.** Consider functions in a p-Rearranged Sobolev space $\mathcal{RS}(\alpha, p, L)$. These functions take the form $f(z) = \sum_{j=1}^{\infty} \theta_j P_j(z)$ such that $\sum_{j=1}^{\infty} |\theta_j| < \infty$ and $\theta \in \Theta^R(\alpha, p, L)$, where

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

Note that $\mathcal{S}(\alpha, L) \subset \mathcal{RS}(\alpha, p, L)$.

The class of rearranged Sobolev functions reduces significantly the relevance of the order of the basis. For each function f, the permutation Υ_f makes the sequence $\{|\theta_{\Upsilon_f(j)}|\}_{j=1}^p$ non-increasing. In particular, this weakly improves upon the error due to truncation in the conventional series estimator described in Example 1 since for any k

$$\sum_{j=k+1}^{p} \theta_j^2 \geq \sum_{j \in \{1,\ldots,p\} \setminus \{\Upsilon_f(1),\ldots,\Upsilon_f(k)\}} \theta_j^2.$$

Next we consider using the solution β_0 of the oracle problem in the main text to approximate a regression function $f \in \mathcal{RS}(\alpha, p, L)$. Recall the oracle problem of choosing the best s-dimensional subspace to approximate the regression function. The oracle problem solves $s \in \arg\min_k c_k^2 + \sigma^2 \frac{k}{n}$ where

$$c_k^2 := \min_{\|\beta\|_0 \leq k} \mathbb{E}_n[(f - \textstyle\sum_{j=1}^{p} \beta_j P_j(z))^2].$$

Lemma 9 below bounds the sparsity s of the oracle solution β_0 and also bounds the mean squared of the approximation errors. Since $\sqrt{\text{Lasso}}$ achieves the oracle rates up to a $\sqrt{\log(p \vee n)}$ factor, this result will allow us to understand the performance of the $\sqrt{\text{Lasso}}$ estimator for regression functions in $\mathcal{RS}(\alpha, p, L)$.

LEMMA 9. *Given a bounded orthonormal basis $\{P_j(\cdot)\}_{j=1}^{\infty}$, consider the nonparametric model (2.1) where $f : [0,1] \to \mathbb{R}$ belongs to a p-Rearranged Sobolev ball $\mathcal{RS}(\alpha, p, L)$ with $\alpha \geq 1$ and $L > 0$, and $z_i \sim \text{Uniform}(0,1)$, independent across $i = 1, \ldots, n$. We have that with probability $1 - o(1)$*

$$s \leq \bar{K} n^{\frac{1}{1+2\alpha}}, \quad c_s^2 + s/n \leq \bar{K} n^{\frac{-2\alpha}{1+2\alpha}}$$

where the constant $\bar{K}$ depends only on α and L.

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

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

However, in many cases the oracle approximation can improve substantially over the standard series approximation associated with projection estimators. For example, suppose that Fourier coefficients feature the following pattern $\theta_j = 0$ for $j \leq j_0$ and $|\theta_j| \leq K j^{-a}$ 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 oracle approximation avoids the first unnecessary n term to achieve consistency. Furthermore, under these regularities conditions (see Theorem 7), without knowing the exact support, the $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$ achieves with probability $1 - o(1)$

$$\begin{aligned} \{\mathbb{E}_n[(f(z) - \textstyle\sum_{j=1}^{p} \widehat{\beta}_j P_j(z))^2]\}^{1/2} &\leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s \\ &\lesssim \sigma \sqrt{s \log(p \vee n)/n} + c_s \\ &\lesssim n^{-\alpha/[2\alpha+1]} \sqrt{\log(p \vee n)}. \end{aligned}$$

In the case of the sparse model described above in which the first n components are not relevant, the adaptivity of $\sqrt{\text{Lasso}}$ allows it to preserve its rate while no series projection estimator will be consistent. ∎

APPENDIX D: DEFERRED PROOFS FOR RESULTS AND ADDITIONAL RESULTS FOR SECTION 4

D.1. Deferred Proofs for Results in Section 4.

PROOF OF COROLLARY 2. Note that

$$\begin{aligned} \widehat{Q}(\widehat{\beta}) - \sigma^2 &= \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) + \left(\mathbb{E}_n[\sigma^2 \epsilon_i^2 + 2\sigma r \epsilon + r^2] - \sigma^2\right) \\ &= \{\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\} + \sigma^2 \mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]] + 2\sigma \mathbb{E}_n[r\epsilon] + \mathbb{E}_n[r^2]. \end{aligned} \tag{D.1}$$

The second term in (D.1) is standard since $\bar{\mathbf{E}}[|\epsilon|^q] \leq C$ for $q > 4$. Because $\mathrm{E}[\{\mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]]\}^2] = \xi_n^2/n$, by Chebyshev's inequality we have $|\mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]]| \leq \xi_n/\sqrt{\delta n}$ with probability $1-\delta$.

Using Theorem 7, for all $n \geq n_0$, with probability at least $1 - \alpha\{1 + \bar{C}/\log n\} - \bar{C}\{n^{-1/2}\log n\}$ the conditions for Theorem 2 hold. These conditions and the bound on the second term in (D.1) imply that for n_0 sufficiently large $|\widehat{Q}^{1/2}(\beta_0) - \sigma| \leq \sigma/2$. Therefore, by Theorem 2 we can bound the first term in (D.1) as

$$\begin{aligned}\left|\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right| &= |\widehat{Q}^{1/2}(\widehat{\beta}) - \widehat{Q}^{1/2}(\beta_0)| \cdot |\widehat{Q}^{1/2}(\widehat{\beta}) + \widehat{Q}^{1/2}(\beta_0)| \\ &\leq 2\widehat{Q}^{1/2}(\beta_0)\tfrac{(\varrho_{\bar{c}}+\bar{\zeta})^2}{1-\bar{\zeta}^2}\left(2\widehat{Q}^{1/2}(\beta_0) + 2\widehat{Q}^{1/2}(\beta_0)\tfrac{(\varrho_{\bar{c}}+\bar{\zeta})^2}{1-\bar{\zeta}^2}\right) \\ &\leq \sigma^2\{C's\log(p/\alpha)/n\}\end{aligned}$$

where the last bound follows from Condition P.

The last term in (D.1) satisfies $\mathbb{E}_n[r^2] \leq c_s^2 \leq \sigma^2 C' s/n$ by Condition P. The third term in (D.1) can be bound using Chebyshev's inequality and that

$$\bar{\mathbf{E}}[\{\mathbb{E}_n[\epsilon r]\}^2] = \bar{\mathbf{E}}[\epsilon^2 r^2]/n \leq c_s^2 \max_{1\leq i\leq n} \mathrm{E}[\epsilon_i^2]/n \leq c_s^2 \bar{\mathbf{E}}[|\epsilon|^q]^{2/q} n^{-1+2/q}.$$

Those imply that $2\sigma|\mathbb{E}_n[r\epsilon]| \leq 2\sigma c_s n^{-1/2+1/q}/\delta^{1/2} \leq \sigma^2 \bar{C} s^{1/2} n^{-1+1/q}/\delta^{1/2}$ with probability $1-\delta$.

For the second result of the theorem, note that Conditions ASM, P and $s^2\log^2(p\vee n) \leq Cn/\log n$, implies that $\varrho_{\bar{c}} + \bar{\zeta} + c_s \leq C'\sqrt{s\log(p\vee n)/n} \leq C'' n^{-1/4}/\log n = o(n^{-1/4})$. We expand as before

$$\begin{aligned}n^{1/2}\left(\widehat{Q}(\widehat{\beta}) - \sigma^2\right) &= n^{1/2}\left(\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right) + n^{1/2}\left(\mathbb{E}_n[\sigma^2\epsilon_i^2 + 2\sigma r\epsilon + r^2] - \sigma^2\right) \\ &= n^{1/2}\left(\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right) + n^{1/2}\left(\sigma^2\mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]] + 2\sigma\mathbb{E}_n[r\epsilon] + \mathbb{E}_n[r^2]\right).\end{aligned}$$

By the first part and $\varrho_{\bar{c}} + \bar{\zeta} + c_s \leq C'\sqrt{s\log(p\vee n)/n} \leq C'' n^{-1/4}/\log n = o(n^{-1/4})$ we have

$$\begin{aligned}n^{1/2}\left|\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right| &\lesssim_P n^{1/2}|\varrho_{\bar{c}} + \bar{\zeta}|^2 = o(1) \\ n^{1/2}\mathbb{E}_n[r^2] &\lesssim_P n^{1/2}c_s^2 \lesssim n^{1/2}s/n = o(1) \\ n^{1/2}|\mathbb{E}_n[r\epsilon]| &\lesssim_P n^{1/2}c_s n^{1/q}/n^{1/2} \lesssim n^{1/2}s^{1/2}n^{-1+1/q} = o(1).\end{aligned}$$

To control the term $\sigma^2 n^{1/2}(\mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]])$, since $\bar{\mathbf{E}}[\epsilon^q] < C$, by Lemma 4 with $\delta = q/4 - 1 > 0$ we have

$$\frac{n^{1/2}\mathbb{E}_n[\epsilon^2 - \mathrm{E}[\epsilon^2]]}{\sqrt{\mathbb{E}_n[(\epsilon^2 - \mathrm{E}[\epsilon^2])^2]}} \Rightarrow N(0,1). \tag{D.2}$$

Since $\bar{\mathbf{E}}[\epsilon^q] < C$ for $q > 4$ (assume w.l.o.g $q \leq 8$), by Vonbahr-Esseen's LLN, we have for any $t > 0$

$$P(|\mathbb{E}_n[\{\epsilon^2 - \mathrm{E}[\epsilon^2]\}^2] - \xi_n^2| > t) \leq \frac{\bar{\mathbf{E}}[|\{\epsilon^2 - \mathrm{E}[\epsilon^2]\}|^{q/2}]}{t^{q/4} n^{q/4-1}} \leq \frac{4C}{t^{q/4} n^{q/4-1}} = o(1).$$

The result follows from (D.2). ∎

D.2. Proofs of Additional Results for Section 4 Concerning Penalty Loadings.

PROOF OF LEMMA 7. Let $t_n = \Phi^{-1}(1 - \alpha/2p)$ and recall we have

$$w_n = (\alpha^{-1} \log n C_q \bar{\mathbf{E}}[|\epsilon|^{q \vee 4}])^{1/q} < 1/2$$

under Condition R. Thus

(D.3)

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

Next we proceed to bound each term. We shall rely on the following lemma:

LEMMA 10. *Let $r_1,\ldots,r_n$ be fixed and assume ϵ_i are independent zero mean random variables such that $\bar{\mathbf{E}}[\epsilon^2] = 1$. Suppose that there is $q > 2$ such that $\bar{\mathbf{E}}[|\epsilon|^q] < \infty$. Then, for $u_n > 0$ we have*

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

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

First Term of (D.3). By Lemma 10 with $v = w_n$ we have that

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

Second Term of (D.3). By Lemma 6 and using that $\widehat{\gamma}_{j,k} \geq 1$,

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

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

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

Third Term of (D.3). Let $\bar{t} = \min_{1\leq j\leq p}(\mathbb{E}_n[x_j^2\mathrm{E}[\epsilon^2]])^{1/2}/(\mathbb{E}_n[|x_j^3|\mathrm{E}[|\epsilon^3|]])^{1/3} > 0$. By Lemma 4 with $\delta = 1$, since $t_n \leq \bar{t}\ n^{1/6} - 1$ by Condition R, we have that there is an universal constant A, such that

$$\begin{aligned} P\left(\max_{1\leq j\leq p}\frac{\sqrt{n}|\mathbb{E}_n[x_j\epsilon]|}{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}} > t_n\right) &\leq p\ \max_{1\leq j\leq p} P\left(\frac{\sqrt{n}|\mathbb{E}_n[x_j\epsilon]|}{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}} > t_n\right) \\ &\leq 2p\ \bar{\Phi}(t_n)\left(1+\tfrac{A}{\ell_n^3}\right) \leq \alpha\left(1+\tfrac{A}{\ell_n^3}\right) \end{aligned}$$

where the last inequality follows from the definition of t_n.
Fourth Term of (D.3). This term determines η. Let $\widehat{\Gamma}_k = \mathrm{diag}(\widehat{\gamma}_{1,k},\ldots,\widehat{\gamma}_{p,k})$. We are interested on the event

$$\sqrt{1+u_n}\widehat{\gamma}_{j,0} \geq \sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}/\sqrt{\mathbb{E}_n[\epsilon^2]} \quad \text{for all } j = 1,\ldots,p. \tag{D.4}$$

For $\widehat{\gamma}_{j,0} = \widehat{\gamma}_{j,0}^{(I)} = \max_{1\leq i\leq n}|x_{ij}|$, (D.4) follows automatically so $\eta = 0$. For the initial choice of $\widehat{\gamma}_{j,0} = \widehat{\gamma}_{j,0}^{(II)} = W(\mathbb{E}_n[x_j^4])^{1/4}$, (D.4) follows provided that $\sqrt{1+u_n}W\sqrt{\mathbb{E}_n[\epsilon^2]} \geq (\mathbb{E}_n[\epsilon^4])^{1/4}$. We bound this probability

$$\begin{aligned} P(\sqrt{1+u_n}W\sqrt{\mathbb{E}_n[\epsilon^2]} < (\mathbb{E}_n[\epsilon^4])^{1/4}) &\leq P(\mathbb{E}_n[\epsilon^4] > W^4) + P\left(\mathbb{E}_n[\epsilon^2] < \tfrac{1}{1+u_n}\right) \\ &\leq \tfrac{C_q\bar{\mathbf{E}}[|\epsilon|^{q\vee 8}]}{v^q n^{q/8}} \wedge \tfrac{2\bar{\mathbf{E}}[|\epsilon|^q]}{n^{1\wedge(q/4-1)}v^{q/4}} + \psi\left(\tfrac{u_n}{1+u_n}\right) \end{aligned}$$

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

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

$$\begin{aligned} \widehat{\gamma}_{j,k} &= 1 \vee \frac{\sqrt{\mathbb{E}_n[x_j^2(y-x'\widehat{\beta})^2]}}{\sqrt{\mathbb{E}_n[(y-x'\widehat{\beta})^2]}} \geq \frac{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}-\sqrt{\mathbb{E}_n[x_j^2\{x'(\widehat{\beta}-\beta_0)\}^2]}/\sigma-\sqrt{\mathbb{E}_n[x_j^2r^2]}/\sigma}{\sqrt{\mathbb{E}_n[\epsilon^2]}+\|\widehat{\beta}-\beta_0\|_{2,n}/\sigma+\sqrt{\mathbb{E}_n[r^2]}/\sigma} \\ &\geq \frac{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}-x_{\infty j}(\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)/\sigma}{\sqrt{\mathbb{E}_n[\epsilon^2]}+(\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)/\sigma}. \end{aligned}$$

Since $\widehat{\gamma}_{j,k} \geq 1$, it suffices to consider the case that $\mathbb{E}_n[\epsilon^2] \leq \mathbb{E}_n[x_j^2\epsilon^2]$. Therefore we have that $(1+\Delta)\gamma_j \geq \sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}/\sqrt{\mathbb{E}_n[\epsilon^2]}$ is implied by

$$\Delta \geq 2(\|\widehat{\beta}-\beta_0\|_{2,n}+c_s)x_{\infty j}/\{\sigma\sqrt{\mathbb{E}_n[\epsilon^2]}\}. \tag{D.5}$$

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 $\widehat{\gamma}_{j,k}$.

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

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

$$\begin{aligned}\widehat{\gamma}_{j,k} &\leq 1 \vee \frac{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}+u_n\sqrt{\mathbb{E}_n[\epsilon^2]}/2}{\sqrt{\mathbb{E}_n[\epsilon^2]}-u_n\sqrt{\mathbb{E}_n[\epsilon^2]}/2} \\ &\leq 1 \vee \frac{1+u_n/2}{1-u_n/2}\frac{\sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}}{\sqrt{\mathbb{E}_n[\epsilon^2]}} \leq \frac{(1+u_n/2)^2}{1-u_n/2}\widehat{\gamma}_{j,0}.\end{aligned}$$

Let $\Gamma^* = \operatorname{diag}(\gamma_1^*, \ldots, \gamma_p^*)$ where $\gamma_j^* = 1 \vee \sqrt{\mathbb{E}_n[x_j^2\epsilon^2]}/\sqrt{\mathbb{E}_n[\epsilon^2]}$, 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 $\tilde{\delta} = \Gamma^{*-1}\widehat{\Gamma}_k\delta$, note that

$$\begin{aligned}\bar{\kappa}(\widehat{\Gamma}_k) &= \min_{\|\widehat{\Gamma}_k\delta_{T^c}\|_1<\|\widehat{\Gamma}_k\delta_T\|_1}\frac{\sqrt{s}\|\delta\|_{2,n}}{\|\widehat{\Gamma}_k\delta_T\|_1-\|\widehat{\Gamma}_k\delta_{T^c}\|_1} \\ &= \min_{\|\Gamma^*\tilde{\delta}_{T^c}\|_1<\|\Gamma^*\tilde{\delta}_T\|_1}\frac{\sqrt{s}\|\widehat{\Gamma}_k^{-1}\Gamma^*\delta\|_{2,n}}{\|\Gamma^*\tilde{\delta}_T\|_1-\|\Gamma^*\tilde{\delta}_{T^c}\|_1} \\ &\geq \bar{\kappa}(\Gamma^*)/\|(\widehat{\Gamma}_k^{-1}\Gamma^*)^{-1}\|_\infty.\end{aligned} \tag{D.6}$$

Thus by Theorem 1 we have that the estimator with $\widehat{\beta}$ based on $\widehat{\Gamma}_k$, $k = 1$, also satisfies (D.5) by the extra condition assumed in the theorem. Thus the same argument established $k > 1$. ∎

PROOF OF LEMMA 10. Let $c_s = (\mathbb{E}_n[r^2])^{1/2}$ and $a_n = 1 - [1/(1+u_n)] = u_n/(1+u_n)$. We have that

$$P(\mathbb{E}_n[\sigma^2\epsilon^2] > (1+u_n)\mathbb{E}_n[(\sigma\epsilon+r)^2]) = P(2\mathbb{E}_n[\epsilon r] < -c_s^2 - a_n\mathbb{E}_n[\sigma^2\epsilon^2]). \tag{D.7}$$

By Lemma 12 we have

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

Thus,

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

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

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

By Chebyshev inequality we have

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

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

Further, we have

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

■

D.3. Proofs of Additional Results for Section 4 Concerning Performance in Sobolev Spaces.

Proof of Lemma 8. The proof used the bound (D.9) in the proof of Lemma 9. With probability $1-o(1)$ $\mathbb{E}_n[(f-\sum_{j=1}^k \theta_j P_j(z))^2] \leq \bar{K}k^{-2\alpha}$ for any $k \leq \sqrt{n/\log n} \wedge n^{1-\frac{1}{2\alpha}}$, and also $\max_{1\leq j\neq j'\leq n}|\mathbb{E}_n[P_j(z)P_{j'}(z)]| \leq C\sqrt{\log n/n}$. Therefore, for $k \leq \sqrt{n/\log n}$, with probability $1-o(1)$ we have

$$\begin{aligned} \mathbb{E}_n[(f-\widehat{f}^{(k)})^2] &\leq 2\mathbb{E}_n[(\textstyle\sum_{j=1}^k(\theta_j-\widehat{\theta}_j P_j(z))^2] + 2\mathbb{E}_n[(f-\sum_{j=1}^k\theta_j P_j(z))^2] \\ &\leq 2\mathbb{E}_n[\textstyle\sum_{j=1}^k(\theta_j-\widehat{\theta}_j)^2]\max_{j,i}|P_j(z_i)|^2 \\ &\quad + \left(\textstyle\sum_{j=1}^k|\widehat{\theta}_j-\theta_j|\right)^2 C\sqrt{\log n/n} + 2\bar{K}k^{-2\alpha} \end{aligned}$$

By Markov's inequality and Lemma 11 we have

$$\mathbb{E}_n[\sum_{j=1}^k(\theta_j-\widehat{\theta}_j)^2] \lesssim_P k^{-2\alpha} + \frac{k}{n}.$$

■

Lemma 11. *Consider the nonparametric model (2.1) where $f:[0,1]\to\mathbb{R}$ belongs to the Sobolev space $\mathcal{S}(\alpha,L)$ with $\alpha \leq 1$ and $L>0$, and $z_i \sim$ Uniform$(0,1)$, independent across $i=1,\ldots,n$. Given a bounded orthonormal basis $\{P_j(\cdot)\}_{j=1}^\infty$, the coefficients of the projection estimator satisfy for any $k \leq n$*

$$\mathrm{E}[\|\widehat{\theta}^{(k)}-\theta\|^2|z_1,\ldots,z_n] \lesssim_P k^{-2\alpha} + \frac{k}{n}$$

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

PROOF OF LEMMA 8. Let $Z = [z_1, \dots, z_n]$ 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 [50] we have $\mathrm{E}[\widehat{\theta}_j|Z] = \theta_j + \gamma_j$, where $\gamma_j = \mathbb{E}_n[f(z)P_j(z)] - \theta_j$, and $\mathrm{E}[(\widehat{\theta}_j - \theta_j)^2|Z] = \mathbb{E}_n[P_j(z)^2]\sigma^2/n + \gamma_j^2$.

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

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

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

(D.8)
$$\begin{aligned}&\textstyle\sum_{m=1}^{\bar{k}}|\theta_m| \leq (\sum_{m=1}^{\bar{k}} m^{2\alpha}\theta_m^2)^{1/2}(\sum_{m=1}^{\bar{k}} m^{-2\alpha})^{1/2} \leq C_{\alpha,L}L^{1/2}, \\ &\textstyle\sum_{m=\bar{k}}^{\infty}|\theta_m| \leq (\sum_{m=1}^{\infty} m^{2\alpha}\theta_m^2)^{1/2}(\sum_{m=\bar{k}}^{\infty} m^{-2\alpha})^{1/2} \leq C_{\alpha,L}L^{1/2}\bar{k}^{-\alpha+1/2}.\end{aligned}$$

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

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

Indeed, note that since the basis is bounded and (D.8) holds, we have

$$|\theta_M' P_M(z_i)P_j(z_i) - \theta_j| \lesssim \|\theta_M\|_1 \lesssim 1,$$

and thus $Z_j := \mathbb{E}_n[\theta_M' P_M(z)P_j(z) - \theta_j]$ satisfies $E_Z[Z_j] = 0$ and $E_Z[Z_j^2] \lesssim 1/n$. Hence, by Markov's inequality we have

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

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

$$\begin{aligned}\textstyle\sum_{j=1}^{k}\mathbb{E}_n[P_j(z)^2]\frac{\sigma^2}{n} &\lesssim \max_{1\leq j\leq k}\mathbb{E}_n[P_j(z)^2]\frac{\sigma^2 k}{n} \lesssim \sigma^2 n^{-1+1/[2\alpha+1]} \lesssim n^{-2\alpha/[2\alpha+1]}, \\ \textstyle\sum_{m=k+1}^{\infty}\theta_j^2 &\lesssim k^{-2\alpha} \lesssim n^{-2\alpha/[2\alpha+1]}, \\ \textstyle\sum_{m=1}^{k}\gamma_m^2 &\lesssim_P \frac{k}{n} + kn^{-2\alpha+1} \lesssim n^{-2\alpha/[2\alpha+1]}\end{aligned}$$

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

$$\begin{aligned}\mathrm{E}[\|\widehat{\theta}^{(k)} - \theta\|^2|Z] &\lesssim \textstyle\sum_{j=1}^{k}\mathbb{E}_n[P_j(z)^2]\frac{\sigma^2}{n} + \sum_{m=1}^{k}\gamma_m^2 + \sum_{m\geq k+1}\theta_m^2 \\ &\lesssim_P \tfrac{k}{n} + k^{-2\alpha}\end{aligned}$$

by the relations above.

■

PROOF OF LEMMA 9. For the sake of exposition, without loss of generality, assume that the components are already rearranged. Note that

$$\begin{aligned} c_k^2 &\leq \mathbb{E}_n[(f - \textstyle\sum_{j=1}^k \theta_j P_j(z))^2] \\ &= \mathbb{E}_n[(\textstyle\sum_{j=k+1}^\infty \theta_j P_j(z))^2] \\ &\leq 2\mathbb{E}_n[(\textstyle\sum_{j=k+1}^n \theta_j P_j(z))^2] + 2\mathbb{E}_n[(\textstyle\sum_{j=n+1}^\infty \theta_j P_j(z))^2] \\ &\leq 2\textstyle\sum_{j=k+1}^n \theta_j^2 \mathbb{E}_n[P_j(z)^2] + 2\left(\textstyle\sum_{j=k+1}^n |\theta_j|\right)^2 \max_{1\leq j\neq j'\leq n} |\mathbb{E}_n[P_j(z)P_{j'}(z)]| \\ &\quad +2\left(\textstyle\sum_{j=n+1}^\infty |\theta_j|\right)^2 \max_{j,z} |P_j(z)|^2 \end{aligned}$$

Since the basis is bounded, $\max_j \mathbb{E}_n[P_j(z)^2] \leq K_1$ and $\max_{j,z} |P_j(z)|^2 \leq K_2$. Because the basis is bounded and orthonormal, $\mathrm{E}[P_j(z)P_{j'}(z)] = 0$ for $j \neq j'$, for some constant K_3, with probability $1 - o(1)$, by Lemma 19 in [4], we have

$$\max_{1\leq j,j'\leq n} \left|\mathbb{E}_n[P_j(z)P_{j'}(z)]\right| \leq K_3\sqrt{\frac{\log n}{n}}.$$

Also, by $f \in \mathcal{RS}(\alpha, p, L)$ we have

$$\sum_{j=k+1}^n \theta_j^2 \mathbb{E}_n[P_j(z)^2] \leq K_1 \sum_{j=k+1}^n \theta_j^2 \leq K_1 k^{-2\alpha} \sum_{j=k+1}^n j^{2\alpha}\theta_j^2 \leq K_1 L^2 k^{-2\alpha}$$

$$\sum_{j=k+1}^\infty |\theta_j| \leq \{\sum_{j=k+1}^\infty j^{2\alpha}\theta_j^2\}^{1/2}\{\sum_{j=k+1}^\infty j^{-2\alpha}\}^{1/2} \leq Lk^{-\alpha+1/2}/\sqrt{2\alpha}$$

Therefore, with probability at least $1 - o(1)$

$$c_k^2 \leq K_1 L^2 k^{-2\alpha} + 2L^2 k^{-2\alpha+1}\sqrt{\log n/n} + 2L^2 n^{-2\alpha+1}$$

Consider the set of indices $\mathcal{I} = \{k \in \mathbf{N} : k\sqrt{\log n/n} \leq 1, \quad k^{2\alpha} \leq n^{2\alpha-1}\}$. For all $k \in \mathcal{I}$, with probability $1 - o(1)$ we have

$$c_k^2 \leq \mathbb{E}_n[(f - \textstyle\sum_{j=1}^k \theta_j P_j(z))^2] \leq \bar{K}k^{-2\alpha}. \tag{D.9}$$

Note that for any constant C, $k_* = Cn^{1/\{1+2\alpha\}} \in \mathcal{I}$ for n sufficiently large since $\alpha \geq 1$. Thus,

$$c_s^2 + s/n \leq \bar{K}k_*^{-2\alpha} + k_*/n \leq C' n^{-2\alpha/\{1+2\alpha\}}.$$

In particular, this implies $s \leq C' n^{1/\{1+2\alpha\}}$. ■

APPENDIX E: VARIOUS TECHNICAL LEMMAS

E.1. Lemmas bounding various empirical moments of ϵ_i.

LEMMA 12. *Let ϵ_i, $i = 1, \dots, n$, be independent random variables such that $\bar{\mathbf{E}}[\epsilon^2] = 1$. Assume that there is $q > 2$ such that $\bar{\mathbf{E}}[|\epsilon|^q] < \infty$. Then there is a constant C_q, that depends on q only, such that for $v > 0$ we have*

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

PROOF OF LEMMA 12. This follows by by the application of either Rosenthal's inequality [47] for the case of $q > 4$ or Vonbahr-Esseen's inequalities [58] for the case of $2 < q \leq 4$, and taking the best bound. ∎

LEMMA 13. *Consider $\epsilon_i \sim t(2)$. Then, for $\tau \in (0, 1)$ we have that:*
(i) $P(\mathbb{E}_n[\epsilon^2] \geq 2\sqrt{2}/\tau + \log(4n/\tau)) \leq \tau/2$.
(ii) For $0 < a < 1/6$, we have $P(\mathbb{E}_n[\epsilon^2] \leq a \log n) \leq \frac{1}{n^{1/2(1/6-a)^2}}$.
(iii) For $u_n \geq 0$ and $0 < a < 1/6$, we have

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

PROOF OF LEMMA 13. To show (i) we will establish a bound on $q(\mathbb{E}_n[\epsilon^2], 1-\tau)$. 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

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

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

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

Thus, by setting $t_n = 4n/\tau$ and $t = 2\sqrt{2}/\tau$ we have by [27], relation (7.5),

$$\begin{aligned} P(|\mathbb{E}_n[\epsilon^2] - \mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \le t_n\}]| \ge t) &\le \tfrac{\mathrm{E}[\epsilon_i^4 1\{\epsilon_i^2 \le t_n\}]}{nt^2} + nP(|\epsilon_i^2| > t_n) \\ &\le \tfrac{t_n}{nt^2} + \tfrac{2n}{2+t_n} \le \tau/2. \end{aligned} \tag{E.3}$$

Thus, (i) is established.

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

$$\begin{aligned} P(\mathbb{E}_n[\epsilon^2] \le a \log n) &\le P(\mathbb{E}_n[\epsilon^2 1\{\epsilon^2 \le n^{1/2}\}] \le a \log n) \\ &\le P(|\mathbb{E}_n[\epsilon^2 1\{\epsilon^2 \le n^{1/2}\}] - \mathrm{E}[\epsilon_i^2 1\{\epsilon_i^2 \le n^{1/2}\}]| \ge (\tfrac{1}{6} - a) \log n) \\ &\le \tfrac{1}{n^{1/2}(1/6-a)^2} \end{aligned} \tag{E.4}$$

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

To show (iii), let $a_n = [(1+u_n)^2 - 1]/(1+u_n)^2 = u_n(2+u_n)/(1+u_n)^2 \ge u_n/(1+u_n)$ and note that by (E.1), (E.3), and (E.4) we have

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

■

APPENDIX F: PROBABILITY INEQUALITIES USED

F.1. Moment Inequalities.

LEMMA 14 (Rosenthal Inequality). *Let $X_1, \ldots, X_n$ be independent zero-mean random variables, then for $r \ge 2$*

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

COROLLARY 3 (Rosenthal LLN). *Let $r \ge 2$, 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 $\ell_n > 0$*

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

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

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

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

so the corollary follows. We refer to [47] for proofs.

LEMMA 15 (Vonbahr-Esseen inequality). *Let $X_1, \ldots, X_n$ be independent zero-mean random variables. Then for $1 \leq r \leq 2$*

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

We refer to [58] for proofs.

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

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

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

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

which implies the corollary.

F.2. Moderate Deviations for Self-Normalized Sums. We shall be using the following result based on Theorem 7.4 in [24]. (This is restated here for completeness.)

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, and $\delta \in (0, 1]$. Let*

$$S_{n,n} = \sum_{i=1}^{n} X_{i,n}, \quad V_{n,n}^2 = \sum_{i=1}^{n} X_{i,n}^2 \quad \text{and} \quad M_n = \frac{(\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}X_{i,n}^2)^{1/2}}{(\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}|X_{i,n}|^{2+\delta})^{1/\{2+\delta\}}} > 0.$$

Suppose that for some $\ell_n \to \infty$ such that $n^{\frac{\delta}{2(2+\delta)}} M_n/\ell_n \geq 1$. Then, for some absolute constant A, uniformly on $0 \leq x \leq n^{\frac{\delta}{2(2+\delta)}} M_n/\ell_n - 1$, we have

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

APPENDIX G: MONTE-CARLO PERFORMANCE OF $\sqrt{\text{LASSO}}$

G.1. Estimation performance of $\sqrt{\text{Lasso}}$, homoscedastic case. 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 homoscedastic 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

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

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 $\Sigma_{jk} = (1/2)^{|j-k|}$.

We present the results of computational experiments for designs a) and b) in Figures 1, 2, 3. The left plot of each figure reports the results for the normal errors, and the right plot of each figure reports the results for $t(4)$ errors. For each model, the figures show the following quantities as a function of 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, $\mathrm{E}[|\text{support}(\widetilde{\beta})|]$.

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 $\sqrt{\text{Lasso}}$ perform similarly in terms of empirical risk, although standard Lasso outperforms somewhat $\sqrt{\text{Lasso}}$. At the same time, ols post Lasso outperforms slightly ols post $\sqrt{\text{Lasso}}$ for larger signal strengths. This is expected since $\sqrt{\text{Lasso}}$ over regularize to simultaneously estimate σ when compared to Lasso (since it essentially uses $\sqrt{\widehat{Q}(\widehat{\beta})}$ as an estimate of σ). In the nonparametric model

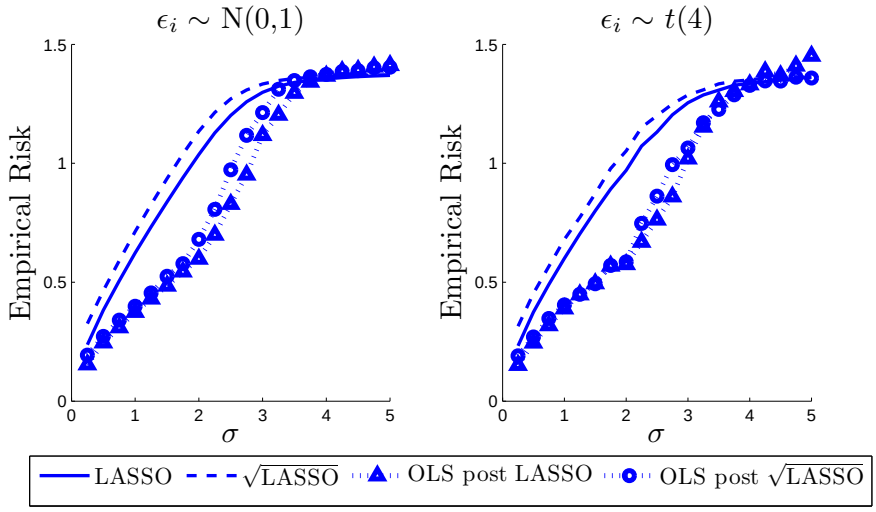

FIG 1. *The average empirical risk of the estimators as a function of the scaling parameter* σ.

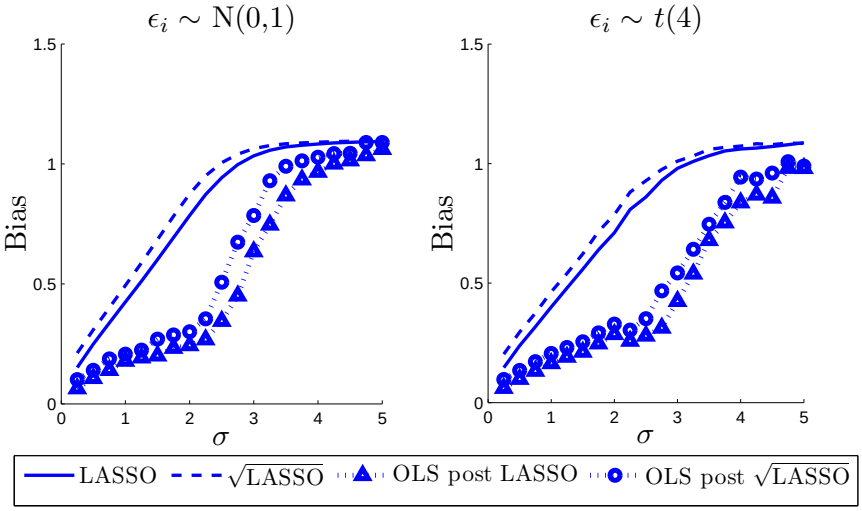

FIG 2. *The norm of the bias of the estimators as a function of the scaling parameter* σ.

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 $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ achieve its goal. Despite not knowing σ, $\sqrt{\text{Lasso}}$ performs comparably to the standard Lasso that knows σ. These results are in close agreement with our

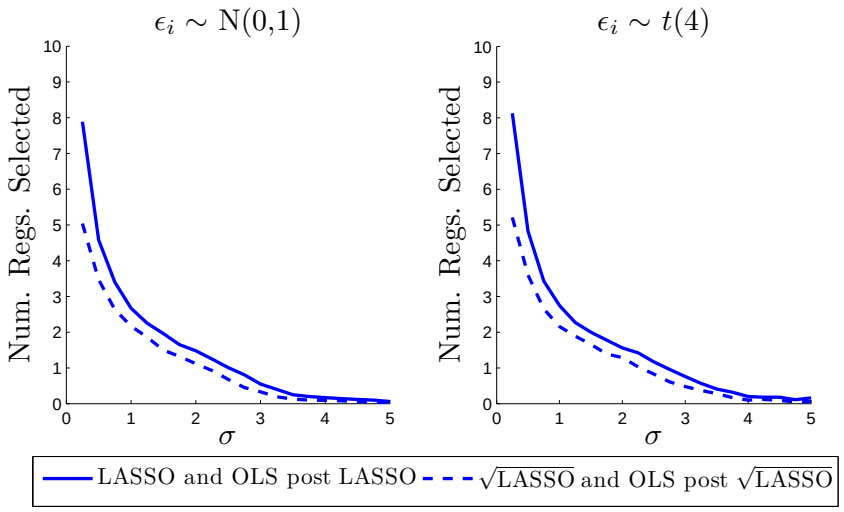

FIG 3. *The average number of regressors selected as a function of the scaling parameter* σ.

theoretical results, which state that the upper bounds on empirical risk for $\sqrt{\text{Lasso}}$ 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 $\sqrt{\text{Lasso}}$ arise primarily due to $\sqrt{\text{Lasso}}$ having a larger bias than standard Lasso. This bias arises because $\sqrt{\text{Lasso}}$ uses an effectively heavier penalty. Figure 3 shows that such heavier penalty translates into $\sqrt{\text{Lasso}}$ 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 $\sqrt{\text{Lasso}}$ under $t(4)$ errors nearly coincides with their performance under Gaussian errors. This is exactly what is predicted by our theoretical results.

G.2. Estimation performance of $\sqrt{\text{Lasso}}$, heteroscedastic case. In this section we use Monte carlo experiments to assess the finite-sample performance under heteroscedastic errors of the following estimators:

- the (infeasible) oracle estimator,
- heteroscedastic $\sqrt{\text{Lasso}}$ (as Algorithm 1),
- ols post heteroscedastic $\sqrt{\text{Lasso}}$, which applies ols to the model selected by heteroscedastic $\sqrt{\text{Lasso}}$.
- the (infeasible) ideal heteroscedastic $\sqrt{\text{Lasso}}$ (which uses exact load-

ings),
- ols post ideal heteroscedastic $\sqrt{\text{Lasso}}$, which applies ols to the model selected by ideal heteroscedastic $\sqrt{\text{Lasso}}$.

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

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

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

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

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 heteroscedastic $\sqrt{\text{Lasso}}$ 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 heteroscedastic $\sqrt{\text{Lasso}}$ performs very close to its ideal counterpart and to the oracle.

APPENDIX H: 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.

H.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

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

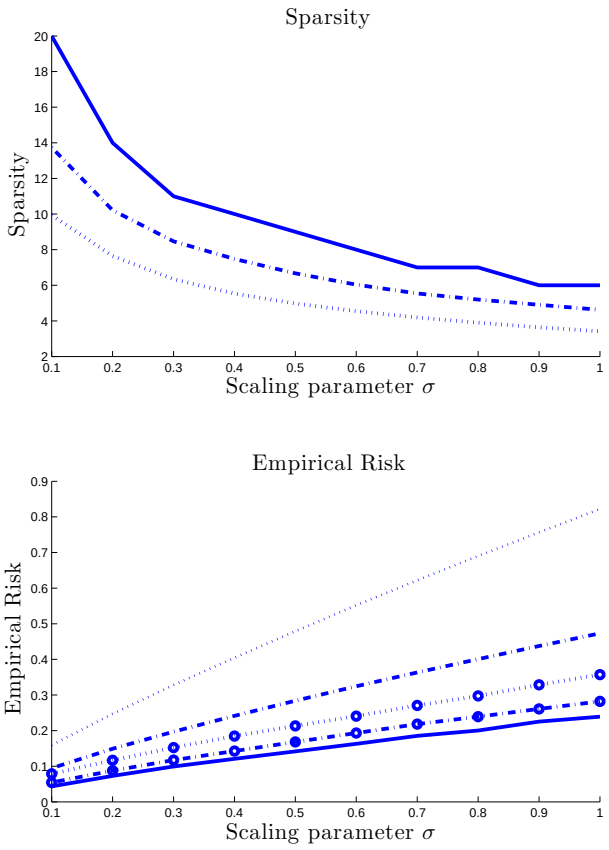

FIG 4. *For each estimator the top figure displays the corresponding sparsity and the bottom figure displays the empirical risk as a function of the scaling parameter σ. The solid line corresponds to the oracle estimator, the dotted line corresponds to the heteroscedastic $\sqrt{Lasso}$, the dashed-dot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

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[5] in several fields including optimal control, learning theory, eigenvalue op-

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

timization, 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 [43, 45], and efficient software implementations [51] 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 [41, 42, 39].

It follows that (2.5) 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 (2.5) as the following conic program:

$$\begin{array}{ll} \min\limits_{t,v,\beta^+,\beta^-} & \dfrac{t}{\sqrt{n}} + \dfrac{\lambda}{n}\displaystyle\sum_{i=1}^{p}\left(\gamma_j\beta_j^+ + \gamma_j\beta_j^-\right) \\ & v_i = y_i - x_i'\beta^+ + x_i'\beta^-,\ i = 1, \ldots, n \\ & (t, v) \in Q^{n+1},\ \beta^+ \geq 0,\ \beta^- \geq 0. \end{array} \tag{H.1}$$

Conic duality immediately yields the following dual problem

$$\begin{array}{ll} \max\limits_{a \in \mathbb{R}^n} & \mathbb{E}_n[ya] \\ & |\mathbb{E}_n[x_j a]| \leq \lambda\gamma_j/n,\ j = 1, \ldots, p \\ & \|a\| \leq \sqrt{n}. \end{array} \tag{H.2}$$

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 different methods to compute $\sqrt{\text{Lasso}}$ with their counterparts to compute Lasso. We note that these methods have different initialization and stopping

$n = 100$, $p = 500$	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.

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 H.1 illustrates that the average computational time to solve Lasso and $\sqrt{\text{Lasso}}$ optimization problems are comparable. Table H.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.

H.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,

$$\min_w c'w \ : \ Aw = b, w \in K. \tag{H.3}$$

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 (H.3).

In order to formulate the optimization problem associated with the Lasso estimator as a conic programming problem (H.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

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

Thus, we have that Lasso optimization problem can be cast

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

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

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

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 [2]. 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

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

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 $\beta_j^+ = \beta_j^k + 2\mathbb{E}_n[x_j(y - x'\beta^k)]/\mu$.

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:

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

Under standard normalization assumptions $\mathbb{E}_n[x_j^2] = 1$ 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 $2\mathbb{E}_n[x_j(y - x'\beta_{-j})] > \lambda\gamma_j/n$ it follows that the optimal choice for β_j is
$$\beta_j = \left(-2\mathbb{E}_n[x_j(y - x'\beta_{-j})] + \lambda\gamma_j/n\right) / \mathbb{E}_n[x_j^2];$$
- if $2\mathbb{E}_n[x_j(y - x'\beta_{-j})] < -\lambda\gamma_j/n$ it follows that the optimal choice for β_j is
$$\beta_j = \left(-2\mathbb{E}_n[x_j(y - x'\beta_{-j})] - \lambda\gamma_j/n\right) / \mathbb{E}_n[x_j^2];$$
- if $2|\mathbb{E}_n[x_j(y - x'\beta_{-j})]| \leq \lambda\gamma_j/n$ 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:

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

As before, under standard normalization assumptions $\mathbb{E}_n[x_j^2] = 1$ 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 $\mathbb{E}_n[x_j(y - x'\beta_{-j})] > (\lambda/n)\gamma_j\sqrt{\widehat{Q}(\beta_{-j})}$, we have

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

- if $\mathbb{E}_n[x_j(y - x'\beta_{-j})] < -(\lambda/n)\gamma_j\sqrt{\widehat{Q}(\beta_{-j})}$, we have

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

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

REFERENCES

[1] Takeshi Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pages 955–968, 1977.

[2] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *ArXiv*, 2010.

[3] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012. Arxiv, 2010.

[4] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[5] A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. *Inverse problems and high dimensional estimation - Stats in the Château summer school in econometrics and statistics, 2009, Springer Lecture Notes in Statistics - Proceedings*, pages 121–156, 2011.

[6] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. Arxiv, 2009.

[7] A. Belloni, V. Chernozhukov, I. Fernandez-Val, and C. Hansen. Program evaluation with high-dimensional data. *arXiv:1311.2645*, 2013.

[8] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. ArXiv, 2010.

[9] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011. forthcoming, The Review of Economic Studies.

[10] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2013. ArXiv, 2011.

[11] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for lad regression models. *ArXiv*, 2013.

[12] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. Arxiv, 2010.

[13] A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *ArXiv:1304.3969*, 2013.

[14] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
[15] P. Bülhmann and S. van de Geer. *Statistics for High-Dimensional Data*. Statistics (New York). Springer-Verlag, Berlin, 2011. Methods, Theory and Applications.
[16] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
[17] 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.
[18] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.
[19] E. Candès and Y. Plan. Near-ideal model selection by l_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.
[20] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.
[21] Gary Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596, 1992.
[22] Y. Chen and A. S. Dalalyan. Fused sparsity and robust estimation for linear models with unknown variance. *Advances in Neural Information Processing Systems*, 25:1268–1276, 2012.
[23] S. Chrétien and S. Darses. Sparse recovery with unknown variance: a LASSO-type approach. *arXiv:1101.04334*, 2012.
[24] V. H. de la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.
[25] 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.
[26] M. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *University of Michigan Department of Economics Working Paper*, 2013.
[27] W. Feller. *An Introduction to Probability Theory and Its Applicaitons, Volume II*. Wiley Series in Probability and Mathematical Statistics, 1966.
[28] E. Gautier and A. Tsybakov. High-dimensional instrumental variables rergession and confidence sets. arXiv:1105.2454v2 [math.ST], 2011.
[29] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *arXiv:1109.5587v2*, 2012.
[30] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.
[31] Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.
[32] B.-Y. Jing, Q.-M. Shao, and Q. Wang. Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.
[33] O. Klopp. High dimensional matrix estimation with unknown variance of the noise. *arXiv*, (arXiv:1112.3055), 2011.
[34] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.
[35] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.
[36] M. C. Veraar L. Duembgen, S. A. van de Geer and J. A. Wellner. Nemirovski's inequalities revisited. *American Mathematical Monthly*, 117:138–160, 2010.

[37] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
[38] 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.
[39] Z. Lu. Gradient based method for cone programming with application to large-scale compressed sensing. *Technical Report*, 2008.
[40] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
[41] Y. Nesterov. Smooth minimization of non-smooth functions, mathematical programming. *Mathematical Programming*, 103(1):127–152, 2005.
[42] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.
[43] Y. Nesterov and A Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1993.
[44] Zhao R., T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.
[45] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2001.
[46] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
[47] 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.
[48] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19:209–285, 2010.
[49] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.
[50] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.
[51] 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.
[52] S. A. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.
[53] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
[54] 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.
[55] S. A. van de Geer, P. Bühlmann, and Y. Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *ArXiv*, 2013.
[56] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
[57] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.
[58] B. von Bahr and C.-G. Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist*, 36:299–303, 1965.
[59] 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.
[60] L. Wang. L1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135151, September 2013.
[61] 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.

[62] C.-H. Zhang and S. S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.