Bernoulli **28**(2), 2022, 713–743
https://doi.org/10.3150/21-BEJ1348

De-biasing the lasso with degrees-of-freedom adjustment

PIERRE C. BELLEC[*] and CUN-HUI ZHANG[†]

Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, USA.
E-mail: [*] *pierre.bellec@rutgers.edu;* [†] *czhang@stat.rutgers.edu*

This paper studies schemes to de-bias the Lasso in sparse linear regression with Gaussian design where the goal is to estimate and construct confidence intervals for a low-dimensional projection of the unknown coefficient vector in a preconceived direction $\boldsymbol{a}_0$. Our analysis reveals that previously analyzed propositions to de-bias the Lasso require a modification in order to enjoy nominal coverage and asymptotic efficiency in a full range of the level of sparsity. This modification takes the form of a degrees-of-freedom adjustment that accounts for the dimension of the model selected by the Lasso. The degrees-of-freedom adjustment (a) preserves the success of de-biasing methodologies in regimes where previous proposals were successful, and (b) repairs the nominal coverage and provides efficiency in regimes where previous proposals produce spurious inferences and provably fail to achieve the nominal coverage. Hence our theoretical and simulation results call for the implementation of this degrees-of-freedom adjustment in de-biasing methodologies.

Let s_0 denote the number of nonzero coefficients of the true coefficient vector and $\boldsymbol{\Sigma}$ the population Gram matrix. The unadjusted de-biasing scheme may fail to achieve the nominal coverage as soon as $s_0 \ggg n^{2/3}$ if $\boldsymbol{\Sigma}$ is known. If $\boldsymbol{\Sigma}$ is unknown, the degrees-of-freedom adjustment grants efficiency for the contrast in a general direction $\boldsymbol{a}_0$ when

$$\frac{s_0 \log p}{n} + \min\left\{\frac{s_\Omega \log p}{n}, \frac{\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1\sqrt{\log p}}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2\sqrt{n}}\right\} + \frac{\min(s_\Omega, s_0)\log p}{\sqrt{n}} \to 0$$

where $s_\Omega = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_0$. The dependence in s_0, s_Ω and $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1$ is optimal and closes a gap in previous upper and lower bounds. Our construction of the estimated score vector provides a novel methodology to handle dense directions $\boldsymbol{a}_0$.

Beyond the degrees-of-freedom adjustment, our proof techniques yield a sharp ℓ_∞ error bound for the Lasso which is of independent interest.

Keywords: Statistical inference; Lasso; semiparametric model; Fisher information; efficiency; confidence interval; p-value; regression; high-dimensional data

1. Introduction

Consider a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

with a sparse coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$, a Gaussian noise vector $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, and a Gaussian design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows. The purpose of this paper is to study the sample size requirement in de-biasing the Lasso for regular statistical inference of a linear contrast

$$\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle \tag{1.2}$$

at the $n^{-1/2}$ rate in the case of $p \gg n$ for both known and unknown $\boldsymbol{\Sigma}$. As a consequence of regularity, the $n^{-1/2}$ rate also corresponds to the length of confidence intervals for θ.

The problem was considered in [29] in a general semi-low-dimensional (LD) approach where high-dimensional (HD) models are decomposed as

$$\text{HD model} = \text{LD component} + \text{HD component} \tag{1.3}$$

in the same fashion as in semi-parametric inference [6]. For the estimation of a real function $\theta = \theta(\boldsymbol{\beta})$ of a HD unknown parameter $\boldsymbol{\beta}$, the decomposition in (1.3) was written in the vicinity of a given $\boldsymbol{\beta}_0$ as

$$\boldsymbol{\beta} - \boldsymbol{\beta}_0 = \boldsymbol{u}_0\big(\theta - \theta_0\big) + \boldsymbol{Q}_0\big(\boldsymbol{\beta} - \boldsymbol{\beta}_0\big), \tag{1.4}$$

where $\boldsymbol{u}_0$ specifies the least favorable one-dimensional local sub-model giving the minimum Fisher information for the estimation of θ, subject to $\langle \boldsymbol{u}_0, \nabla\theta(\boldsymbol{\beta}_0)\rangle = 1$, and $\boldsymbol{Q}_0 = \boldsymbol{I}_{p\times p} - \boldsymbol{u}_0(\nabla\theta(\boldsymbol{\beta}_0))^\top$ projects $\boldsymbol{\beta} - \boldsymbol{\beta}_0$ to a space of nuisance parameters. [29] went on to propose a low-dimensional projection estimator (LDPE) as a one-step maximum likelihood correction of an initial estimator $\hat{\boldsymbol{\beta}}^{(\text{init})}$ in the direction of the least favorable one-dimensional sub-model,

$$\widehat{\theta} = \theta\big(\hat{\boldsymbol{\beta}}^{(\text{init})}\big) + \operatorname*{arg\,max}_{\phi\in\mathbb{R}} \text{ log-likelihood}\big(\hat{\boldsymbol{\beta}}^{(\text{init})} + \boldsymbol{u}_0\phi\big), \tag{1.5}$$

and stated without proof that the asymptotic variance of such a one-step estimator achieves the lower bound given by the reciprocal of the Fisher information.

For the estimation of a contrast (1.2) in linear regression (1.1), we have $\nabla\theta(\boldsymbol{\beta}_0) = \boldsymbol{a}_0$, the Fischer information in the one dimension sub-model $\{\boldsymbol{\beta} + \phi\boldsymbol{u}, \phi\in\mathbb{R}\}$ is $\langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle\sigma^{-2}$, the least favorable sub-model is given by

$$\boldsymbol{u}_0 = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0}{\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle}, \quad \text{i.e., the minimizer} \quad \boldsymbol{u}_0 = \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^p:\langle \boldsymbol{u},\boldsymbol{a}_0\rangle=1} \frac{\langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle}{\sigma^2} \tag{1.6}$$

and the Fisher information for the estimation of θ is

$$F_\theta = 1\Big/\Big(\sigma^2\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle\Big). \tag{1.7}$$

In the linear model (1.1), the log-likelood function is $\boldsymbol{b}\to -\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2\sigma^2)$ up to a constant term and the one-step log-likelihood correction (1.5) can be explicitly written as a linear bias correction,

$$\widehat{\theta} = \Big\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{init})}\Big\rangle + \frac{\Big\langle z_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{init})}\Big\rangle}{\|z_0\|_2^2} \quad \text{with} \quad z_0 = \boldsymbol{X}\boldsymbol{u}_0. \tag{1.8}$$

Here, $z_0 = \boldsymbol{X}\boldsymbol{u}_0$ can be viewed as an efficient score vector for the estimation of θ.

In the case of unknown $\boldsymbol{\Sigma}$, the efficient score vector z_0 has to be estimated from the data. For statistical inference of a preconceived regression coefficient β_j or a linear combination of a small number of β_j, such one-step linear bias correction was considered in [4,8,13,15,24,31] among others. The focus of the present paper is to find sharper sample size requirements, in the case of Gaussian design, than the typical $n \gg (s_0\log p)^2$ required in the aforementioned previous studies. Here and in the sequel,

$$s_0 = \big|S\big| \ \text{ with } \ S = \operatorname{supp}(\boldsymbol{\beta}). \tag{1.9}$$

Our results study both known $\boldsymbol{\Sigma}$—in that case the ideal score vector z_0 can be used—and unknown $\boldsymbol{\Sigma}$ where estimated score vectors $\widehat{z} \approx z_0$ are used. The results of [10] show that for unknown $\boldsymbol{\Sigma}$ with

bounded condition number, it is impossible to construct confidence intervals for $\theta = \boldsymbol{a}_0^\top \boldsymbol{\beta}$ with length of order $n^{-1/2}\|\boldsymbol{a}_0\|_2$ in the sparsity regime $s_0 \ggg \sqrt{n}$. Proposition 4.2 in [15] extends the lower bound from [10] to account for the sparsity and ℓ_1 norm of $\boldsymbol{u}_0$ in (1.6) as follows. Let $\Theta(s_0, s_\Omega, \rho)$ be the collection of all pairs $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ such that $\lambda_{\min}(\boldsymbol{\Sigma})^{-1} \vee \lambda_{\max}(\boldsymbol{\Sigma}) \le c_0$ for some absolute constant $c_0 > 1$ and

$$\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0 \le s_\Omega, \qquad \|\boldsymbol{\beta}\|_0 \le s_0, \qquad \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1 \le 1.02 \vee \rho.$$

When $\boldsymbol{a}_0 = \boldsymbol{e}_j$ for a fixed canonical basis vector and $s_0 \le c_1 \min(p^{0.49}, n/\log p)$,

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\Sigma})\in\Theta(s_0,s_\Omega,\rho)} \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\Sigma}}\left[n^{1/2}\sigma^{-1}|\hat{\beta}_j - \beta_j|\right] \ge c_2 + c_3 r_n(s_0, s_\Omega, \rho) \tag{1.10}$$

$$\text{with } r_n(s_0, s_\Omega, \rho) = \min\left\{\min(s_0, s_\Omega)\log(p)n^{-1/2}, (\rho \vee 1.02)\sqrt{\log p}\right\}$$

for any estimator $\hat{\beta}_j$ as a measurable function of $(\boldsymbol{y}, \boldsymbol{X})$, where $c_1, c_2, c_3 > 0$ are absolute constants.

Hence the minimax rate of estimation of β_j over $\Theta(s_0, s_\Omega, \rho)$ is at least $\sigma n^{-1/2}(1 + r_n(s_0, s_\Omega, \rho))$, and any $(1-\alpha)$-confidence interval[1] for β_j valid uniformly over $\Theta(s_0, s_\Omega, \rho)$ must incur a length of order $\sigma n^{-1/2}(1 + r_n(s_0, s_\Omega, \rho))$ up to a constant depending on α. Since the focus of the present paper is on efficiency results and other phenomena for sparsity $s_0 \ggg \sqrt{n}$, these impossibility results from [10,15] motivate either the known $\boldsymbol{\Sigma}$ assumption (in Sections 2.1 and 3 below) or the sparsity assumptions on $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ for unknown $\boldsymbol{\Sigma}$ in Section 2.2 where we prove that the lower bound (1.10) is sharp. For known $\boldsymbol{\Sigma}$, our analysis reveals that the de-biasing scheme (1.8) needs to be modified to enjoy efficiency in the regime $s_0 \ggg n^{2/3}$ when the initial estimator is the Lasso. For unknown $\boldsymbol{\Sigma}$, the modification of (1.8) is also required for efficiency when s_0, s_Ω satisfy the conditions in Theorem 2.6 of Section 2.2.

The required modification of (1.8) takes the form of a multiplicative adjustment to account for the degrees-of-freedom of the initial estimator. Interestingly, [14] proved that for the Gaussian design with known $\boldsymbol{\Sigma} = \boldsymbol{I}_{p\times p}$, the sample size $n \ge Cs_0\log(p/s_0)$ is sufficient in de-biasing the Lasso for the estimation of β_j at the $n^{-1/2}$ rate. More recently, [15] extended this result and showed that $n \ge Cs_0(\log p)^2$ is sufficient to de-bias the Lasso for the estimation of β_j at the $n^{-1/2}$ rate for Gaussian designs with known covariance matrices $\boldsymbol{\Sigma}$ when the ℓ_1 norm of each column of $\boldsymbol{\Sigma}^{-1}$ is bounded, i.e., for some constant $\rho > 0$

$$\max_{j=1,\ldots,p} \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1 \le \rho \tag{1.11}$$

holds, where $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$ is the canonical basis in $\mathbb{R}^p$. From this perspective, the present paper provides an extension of these results to more general $\boldsymbol{\Sigma}$: We will see below that for $n \ge Cs_0\log(p)^2$, the efficiency of the de-biasing scheme (1.8) is specific to assumption (1.11) and that the de-biasing scheme (1.8) requires a modification to be efficient in cases where (1.11) is violated.

The paper is organized as follows. Section 2 provides a description of our proposed estimator, which is a modification of the de-biasing scheme (1.8) that accounts for the degrees-of-freedom of the initial estimator. Section 3 describes our strongest results in linear regression with known covariance matrix for the Lasso. This includes several efficiency results for the de-biasing scheme modified with

[1](footnote) Note that (1.10) is stated slightly differently than in [15]: It can be equivalently stated as a lower bound on the expected length of $(1-\alpha)$-confidence intervals for β_j valid uniformly over $\Theta(s_0, s_\Omega, \rho)$ up to constants depending on α. This follows by picking as $\hat{\beta}_j$ any point in the confidence interval, or by constructing a confidence interval from an estimate $\hat{\beta}_j$ and its maximal expected length over $\Theta(s_0, s_\Omega, \rho)$ by Markov's inequality.

degrees-of-freedom adjustment and a characterization of the asymptotic regime where this adjustment is necessary. Section 4 studies the specific situation where bounds on the ℓ_1 norm of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ are available, similarly to (1.11) when $\boldsymbol{a}_0$ is a canonical basis vector. The additional assumptions on $\boldsymbol{\Sigma}^{-1}$ and the results of Section 4 explain why the necessity of degrees-of-freedom adjustment did not appear in some previous works. Section 5 provides a new ℓ_∞ bound for estimation of $\boldsymbol{\beta}$ by the Lasso under assumptions similar to (1.11). Section 6 discusses efficiency and regularity, and shows that asymptotic normality remains unchanged under non-sparse $n^{-1/2}$-perturbations of $\boldsymbol{\beta}$. Section 7 shows that the degrees-of-freedom adjustment is also needed for certain non-Gaussian designs. The proofs of the main results are given in Section 8 and in Appendices A, B, C and H of the supplement [2]. The proofs of intermediary lemmas and propositions can be found in Appendices D to G of the supplement [2]. Our main technical tool is a carefully constructed Gaussian interpolation path described in Section 8.1.

Notation

We use the following notation throughout the paper. Let $\boldsymbol{I}_d$ be the identity matrix of size $d \times d$, e.g. $d = n, p$. For any $p \geq 1$, let $[p]$ be the set $\{1, \ldots, p\}$. For any vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$ and any set $A \subset [p]$, the vector $\boldsymbol{v}_A \in \mathbb{R}^{|A|}$ is the restriction $(v_j)_{j\in A}$. For any $n \times p$ matrix $\boldsymbol{M}$ with columns $(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_p)$ and any subset $A \subset [p]$, let $\boldsymbol{M}_A = (\boldsymbol{M}_j, j \in A)$ be the matrix composed of columns of $\boldsymbol{M}$ indexed by A, and $\boldsymbol{M}_A^\dagger$ be the Moore-Penrose generalized inverse of $\boldsymbol{M}_A$. If $\boldsymbol{M}$ is a symmetric matrix of size $p \times p$ and $A \subset [p]$, then $\boldsymbol{M}_{A,A}$ denotes the sub-matrix of $\boldsymbol{M}$ with rows and columns in A, and $\boldsymbol{M}_{A,A}^{-1}$ is the inverse of $\boldsymbol{M}_{A,A}$. Let $\|\cdot\|_q$ denote the ℓ_q norm of vectors, $\|\cdot\|_{op}$ the operator norm (largest singular value) of matrices and $\|\cdot\|_F$ the Frobenius norm. We use the notation $\langle\cdot,\cdot\rangle$ for the canonical scalar product of vectors in $\mathbb{R}^n$ or $\mathbb{R}^p$, i.e., $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}^\top \boldsymbol{b}$ for two vectors $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension.

Throughout the paper, $C_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2$, $\boldsymbol{u}_0$ is as in (1.6) and F_θ as in (1.7). The score vector $\boldsymbol{z}_0$ is always defined as $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{u}_0$ and $\boldsymbol{Q}_0$ is the matrix $\boldsymbol{Q}_0 = \boldsymbol{I}_{p\times p} - \boldsymbol{u}_0\boldsymbol{a}_0^\top$, so that

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + \boldsymbol{z}_0\boldsymbol{a}_0^\top$$

always holds. As in (1.9), S and s_0 are the support and number of nonzero coefficients of the unknown coefficient vector $\boldsymbol{\beta}$. For any event Ω, denote by I_Ω its indicator function and $a_+ = \max(0, a)$ for $a \in \mathbb{R}$.

2. Degrees of freedom adjustment

2.1. Known $\boldsymbol{\Sigma}$

In addition to the de-biasing scheme (1.8), we consider the following degrees-of-freedom adjusted version of it. Suppose that the Lasso estimator $\hat{\boldsymbol{\beta}}^{(\text{lasso})}$ is used as the initial estimator $\hat{\boldsymbol{\beta}}^{(\text{init})}$, where

$$\hat{\boldsymbol{\beta}}^{(\text{lasso})} = \mathop{\arg\min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2n) + \lambda\|\boldsymbol{b}\|_1 \right\}. \tag{2.1}$$

The degrees-of-freedom adjusted LDPE is defined as

$$\widehat{\theta}_\nu = \left\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{lasso})} \right\rangle + \frac{\left\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})} \right\rangle}{\|\boldsymbol{z}_0\|_2^2(1 - \nu/n)}, \tag{2.2}$$

where z_0 is as in (1.8) and $\nu \in [0, n)$ is a degrees-of-freedom adjustment; ν is allowed to be random. Our theoretical results will justify the degrees-of-freedom adjustment $\nu = |\widehat{S}|$ where $\widehat{S} = \mathrm{supp}(\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})})$. The size of the selected model has the interpretation of degrees of freedom for the Lasso estimator in the context of Stein's Unbiased Risk Estimate (SURE) [20,28,37].

We still retain other possibilities for ν such as $\nu = 0$ in order to analyse the unadjusted de-biasing scheme (1.8). With some abuse of notation, in order to avoid any ambiguity we may sometimes use the notation $\widehat{\theta}_{\nu=0}$ for the unadjusted (1.8) and $\widehat{\theta}_{\nu=|\widehat{S}|}$ for (2.2) with $|\widehat{S}|$ being the size of the support of the Lasso.

Our main results will be developed in Section 3. Here is a simpler version of the story.

Theorem 2.1. *Let s_0, n and p be positive integers satisfying $p/s_0 \to \infty$ and $(s_0/n)\log(p/s_0) \to 0$. Assume that $\boldsymbol{\Sigma}_{jj} \le 1$ for all $j \in [p]$ and that the spectrum of $\boldsymbol{\Sigma}$ is uniformly bounded away from 0 and ∞; e.g. $\max(\|\boldsymbol{\Sigma}\|_{op}, \|\boldsymbol{\Sigma}^{-1}\|_{op}) \le 2$. Let $\lambda = 1.01\sigma\sqrt{2\log(8p/s_0)/n}$.*

(i) Then $|\widehat{S}| = O_{\mathbb{P}}(s_0) = o_{\mathbb{P}}(n)$ and for $\nu = |\widehat{S}|$ we have for every $\boldsymbol{a}_0$

$$\sqrt{nF_\theta}\left(1 - |\widehat{S}|/n\right)\left(\widehat{\theta}_{\nu=|\widehat{S}|} - \theta\right) = T_n + o_{\mathbb{P}}(1) \tag{2.3}$$

where $T_n = \sqrt{nF_\theta}\langle z_0, \boldsymbol{\varepsilon}\rangle/\|z_0\|_2^2$ has the t-distribution with n degrees of freedom. Thus the estimator (2.2) enjoys asymptotic efficiency when $\nu = |\widehat{S}|$.

(ii) The quantity $\sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) - T_n$ is unbounded for certain $\boldsymbol{\beta}$ satisfying $n/\log(p/s_0) \gg s_0 \gg n^{2/3}/\log(p/s_0)^{1/3}$ and $\boldsymbol{a}_0$ depending on S and $\boldsymbol{\Sigma}$ only. Consequently, the unadjusted (1.8) cannot be efficient.

Theorem 2.1(ii) implies that with $\nu = 0$, the unadjusted (1.8) cannot be efficient in the whole range $\{s_0 : s_0\log(p/s_0) \lll n\}$ of sparsity levels unless extra assumptions are made on the covariance matrix $\boldsymbol{\Sigma}$ such as (1.11). Theorem 2.1(i) shows that using the adjustment $\nu = |\widehat{S}|$ repairs this: The efficiency in (2.3) then holds in the whole range $\{s_0 : s_0\log(p/s_0) \lll n\}$ of sparsity levels. Theorem 2.1(i) is proved after Corollary 3.2 below while (ii) is a consequence of the following proposition.

Proposition 2.2. *Let the setting and assumptions of Theorem 2.1 be fulfilled and let ν be a random variable with $\nu \in [0, n)$ almost surely. Then*

$$\sqrt{nF_\theta}(\widehat{\theta}_\nu - \theta) = T_n + o_{\mathbb{P}}(1) + n^{-1}(\nu - |\widehat{S}|)\Lambda_\nu \tag{2.4}$$

where $\Lambda_\nu = \sqrt{nF_\theta}(1 - \nu/n)^{-1}(1 - |\widehat{S}|/n)^{-1}\|z_0\|^{-2}\langle z_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}\rangle$. Furthermore for any (s_Ω, s_0) with $s_\Omega \le s_0 = o(n/\log(p/s_0))$, and any $\boldsymbol{a}_0$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = 1$ and $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_0 = s_\Omega$, there exists $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_0 = s_0$ such that

$$\mathbb{P}\big[|\Lambda_\nu| \ge \|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1\sqrt{\log(8p/s_0)}\big] \to 1, \quad \mathbb{P}\big[|\widehat{S}| \ge s_0\big] \to 1. \tag{2.5}$$

In particular, it is possible to pick $\boldsymbol{a}_0$ satisfying in addition $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1 \ge s_\Omega^{1/2}/\|\boldsymbol{\Sigma}\|_{op}^{1/2}$.

Proposition 2.2 is proved in Appendix C of the supplement [2]. Theorem 2.1(ii) is implied by Proposition 2.2 with $s_\Omega = s_0$: If $s_0^{3/2}\sqrt{\log(8p/s_0)}/n \to +\infty$ then $n^{-1}|\widehat{S}|\Lambda_{\nu=0}$ is unbounded with probability approaching one by (2.5), while the other terms in (2.4) are stochastically bounded.

Example 2.1. It is informative to unpack from the proof of Proposition 2.2 how $(\boldsymbol{a}_0, \boldsymbol{u}_0, \boldsymbol{\beta})$ is constructed so that (2.5) holds. Theorem 2.1 and Proposition 2.2 apply to any $\boldsymbol{\Sigma}$ with bounded spectrum

and $\Sigma_{ii} \le 1$. Let $\boldsymbol{\beta}$ be an s_0-sparse vector with large enough non-zero coefficients and $\beta_j > 0$ for some index $j = 1, \dots, p$ that is fixed throughout this example. Then let $\boldsymbol{v} \in \{-1, 0, 1\}^p$ be an s_Ω-sparse vector with $\operatorname{supp}(\boldsymbol{v}) \subset \operatorname{supp}(\boldsymbol{\beta})$ and $v_k = \operatorname{sgn}(\beta_k)$ for all $k \in \operatorname{supp}(\boldsymbol{v})$. Consider

$$\boldsymbol{\Omega} = \boldsymbol{I}_p + (1/4)s_\Omega^{-1/2}[\boldsymbol{e}_j \boldsymbol{v}^\top + \boldsymbol{v}\boldsymbol{e}_j^\top].$$

which has bounded spectrum since $\|\boldsymbol{\Omega} - \boldsymbol{I}_p\|_{op} \le 1/2$ and set $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}\kappa$ for some constant $\kappa > 0$ such that $\max_{j=1,\dots,p} \Sigma_{jj} = 1$. Since $\boldsymbol{\Omega}$ has bounded spectrum, κ is also bounded and the spectrum of $\boldsymbol{\Sigma}$ is bounded as required. From the proof of Proposition 2.2, we see that the requirement for $\boldsymbol{u}_0$ is that

$$\langle \boldsymbol{u}_0, \operatorname{sgn}(\boldsymbol{\beta}) \rangle = \|\boldsymbol{u}_0\|_1 \tag{2.6}$$

must hold. For the $\boldsymbol{\Sigma}$ just defined, set $\boldsymbol{u}_0 = \boldsymbol{\Omega}\boldsymbol{e}_j = \big(1 + (1/4)s_\Omega^{-1/2} v_j\big)\boldsymbol{e}_j + (1/4)s_\Omega^{-1/2}\boldsymbol{v}$. Since $\boldsymbol{\beta}$ was chosen with $\beta_j > 0$, we have $v_j \ge 0$ by definition of $\boldsymbol{v}$ and $\boldsymbol{u}_0$ satisfies (2.6). These quantities $(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{u}_0)$, when $\boldsymbol{\beta}$ has large enough coefficients, satisfy (2.5) by the proof of Proposition 2.2. Finally, from (1.6) there is a one-to-one correspondence between $\boldsymbol{u}_0$ and $\boldsymbol{a}_0$ given by $\boldsymbol{a}_0 = \boldsymbol{\Sigma}\boldsymbol{u}_0/\langle \boldsymbol{u}_0, \boldsymbol{\Sigma}\boldsymbol{u}_0 \rangle$. This implies $\boldsymbol{a}_0 = \boldsymbol{\Omega}^{-1}\boldsymbol{u}_0/\langle \boldsymbol{u}_0, \boldsymbol{\Omega}^{-1}\boldsymbol{u}_0 \rangle$ and since $\boldsymbol{u}_0 = \boldsymbol{\Omega}\boldsymbol{e}_j$, the direction $\boldsymbol{a}_0$ for this example is proportional to the canonical basis vector $\boldsymbol{e}_j$. Proposition 2.2 thus proves the necessity of the degrees-of-freedom adjustment with $\boldsymbol{a}_0$ proportional to $\boldsymbol{e}_j$. Figure 2 illustrates this phenomenon on simulated data.

The adjustment in (2.2) was proposed by [14] in the form of

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\text{lasso})} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})})}{n - \nu} \tag{2.7}$$

based on heuristics of the replica method from statistical physics and a theoretical justification in the case of $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. As $z_0 = \boldsymbol{X}\boldsymbol{u}_0$ with $\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0 \rangle$ in (1.8), $\mathbb{E}\|z_0\|_2^2/n = 1/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0 \rangle$ and

$$\frac{\langle z_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{init})} \rangle}{(\mathbb{E}\|z_0\|_2^2)(1 - \nu/n)} = \left\langle \boldsymbol{a}_0, \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})})}{n - \nu} \right\rangle.$$

Thus, the plug-in estimator

$$\widehat{\theta}_\nu = \langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}} \rangle \text{ with the } \hat{\boldsymbol{\beta}} \text{ in (2.7)}, \tag{2.8}$$

is equivalent to replacing $\|z_0\|_2^2$ with its expectation in the denominator of the bias correction term in (2.2). Another version of the estimator, akin to the version of the LDPE proposed in [31], is

$$\widehat{\theta}_\nu = \langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{lasso})} \rangle + \frac{\langle z_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})} \rangle}{\langle z_0, \boldsymbol{X}\boldsymbol{u} \rangle (1 - \nu/n)} \tag{2.9}$$

with a vector $\boldsymbol{u} \in \mathbb{R}^p$ satisfying $\langle \boldsymbol{u}, \boldsymbol{a}_0 \rangle = 1$. Since $\mathbb{E}\langle z_0, \boldsymbol{X}\boldsymbol{u} \rangle = \mathbb{E}\|z_0\|_2^2$, the estimator (2.7) also corresponds to (2.9) with $\langle z_0, \boldsymbol{X}\boldsymbol{u} \rangle$ replaced by its expectation in the denominator of the bias correction term.

Let $\boldsymbol{h}^{(\text{lasso})} = (\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta})$. It is worthwhile to mention here that when $\|\boldsymbol{X}\boldsymbol{h}^{(\text{lasso})}\|_2/\sqrt{n} = o_{\mathbb{P}}(1)$ based on existing results on the Lasso, the asymptotic distribution of (2.2) adjusted at the $n^{-1/2}$ rate does not

change when $\|z_0\|_2^2$ is replaced by a quantity of type $\|z_0\|_2^2(1+O(n^{-1/2}))$ in the denominator of the bias correction term. Indeed,

$$\sqrt{nF_\theta}\left|\frac{\left\langle z_0,\, \boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})}\right\rangle}{\|z_0\|_2^2(1-\nu/n)}-\frac{\left\langle z_0,\, \boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})}\right\rangle}{\|z_0\|_2^2(1+O(n^{1/2}))(1-\nu/n)}\right| \tag{2.10}$$
$$\le O(1)(1-\nu/n)^{-1}\left(|T_n|n^{-1/2}+\|\boldsymbol{X}\boldsymbol{h}^{(\text{lasso})}\|_2/(\sigma C_0\|z_0\|_2)\right).$$

The right-hand side converges to 0 in probability if $(1-\nu/n)^{-1}=O_{\mathbb{P}}(1)$ and $\|\boldsymbol{X}\boldsymbol{h}^{(\text{lasso})}\|_2/\sqrt{n}=o_{\mathbb{P}}(1)$ since T_n has the t-distribution with n degrees of freedom. Thus, as (2.2), (2.8) and (2.9) are asymptotically equivalent, the most notable feature of these estimators is the degrees-of-freedom adjustment with the choice $\nu=|\widehat{S}|$, as proposed in [14], compared with earlier proposals with $\nu=0$. While the properties of these estimators for general $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ will be studied in the next section, we highlight in the following theorem the requirement of either a degrees-of-freedom adjustment or some extra condition on the bias of the Lasso in the special case where the Lasso is sign consistent.

Theorem 2.3. *Suppose that the Lasso is sign consistent in the sense of*

$$\mathbb{P}\left\{\operatorname{sgn}(\hat{\boldsymbol{\beta}}^{(\text{lasso})})=\operatorname{sgn}(\boldsymbol{\beta})\right\}\to 1. \tag{2.11}$$

Let $C_0=\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2$ *and* $C_{\boldsymbol{\beta}}=\|\boldsymbol{\Sigma}_{S,S}^{-1/2}\operatorname{sgn}(\boldsymbol{\beta}_S)\|_2/\sqrt{s_0}$. *Suppose that* $\sqrt{(1\vee s_0)/n}+C_{\boldsymbol{\beta}}\sqrt{s_0}(\lambda/\sigma)\big)\le\eta_n$ *for a sufficiently small* $\eta_n<1$. *Let* $F_\theta=1/(\sigma C_0)^2$ *be the Fisher information as in* (1.7), *and* $T_n=\sqrt{nF_\theta}\langle z_0,\boldsymbol{\varepsilon}\rangle/\|z_0\|_2^2$ *so that* T_n *has the t-distribution with n degrees of freedom. Let* $\widehat{\theta}_\nu$ *be as in* (2.2) *or* (2.8). *Then*,

$$(1-\nu/n)\sqrt{nF_\theta}\big(\widehat{\theta}_\nu-\theta\big)=T_n+O_{\mathbb{P}}(\eta_n) \tag{2.12}$$

for a random variable $\nu\in[0,s_0]$ *if and only if*

$$\sqrt{F_\theta/n}\big(s_0-\nu\big)\big\langle\boldsymbol{a}_0,\hat{\boldsymbol{\beta}}^{(\text{lasso})}-\boldsymbol{\beta}\big\rangle=O_{\mathbb{P}}(\eta_n), \tag{2.13}$$

if and only if

$$\sqrt{F_\theta/n}\big(s_0-\nu\big)\big\langle(\boldsymbol{a}_0)_S,\lambda(\boldsymbol{X}_S^\top\boldsymbol{X}_S/n)^{-1}\operatorname{sgn}(\boldsymbol{\beta}_S)\big\rangle=O_{\mathbb{P}}(\eta_n). \tag{2.14}$$

The conclusion also holds for the $\widehat{\theta}_\nu$ *in* (2.9) *when* $C_0\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|_2=O(1)$.

The proof is given in Appendix H of the supplement [2]. Theorem 2.3 provides an alternative negative result, similar in flavor to Theorem 2.1(ii) and Proposition 2.2 above. The settings may not match exactly since the tuning parameter λ required for sign consistency is larger than the one featured in Theorem 2.1. Compared with Proposition 2.2, the sign consistency lets us derive the two explicit conditions (2.13)-(2.14) for efficiency that are useful to pinpoint situations, such as those described in the next two paragraphs, where efficiency does not hold.

Theorem 2.3 implies that for efficient statistical inference of θ at the $n^{-1/2}$ rate, the unadjusted de-biasing scheme (1.8) requires either a degrees-of-freedom adjustment or the extra condition that the bias of the initial Lasso estimator of θ, given by $\big\langle(\boldsymbol{a}_0)_S,\lambda(\boldsymbol{X}_S^\top\boldsymbol{X}_S/n)^{-1}\operatorname{sgn}(\boldsymbol{\beta}_S)\big\rangle$, is of order $o_{\mathbb{P}}(n^{1/2}/s_0)$, even when the initial Lasso estimator is sign-consistent. For example, if $(\boldsymbol{a}_0)_{S^c}=0$

and $(\boldsymbol{a}_0)_S = \mathrm{sgn}(\boldsymbol{\beta}_S)/\|\boldsymbol{\Sigma}^{-1/2}\mathrm{sgn}(\boldsymbol{\beta}_S)\|_2$, then $\boldsymbol{a}_0$ is standardized with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = 1$ and condition (2.14) on the bias can be written as

$$(\lambda/\sigma)n^{-1/2}(s_0 - \nu)\|\boldsymbol{\Sigma}_{S,S}^{-1/2}\mathrm{sgn}(\boldsymbol{\beta}_S)\|_2 = O_{\mathbb{P}}(\eta_n)$$

because the singular values of the Wishart matrix $\boldsymbol{\Sigma}_{S,S}^{-1/2}(\boldsymbol{X}_S^\top \boldsymbol{X}_S/n)\boldsymbol{\Sigma}_{S,S}^{-1/2}$ are bounded away from 0 and $+\infty$ with high probability. For $\nu = 0$, this is equivalent to $C_{\boldsymbol{\beta}}(\lambda/\sigma)s_0^{3/2}/\sqrt{n} = O(\eta_n)$. If $C_{\boldsymbol{\beta}}$ is of order of a constant and $\eta_n < 1$, this implies that the unadjusted de-biasing scheme (1.8) cannot be efficient in the asymptotic regime when

$$(\lambda/\sigma)s_0^{3/2}/\sqrt{n} \gg 1. \tag{2.15}$$

Interestingly, the condition $(\lambda/\sigma)s_0^{3/2}/\sqrt{n} = O(1)$ is weaker than the typical sample size requirement $n \gg (s_0 \log p)^2$ in the case of unknown $\boldsymbol{\Sigma}$.

Another enlightening situation is $\boldsymbol{a}_0 = \boldsymbol{e}_j$ the j-th canonical basis vector for some $j \in S$, $S = \{1, \ldots, s_0\}$ and $\boldsymbol{\Sigma}^{-1}$ diagonal by block with two blocks

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{I}_{|S|} + (1/4)|S|^{-1/2}[\mathrm{sgn}(\boldsymbol{\beta}_S)\boldsymbol{e}_j^\top + \boldsymbol{e}_j\mathrm{sgn}(\boldsymbol{\beta}_S)^\top] & \boldsymbol{0}_{|S|,p-|S|} \\ \boldsymbol{0}_{p-|S|,|S|} & \boldsymbol{I}_{p-|S|} \end{pmatrix}.$$

The eigenvalues of $\boldsymbol{\Sigma}^{-1/2}$ belong to $[1/2, 3/2]$ by construction since $(|S|)^{-1/2}\mathrm{sgn}(\boldsymbol{\beta}_S)\boldsymbol{e}_j^\top$ has operator norm equal to one. Again using properties of the singular values of the Wishart matrix $\boldsymbol{\Sigma}_{S,S}^{-1/2}(\boldsymbol{X}_S^\top \boldsymbol{X}_S/n)\boldsymbol{\Sigma}_{S,S}^{-1/2}$, the left hand-side of condition (2.14) is of order

$$\sqrt{F_\theta/n}(s_0 - \nu)\lambda\boldsymbol{e}_j^\top(\boldsymbol{\Sigma}_{S,S})^{-1}\mathrm{sgn}(\boldsymbol{\beta}_S) \asymp (\lambda/\sigma)n^{-1/2}(s_0 - \nu)\sqrt{s_0}.$$

Similarly to the previous paragraph, this implies that with $\nu = 0$ the unadjusted de-biasing scheme (1.8) cannot be efficient if (2.15) holds. Up to a multiplicative constant in $\boldsymbol{\Sigma}$, this example is similar to Example 2.1 with $s_\Omega = s_0$.

The novelty of our contributions resides in the $s_0^2 \gg n$ regime up to logarithmic factor, in the sparsity range where the transition (2.15) happens. The necessity of the degrees-of-freedom adjustment can be seen in simulated data as follows. Figure 1 presents the distribution of $\sqrt{n}(\hat{\theta}_\nu - \theta)$ with and without the adjustment for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\sigma = 1$ for $(n, p) = (4000, 6000)$ and $s_0 = 20, 40, 80, 120$. Although classical results on de-biasing in the regime $s_0^2 \ll n$ proves that $\sqrt{n}(\hat{\theta}_\nu - \theta) \approx N(0, 1)$ [13,24,31] with $\nu = 0$, simulations reveal that $\sqrt{n}(\hat{\theta}_\nu - \theta)$ is substantially biased (downward in Figure 1), and any confidence interval constructed from $\sqrt{n}(\hat{\theta}_\nu - \theta) \approx N(0, 1)$ would not correctly control Type-I error due to this substantial bias. This substantial bias is present for sparsity as small as $s_0 = 20$ (for which $s_0^2/n = 0.1$). On the other hand, the adjustment $\nu = |\widehat{S}|$ repairs this, as shown both in the simulation in Figure 1 and by the theory in Theorem 2.1 and in the next sections. Thus our novel results on the necessity of the degrees-of-freedom adjustment is not only theoretical; It also explains the gap between simulations and the predictions from the early literature on de-biasing [13,24,31] where the degrees-of-freedom adjustment is absent.

Guided by Theorem 2.3, one can easily exhibit situations with correlated $\boldsymbol{\Sigma}$ and $\boldsymbol{a}_0$ proportional to $\boldsymbol{e}_j$ (a canonical basis vector), such that the unadjusted estimate leads to spurious inference: One just needs to find problem instances such that (2.14) is large. As an example, Figure 2 shows boxplots of the situation with $\boldsymbol{a}_0 = \boldsymbol{e}_j/(\boldsymbol{\Sigma}^{-1})_{jj}^{1/2}$ sparsity $s_0 = \|\boldsymbol{\beta}\|_0 = 120$, $p = 6000$, $n = 4000$, $\sigma = 1$ and $\boldsymbol{\Sigma}$ is correlated of the form $\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_p + 0.07(\mathrm{sgn}(\boldsymbol{\beta})\boldsymbol{e}_j^\top + \boldsymbol{e}_j\mathrm{sgn}(\boldsymbol{\beta})^\top)$. In the un-adjusted case, the pivotal

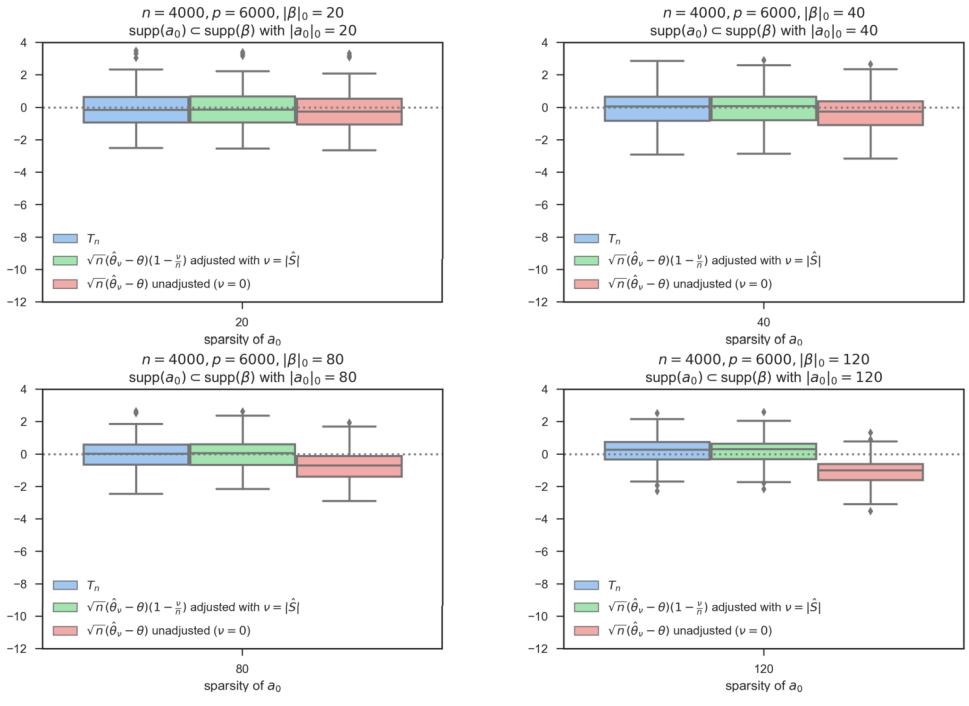

Figure 1. Distribution of $\sqrt{n}(\hat{\theta}_\nu - \theta)$ in the adjusted ($\nu = |\widehat{S}|$) and unadjusted $\nu = 0$ cases. For comparison, T_n has the t-distribution with n degrees-of-freedom. Here $\boldsymbol{a}_0$ is proportional to $\mathrm{sgn}(\boldsymbol{\beta})$ normalized with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = 1$. Experiments were replicated 200 times. A two-sided t-test rejects that the mean of $\sqrt{n}(\hat{\theta}_\nu - \theta)$ is zero in the unadjusted case ($\nu = 0$) with p-value 0.0048 for $s_0 = 20$, p-value 0.00028 for $s_0 = 40$, p-value $7 \cdot 10^{-22}$ for $s_0 = 80$ and p-value $2 \cdot 10^{-31}$ for $s_0 = 120$.

quantity $n^{1/2}(\hat{\theta}_{\nu=0} - \theta)$ is biased downward and would produce misleading confidence intervals with incorrect coverage. The adjustment $\nu = |\widehat{S}|$ exactly repairs this.

Theorem 2.3 requires sign consistency of the Lasso in (2.11). Sufficient conditions for the sign consistency of the Lasso were given in [17,21,26,33]. In particular, [26] gave the following sufficient conditions for (2.11) in the case of linear regression (1.1) with Gaussian design: For certain positive γ, δ and $\phi_p \geq 2$,

$$\left\|\boldsymbol{\Sigma}_{S^c,S}\boldsymbol{\Sigma}_{S,S}^{-1}\mathrm{sgn}(\boldsymbol{\beta}_S)\right\|_\infty \leq 1 - \gamma,$$

$$\lambda = \gamma^{-1}\sigma\sqrt{\phi_p \rho(2/n)\log p},$$

$$\rho(C_{\min}\gamma^2)^{-1}(2s_0/n)\log(p - s_0) + (\phi_p \log p)^{-1}\log(p - s_0) < 1 - \delta,$$

with $\rho = \max_{j \in S^c}\left(\boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,S}\boldsymbol{\Sigma}_{S,S}^{-1}\boldsymbol{\Sigma}_{S,j}\right)$ and $C_{\min} = \min_{\|\boldsymbol{u}\|_2=1}\left\|\boldsymbol{\Sigma}_{S,S}^{-1/2}\boldsymbol{u}\right\|_2$, and

$$\min_{j \in S}|\beta_j| \geq \left(1 + n^{-1/2}c_n\right)\lambda \max_{\|\boldsymbol{u}\|_\infty=1}\left\|\boldsymbol{\Sigma}_{S,S}^{-1/2}\boldsymbol{u}\right\|_\infty^2 + 20\left(\sigma^2(C_{\min}n)^{-1}\log s_0\right)^{1/2},$$

for some $c_n \to \infty$.

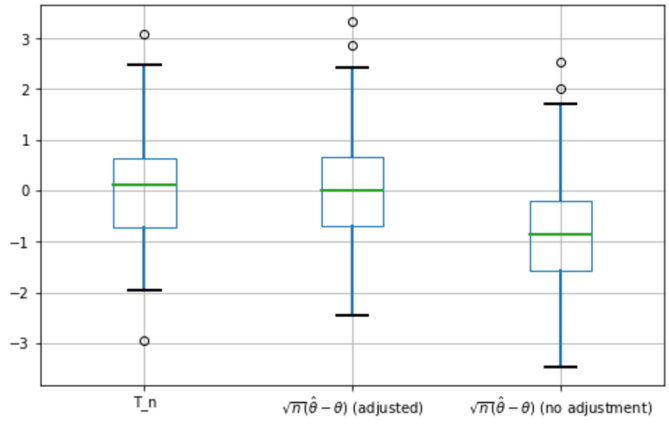

Figure 2. Boxplots of pivotal random variables $\sqrt{n}(\hat{\theta}_n - \theta)$ for $v = 0$ (unadjusted) and $v = |\widehat{S}|$ (adjusted) when $\boldsymbol{a}_0 = \boldsymbol{e}_j/(\boldsymbol{\Sigma}^{-1})_{jj}^{1/2}$, $s_0 = \|\boldsymbol{\beta}\|_0 = 120$, $p = 6000$, $n = 4000$, $\sigma = 1$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_p + 0.07(\mathrm{sgn}(\boldsymbol{\beta})\boldsymbol{e}_j^\top + \boldsymbol{e}_j \mathrm{sgn}(\boldsymbol{\beta})^\top)$. For comparison, T_n has the t-distribution with n degrees-of-freedom.

2.2. Unknown $\boldsymbol{\Sigma}$

In the case of unknown $\boldsymbol{u}_0$, one needs to estimate the ideal score vector $z_0 = \boldsymbol{X}\boldsymbol{u}_0$ as well as the variance level $\|z_0\|^2/n$ in (1.8). In view of (1.6), we consider

$$z = \boldsymbol{X}\boldsymbol{u}, \quad \boldsymbol{Q} = \boldsymbol{I}_p - \boldsymbol{u}\boldsymbol{a}_0^\top \text{ with } \boldsymbol{u} \text{ satisfying } \langle \boldsymbol{u}, \boldsymbol{a}_0 \rangle = 1. \tag{2.16}$$

As $\boldsymbol{Q}^2 = \boldsymbol{Q}$, by algebra and the definitions of $\boldsymbol{u}_0$ and z_0 in (1.6) and (1.8),

$$z = -\boldsymbol{X}\boldsymbol{Q}\boldsymbol{u}_0 + z_0 = \boldsymbol{X}\boldsymbol{Q}\boldsymbol{\gamma} + z_0 \tag{2.17}$$

with $\boldsymbol{\gamma} = -\boldsymbol{Q}\boldsymbol{u}_0$ and $\mathbb{E}\big[(\boldsymbol{X}\boldsymbol{Q})^\top z_0\big] = \boldsymbol{Q}^\top \boldsymbol{\Sigma}\boldsymbol{u}_0 = \boldsymbol{0}$. Hence, (2.17) is a linear model with response vector $z \in \mathbb{R}^n$, Gaussian design matrix $\boldsymbol{X}\boldsymbol{Q} \in \mathbb{R}^{n\times p}$ with n independent rows, true coefficient vector $\boldsymbol{\gamma}$, and Gaussian noise $z_0 \sim N(\boldsymbol{0}, C_0^{-2}\boldsymbol{I}_n)$ independent of $\boldsymbol{X}\boldsymbol{Q}$, where $C_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2$. Note that since $\boldsymbol{Q}$ is rank deficient, the linear model (2.17) is unidentifiable: For both $\widetilde{\boldsymbol{\gamma}} = -\boldsymbol{u}_0$ and $\boldsymbol{\gamma} = -\boldsymbol{Q}\boldsymbol{u}_0$ we have $\boldsymbol{X}\boldsymbol{Q}\widetilde{\boldsymbol{\gamma}} = \boldsymbol{X}\boldsymbol{Q}\boldsymbol{\gamma}$ so that both $\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\gamma}$ can be regarded as the true coefficient vector in the model (2.17). To solve this identifiability issue, we view the parameter space of (2.17) as the image of $\boldsymbol{Q}$ and the true coefficient vector as $\boldsymbol{\gamma} = -\boldsymbol{Q}\boldsymbol{u}_0$.

It is thus natural to estimate z_0 in the linear model (2.17), as was already suggested previously for $\boldsymbol{a}_0 = \boldsymbol{e}_j$ [15,24,31]. Given an estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, we define the estimated score vector

$$\widehat{z} = z - \boldsymbol{X}\boldsymbol{Q}\widehat{\boldsymbol{\gamma}} \tag{2.18}$$

and the corresponding de-biased estimate

$$\hat{\theta}_{v,\widehat{z}} = \langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} \rangle + \frac{\langle \widehat{z}, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} \rangle}{(1 - v/n)\langle \widehat{z}, z \rangle}. \tag{2.19}$$

This corresponds to (2.9) with the ideal score vector z_0 replaced by $\widehat{z}$.

The vector $\boldsymbol{u}$ in (2.16) that defines the linear model (2.17) should be picked carefully to yield small prediction error $\|z_0 - \widehat{z}\|_2^2/n = \|\boldsymbol{X}\boldsymbol{Q}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2/n$ in the linear model (2.16). As $z_0 = \boldsymbol{X}\boldsymbol{u}_0$ with a

high-dimensional $\boldsymbol{u}_0$, it would be reasonable to expect that a sparsity condition on $\boldsymbol{u}_0$ would ensure proper convergence of $\widehat{z}$ to z_0. However, this requires a connection between the sparsity of $\boldsymbol{\gamma} = -\boldsymbol{Q}\boldsymbol{u}_0$ to that of $\boldsymbol{u}_0$. To this end, we pick

$$\boldsymbol{u} = \boldsymbol{e}_{j_0}/(a_0)_{j_0} \text{ with } j_0 = \underset{j=1,\ldots,p}{\arg\max} \left|(a_0)_j\right|. \tag{2.20}$$

For the above choice of $\boldsymbol{u}$,

$$\langle \boldsymbol{u}, \boldsymbol{a}_0 \rangle = 1, \quad \|\boldsymbol{Q}\boldsymbol{h}\|_0 \le 1 + \|\boldsymbol{h}\|_0, \quad \|\boldsymbol{Q}\boldsymbol{h}\|_1 \le 2\|\boldsymbol{h}\|_1, \quad \forall\, \boldsymbol{h} \in \mathbb{R}^p, \tag{2.21}$$

so that the sparsity of $\boldsymbol{u}_0$ implies that of $\boldsymbol{\gamma}$. This leads to the Lasso estimator

$$\widehat{\boldsymbol{\gamma}} = \boldsymbol{Q}\widehat{\boldsymbol{b}} \text{ with } \widehat{\boldsymbol{b}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|z - \boldsymbol{X}\boldsymbol{Q}\boldsymbol{b}\|_2^2/(2n) + \widehat{\tau} A \lambda_{\text{univ}} \|\boldsymbol{Q}\boldsymbol{b}\|_1 \right\} \tag{2.22}$$

where $\lambda_{\text{univ}} = \sqrt{(2/n)\log p}$, A is an upper bound for $\max_{j=1,\ldots,p} \|\boldsymbol{X}\boldsymbol{Q}\boldsymbol{e}_j\|_2/n^{1/2}$ and $\widehat{\tau}$ is an estimate of the noise level C_0^{-1} in the regression model (2.17). We note the delicate difference between (2.22) and the usual Lasso as the estimator and penalty are both restricted to the image of $\boldsymbol{Q}$. To the best of our knowledge, the regression model (2.17) in the direction (2.20), which plays a crucial role in our analysis, provides a new way of dealing with dense direction $\boldsymbol{a}_0$ in de-biasing the Lasso. We note that the natural choice $\widetilde{\boldsymbol{u}} = \boldsymbol{a}_0\|\boldsymbol{a}_0\|_2^{-2}$ satisfies $\langle \boldsymbol{a}_0, \widetilde{\boldsymbol{u}} \rangle = 1$, but for certain dense $\boldsymbol{a}_0$ the corresponding projection matrix $\widetilde{\boldsymbol{Q}} = \boldsymbol{I}_p - \widetilde{\boldsymbol{u}}\boldsymbol{a}_0^\top$ does not preserve sparsity as in (2.21).

For the purpose of the asymptotic normality result in Theorem 2.5 below, we will consider estimators $\widehat{\boldsymbol{\gamma}}$ satisfying

$$\|\boldsymbol{Q}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_1 = O_{\mathbb{P}}(C_0^{-1}) \min\left\{ \|\boldsymbol{\gamma}\|_0 \lambda_{\text{univ}}, C_0 \|\boldsymbol{\gamma}\|_1 \right\}, \tag{2.23}$$

$$\sup\left\{ \langle \widehat{z}, \boldsymbol{X}\boldsymbol{Q}\boldsymbol{h} \rangle / n : \|\boldsymbol{Q}\boldsymbol{h}\|_1 = 1 \right\} = O_{\mathbb{P}}(C_0^{-1}\lambda_{\text{univ}}). \tag{2.24}$$

Inequality (2.23) is the usual ℓ_1 estimation rate when $\boldsymbol{\gamma}$ is sparse or $\boldsymbol{\gamma}$ has small ℓ_1 norm. Condition (2.24) holds automatically for the Lasso estimator (2.22) when $C_0\widehat{\tau} = O_{\mathbb{P}}(1)$ as a consequence of the KKT conditions as explained in the following proposition.

Proposition 2.4. *Let $z = \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{Q} = \boldsymbol{I}_p - \boldsymbol{u}\boldsymbol{a}_0^\top$ be as in (2.16) with the $\boldsymbol{u}$ in (2.20). Assume that $\boldsymbol{\Sigma}_{j,j} \le 1\, \forall j$, $\phi_{\min}(\boldsymbol{\Sigma})$ is bounded away from 0, and $\min\{\|\boldsymbol{Q}\boldsymbol{u}_0\|_0 \log(p)/n, C_0\|\boldsymbol{Q}\boldsymbol{u}_0\|_1\sqrt{\log(p)/n}\} = o(1)$. Let $\boldsymbol{Q}\widehat{\boldsymbol{\gamma}}$ be as in (2.22) with a constant $A > 2$ and $\widehat{\tau}$ satisfying one of the following conditions*:

(i) *$\widehat{\tau} = \|z - \boldsymbol{X}\boldsymbol{Q}\widehat{\boldsymbol{\gamma}}\|_2/n^{1/2}$ is the recursive solution of (2.22) as scaled Lasso [19],*
(ii) *or $\widehat{\tau}$ is any estimator satisfying $1 + o_{\mathbb{P}}(1) \le C_0\widehat{\tau} \le O_{\mathbb{P}}(1)$.*

Then, the requirements (2.23)-(2.24) are satisfied.

Proposition 2.4 is proved in Appendix C of the supplement [2]. The following is our main result for unknown $\boldsymbol{\Sigma}$.

Theorem 2.5. *Assume that $\boldsymbol{\Sigma}_{jj} \le 1$ for all $j \in [p]$ and that the spectrum of $\boldsymbol{\Sigma}$ is uniformly bounded away from 0 and ∞; e.g. $\max(\|\boldsymbol{\Sigma}\|_{op}, \|\boldsymbol{\Sigma}^{-1}\|_{op}) \le 2$. Let $\lambda = 1.01\sigma\sqrt{2\log(8p/s_0)/n}$ for the Lasso (2.1) in the linear model (1.1). Let $\epsilon_n > 0$ with $\epsilon_n = o(1)$ and $\mathscr{B}_n$ be the class of $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathbb{R}^{p\times 2}$ satisfying*

$$\frac{\|\boldsymbol{\beta}\|_0 \log p}{n} + \min\left\{ \frac{\|\boldsymbol{u}_0\|_0 \log p}{n}, \frac{C_0\|\boldsymbol{u}_0\|_1\sqrt{\log p}}{\sqrt{n}} \right\} \le \epsilon_n \tag{2.25}$$

and $\boldsymbol{a}_0 \neq \boldsymbol{0}$, *where* $\boldsymbol{u}_0$ *is as in* (1.6). *Given* $\boldsymbol{a}_0 \neq \boldsymbol{0}$, *let* $\boldsymbol{u}$ *be as in* (2.20), $\boldsymbol{Q} = \boldsymbol{I}_p - \boldsymbol{u}\boldsymbol{a}_0^\top$, $\boldsymbol{Q}\widehat{\boldsymbol{\gamma}}$ *an estimator of* $\boldsymbol{\gamma} = -\boldsymbol{Q}\boldsymbol{u}_0$ *in the linear model* (2.17) *satisfying* (2.23)-(2.24), $\widehat{\boldsymbol{z}}$ *the estimated score vector in* (2.18), *and* $\hat{\theta}_{v,\widehat{\boldsymbol{z}}}$ *the de-biased estimate in* (2.19). *If* $v = \|\hat{\boldsymbol{\beta}}^{(lasso)}\|_0$, *then uniformly for* $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathscr{B}_n$

$$\sqrt{nF_\theta}(\widehat{\theta}_{v=|\widehat{S}|,\widehat{\boldsymbol{z}}} - \theta) = Z_n + O_{\mathbb{P}}(r_n)$$

holds, where $Z_n \to^d N(0,1)$ *and*

$$r_n = r_{n,p}(\boldsymbol{\beta}, \boldsymbol{a}_0) = \min\left\{ \frac{(\|\boldsymbol{\beta}\|_0 \wedge \|\boldsymbol{u}_0\|_0)\log(p)}{\sqrt{n}}, C_0\|\boldsymbol{u}_0\|_1\sqrt{\log p} \right\}. \tag{2.26}$$

Consequently, for all $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathscr{B}_n$ *satisfying* $r_n \to 0$,

$$\sqrt{nF_\theta}(\widehat{\theta}_{v=|\widehat{S}|,\widehat{\boldsymbol{z}}} - \theta) \to^d N(0,1).$$

Theorem 2.5 is proved in Appendix C the supplement [2]. The sparsity condition (2.25) is mild: it only requires that the squared prediction rate for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ converge to 0. Under this condition, Theorem 2.5 shows that estimation of θ is possible, for general directions $\boldsymbol{a}_0 \neq 0$, at the rate $n^{-1/2}(1 + r_n)$ where r_n is given by (2.26). The rate $n^{-1/2}(1 + r_n)$ is optimal as it matches the lower bound in Proposition 4.2 of [15] for the estimation of $\theta = \beta_j$ in the canonical basis directions $\boldsymbol{a}_0 = \boldsymbol{e}_j$ stated in (1.10). Before Theorem 2.5, it was unknown whether the lower bound (1.10) can be attained (cf. for instance the discussion in Remark 4.3 of [15]). Theorem 2.5 closes this gap, extends the upper bound to general direction $\boldsymbol{a}_0$, and relaxes the ℓ_1 bound on $\boldsymbol{\Sigma}$ imposed in [15].

The recent work [9] proposes an alternative construction of a score vector for general direction $\boldsymbol{a}_0$ based on a quadratic program. This quadratic program is similar to the construction in [13,31], with a modification to handle general direction $\boldsymbol{a}_0$, see [9], equation (7), (8) and (10). The upper bounds in [9], Corollaries 3 and 4, require $\|\boldsymbol{\beta}\|_0 \lesssim \sqrt{n}/\log p$ in contrast with Theorem 2.5 where $\|\boldsymbol{\beta}\|_0 \ggg \sqrt{n}$ is allowed.

Another recent line of research [34–36] consider the construction of confidence intervals for $\boldsymbol{a}_0^\top\boldsymbol{\beta}$ for general directions $\boldsymbol{a}_0$ without sparsity assumption on $\boldsymbol{\beta}$. These works consider the setting where $\boldsymbol{\beta}$ is arbitrary but bounded in the sense that $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 \leq C$ for some constant $C \asymp \sigma^2$ independent of n, p. In this setting, $\|\boldsymbol{\beta}\|_0 \log(p)/n \to 0$ is violated and consistent estimation of $\boldsymbol{\beta}$ or $\boldsymbol{Q}_0\boldsymbol{\beta}$ is not possible. Assuming $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 \leq C$ instead of a sparsity assumption on $\boldsymbol{\beta}$ leads to different minimax rates: The rate in [34], Corollary 5, does not depend on $\|\boldsymbol{\beta}\|_0$ but depends implicitly on $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2$ instead; hence the rate in Theorem 2.5 and (1.10) is not directly comparable to theirs. On a higher level, this line of research is fundamentally different than the present work: [34–36] leverage the assumption that the nuisance part of the signal, $\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{\beta}$, is bounded with componentwise variance of the same order as that of the noise, without attempting to estimate the nuisance part of the signal. In contrast, Theorem 2.5 attemps to estimate the nuisance parameter and the nuisance part of the signal $\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{\beta}$ is allowed to have arbitrarily large componentwise variance.

Next, we prove that the de-biased estimator in Theorem 2.5 for unknown $\boldsymbol{\Sigma}$, and the ideal $\widehat{\theta}_v$ in (2.2) for known $\boldsymbol{\Sigma}$ as well, would not achieve the same rate without the degrees-of-freedom adjustment. Compared with Theorem 2.3, Theorem 2.6 below is somewhat less explicit but the sign consistency of the Lasso is no longer required.

Theorem 2.6. *Let* $\boldsymbol{\Sigma}$, ϵ_n, $\mathscr{B}_n$, r_n, $\widehat{\boldsymbol{z}}$ *and* $\hat{\theta}_{v,\widehat{\boldsymbol{z}}}$ *be as in Theorem* 2.5. *Let* s_0 *and* s_Ω *be positive integers satisfying* $s_0\log(p)/n \leq \epsilon_n$, $s_\Omega \leq s_0$ *and*

$$n^{-1}s_0 s_\Omega^{1/2}\sqrt{\log(8p/s_0)} \ggg 1 + s_\Omega\log(p)/n^{1/2}. \tag{2.27}$$

If $\nu = 0$, which means no degrees-of-freedom adjustment in (2.19), then there exist $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathscr{B}_n$ such that $\|\boldsymbol{\beta}\|_0 = s_0$, $\|\boldsymbol{u}_0\|_0 = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_0 = s_\Omega$, and $\sqrt{nF_\theta}(\widehat{\theta}_{\nu,\widehat{\boldsymbol{z}}} - \theta)/(1 + r_n)$ is stochastically unbounded. Moreover, the above statement also holds when $\hat{\theta}_{\nu,\widehat{\boldsymbol{z}}}$ is replaced by $\widehat{\theta}_\nu$ in (2.2).

Theorem 2.6 is proved in Appendix C of the supplement [2]. As an example, if $s_\Omega = \epsilon_n\sqrt{n}/\log(p)$ and $s_0 \geq \epsilon_n^{-1}n^{3/4}$ for some $\epsilon_n \to 0$, then

- (2.27) holds so that, without adjustment, $\sqrt{nF_\theta}(\widehat{\theta}_{\nu=0,\widehat{\boldsymbol{z}}} - \theta)$ is unbounded by Theorem 2.6 for some $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathscr{B}_n$ with $\|\boldsymbol{\beta}\|_0 = s_0$ and $\|\boldsymbol{u}_0\|_0 = s_\Omega$.
- $r_n \to 0$ hence $\sqrt{nF_\theta}(\widehat{\theta}_{\nu=|\widehat{S}|,\widehat{\boldsymbol{z}}} - \theta) \to^d N(0, 1)$ by Theorem 2.5 and the de-biased estimate adjusted with $\nu = \|\hat{\boldsymbol{\beta}}^{(\text{lasso})}\|_0$ is efficient for all $(\boldsymbol{\beta}, \boldsymbol{a}_0) \in \mathscr{B}_n$ with $\|\boldsymbol{\beta}\|_0 \leq s_0$, $\|\boldsymbol{u}_0\|_0 \leq s_\Omega$.

2.3. Unknown $\boldsymbol{\Sigma}$ and canonical basis directions $\boldsymbol{a}_0 = \boldsymbol{e}_j$

For convenience we provide here the notation and corollary of Theorem 2.5 in the case of canonical basis vector $\boldsymbol{a}_0 = \boldsymbol{e}_j$ for some $j \in \{1, \ldots, p\}$. We denote $(\boldsymbol{u}_0, z_0, \widehat{\boldsymbol{z}})$ by $(\boldsymbol{u}_j, z_j, \widehat{\boldsymbol{z}}_j)$ and write the linear model (2.17) as

$$\boldsymbol{X}\boldsymbol{e}_j = \boldsymbol{X}^{(-j)}\boldsymbol{\gamma}^{(j)} + \boldsymbol{z}_j \tag{2.28}$$

where $\boldsymbol{X}^{(-j)} \in \mathbb{R}^{n\times(p-1)}$ is the matrix $\boldsymbol{X}$ with j-th column removed, $\boldsymbol{\gamma}^{(j)} \in \mathbb{R}^{p-1}$. The corresponding vector $\boldsymbol{u}_0$ is $\boldsymbol{u}_j = (\boldsymbol{\Sigma}^{-1})_{jj}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$ which is related to $\boldsymbol{\gamma}^{(j)}$ by $(\boldsymbol{u}_j)_j = 1$ and $(\boldsymbol{u}_j)_{(-j)} = -\boldsymbol{\gamma}^{(j)}$. The ideal score vector $\boldsymbol{z}_0$ becomes $\boldsymbol{z}_j = (\boldsymbol{\Sigma}^{-1})_{jj}^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$ and has iid $N(0, (\boldsymbol{\Sigma}^{-1})_{jj}^{-1})$ entries independent of $\boldsymbol{X}^{(-j)}$. For a given estimator $\widehat{\boldsymbol{\gamma}}^{(j)}$ of $\boldsymbol{\gamma}^{(j)}$, the score vector (2.18) is then $\widehat{\boldsymbol{z}}_j = \boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}^{(-j)}\widehat{\boldsymbol{\gamma}}^{(j)}$ and the de-biased estimate (2.19) reduces to

$$\hat{\beta}_j^{(\text{de-biased})} = (\hat{\boldsymbol{\beta}}^{(\text{lasso})})_j + \frac{\langle \widehat{\boldsymbol{z}}_j, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\text{lasso})}\rangle}{(1 - \nu/n)\langle \widehat{\boldsymbol{z}}_j, \boldsymbol{X}\boldsymbol{e}_j\rangle}. \tag{2.29}$$

which corresponds to the proposal in [31] modified with the degrees-of-freedom adjustment $(1 - \nu/n)$. For $\boldsymbol{a}_0 = \boldsymbol{e}_j$, the Lasso estimator (2.22) becomes

$$\widehat{\boldsymbol{\gamma}}^{(j)} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\arg\min}\left\{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}^{(-j)}\boldsymbol{\gamma}\|_2^2 + \hat{\tau}_j\overline{\lambda}\|\boldsymbol{\gamma}\|_1\right\}. \tag{2.30}$$

with recursive solution $\hat{\tau}_j = \|\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}^{(-j)}\boldsymbol{\gamma}\|_2/n^{1/2}$ in the scaled Lasso [19] or any estimate $\hat{\tau}_j$ satisfying $1 + o_{\mathbb{P}}(1) \leq (\boldsymbol{\Sigma}^{-1})_{j,j}\hat{\tau}_j^2 \leq O_{\mathbb{P}}(1)$. As the choice of $\boldsymbol{u}$ in (2.20) for $\boldsymbol{a}_0 = \boldsymbol{e}_j$ is $\boldsymbol{u} = \boldsymbol{e}_j$, the proof of Theorem 2.5 can be modified to allow $\overline{\lambda} = A\lambda_{\text{univ}}$ with $A > 1$, since in this case $\mathbb{E}\|\boldsymbol{X}\boldsymbol{Q}\boldsymbol{a}_0\|_2^2/n$ is bounded from the above by 1.

Corollary 2.7. *Assume that $\boldsymbol{\Sigma}_{jj} \leq 1$ for all $j \in [p]$ and that the spectrum of $\boldsymbol{\Sigma}$ is uniformly bounded away from 0 and ∞; e.g. $\max(\|\boldsymbol{\Sigma}\|_{op}, \|\boldsymbol{\Sigma}^{-1}\|_{op}) \leq 2$. Let $\lambda = 1.01\sigma\sqrt{2\log(8p/s_0)/n}$ for the Lasso (2.1). Consider the Scaled Lasso in (2.30) with $\overline{\lambda} = 1.01\sqrt{2\log(p)/n}$, the corresponding score vector $\widehat{\boldsymbol{z}}_j$ and de-biased estimate $\widehat{\beta}_j^{(\text{de-biased})}$ in (2.29) with $\nu = \|\hat{\boldsymbol{\beta}}^{(\text{lasso})}\|_0$. Then for any j,*

$$\frac{(\|\boldsymbol{\beta}\|_0 \vee \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0)\log(p)}{n} \to 0 \text{ and } \frac{(\|\boldsymbol{\beta}\|_0 \wedge \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0)\log(p)}{\sqrt{n}} \to 0 \tag{2.31}$$

implies $\sqrt{n}(\boldsymbol{\Sigma}^{-1})_{jj}^{-1/2}(\widehat{\beta}_j^{(\text{de-biased})} - \beta_j) \to^d N(0, \sigma^2)$.

Remark 2.1. The tuning parameters of the present section are chosen as $\lambda = 1.01\sigma\sqrt{2\log(8p/s_0)/n}$ for simplicity of the presentation. As the results of the present section are consequences of Theorem 3.1 in the next section, more general tuning parameters of the form (3.4) are also allowed and the resulting constants in the theorems would then depend on certain constants $\eta_2 \in (0,1)$, $\eta_3 > 0$.

3. Theoretical results for known Σ

In this section, we prove that the degrees-of-freedom adjusted LDPE in (2.2) indeed removes the bias of the Lasso for the estimation of a general linear functional $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ when $(s_0/n)\log(p/s_0)$ is sufficiently small and a sparse Riesz condition (SRC) [30] holds on the population covariance matrix $\boldsymbol{\Sigma}$ of the Gaussian design.

The SRC is closely related to the restricted isometry property (RIP) [11,12]. While the RIP is specialized for nearly uncorrelated design variables in the context of compressed sensing, the SRC is more suitable in analysis of data from observational studies or experiments with higher correlation in the design. For example, the SRC allows an upper sparse eigenvalue greater than 2. For $p \times p$ positive semi-definite matrices $\boldsymbol{M}$, integers $1 \le m \le p$ and a support set $B \subset \{1, \ldots, p\}$, define a lower sparse eigenvalue as

$$\phi_{\min}(m, B; \boldsymbol{M}) = \min_{A \subset [p]: |A \setminus B| = m} \phi_{\min}\Big(\boldsymbol{M}_{A,A}\Big) \tag{3.1}$$

and an upper sparse eigenvalue as

$$\phi_{\max}(m, B; \boldsymbol{M}) = \max_{A \subset [p]: |A \setminus B| = m} \phi_{\max}\Big(\boldsymbol{M}_{A,A}\Big), \tag{3.2}$$

where $\phi_{\min}(\boldsymbol{M})$ and $\phi_{\max}(\boldsymbol{M})$ are respectively the smallest and largest eigenvalues of symmetric matrix $\boldsymbol{M}$. Define similarly the sparse condition number by

$$\phi_{\mathrm{cond}}(m; B, \boldsymbol{M}) = \max_{A \subset [p]: |A \setminus B| \le (1 \vee m)} \big\{\phi_{\max}(\boldsymbol{M}_{A,A})/\phi_{\min}(\boldsymbol{M}_{A,A})\big\}. \tag{3.3}$$

Recall that S is the support of $\boldsymbol{\beta}$ and $s_0 = |S|$. For a precise statement of the sample size requirement for our main results, we will assume the following.

Assumption 3.1. Assume that $\boldsymbol{\Sigma}$ is invertible with diagonal elements at most 1, i.e., $\max_{j=1,\ldots,p} \Sigma_{jj} \le 1$. Consider positive integers $\{m, n, p, k\}$ and positive constants $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ with $\eta_2, \eta_3 \in (0,1)$. Set the tuning parameter of the Lasso by

$$\lambda = \eta_2^{-1}(1+\eta_3)\sigma\lambda_0, \qquad \text{where} \qquad \lambda_0 = \sqrt{(2/n)\log(8p/k)}. \tag{3.4}$$

Define $\{\tau_*, \tau^*\}$ by $\tau_* = (1-\epsilon_1-\epsilon_2)^2$, $\tau^* = (1+\epsilon_1+\epsilon_2)^2$ and assume that

$$s_0 + k < \frac{(1-\eta_2)^2 2m}{(1+\eta_2)^2\big\{(\tau^*/\tau_*)\phi_{\mathrm{cond}}(m+k; S, \boldsymbol{\Sigma}) - 1\big\}} \tag{3.5}$$

and $\rho_* \le \phi_{\min}(m+k, S; \boldsymbol{\Sigma})$ hold. Assume that $\lambda_0\sqrt{s_*} \le 1$ where $s_* = s_0 + m + k$, as well as

$$2(m+k) + s_0 + 1 \le (n-1) \wedge (p+1), \tag{3.6}$$

$$\epsilon_1 + \epsilon_2 < 1, \quad \epsilon_3 + \epsilon_4 = \epsilon_2^2/8, \tag{3.7}$$

$$s_0 + m + k + 1 \le \min(p+1, \epsilon_1^2 n/2), \quad \log\binom{p-s_0}{m+k} \le \epsilon_3 n. \tag{3.8}$$

Typical values of k, m and $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ are given after Corollary 3.2 below. As will become clear in the proofs in Appendix A of the supplement [2], the integer k above is an upper bound on the cardinality of the set

$$B = \{j \in [p] : |\boldsymbol{\varepsilon}^\top \boldsymbol{x}_j|/n \ge \eta_2 \lambda\}, \tag{3.9}$$

i.e., the set of covariates that correlate highly with the noise. If $k = 1$ then $\lambda = \eta_2^{-1}(1+\eta_3) \times \sigma\sqrt{(2/n)\log(8p)}$ and the set B is empty with high probability. The integer m is, with high probability, an upper bound on the cardinality of the set $\mathrm{supp}(\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}) \setminus (S \cup B)$. In other words, the support of $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$ contains at most m variables that are neither in the true support S nor in the set B of highly correlated covariates. These statements are made rigorous in Appendices A.1 and A.2 of the supplement [2]. Results of the form $|\widehat{S}| = O_{\mathbb{P}}(s_0)$ have appeared before for the Lasso, see for instances [30], Theorem 1, [7], Eq. (7.9), [32], Corollary 2 (ii), and [5], Theorem 3. Among these existing bounds, the theory derived in the present paper is closest to [32], Corollary 2 (ii), where a bound of the form $|\widehat{S}| = O_{\mathbb{P}}(s_0)$ is derived under a condition on the upper sparse eigenvalue (3.2) after a prediction error bound under a weak restricted eigenvalue condition. They depart from other existing bounds of the form $|\widehat{S}| = O_{\mathbb{P}}(s_0)$ in several ways. The bounds in [30], Theorem 1, requires the tuning parameter to be set as a function of the sparse eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}/n$. The bound from [7] involves $\phi_{\max}(\boldsymbol{X}^\top \boldsymbol{X}/n)$ which is unbounded if $p/n \to +\infty$ for Gaussian designs. The bound [5], Theorem 3, tackles tuning parameters larger than $\sigma\sqrt{2\log(p)/n}$ but does not provide guarantees for smaller tuning parameters of order $\sigma\sqrt{2\log(8p/k)/n}$. The theory developed for the present paper in Appendix A of the supplement [2] improves upon these aforementioned references: The theory only requires bounds on sparse condition number (cf. the SRC condition (3.5)), the tuning parameters need not depend on the sparse eigenvalues, and small tuning parameters of order $\sigma\sqrt{2\log(8p/k)/n}$ are allowed. Furthermore, the theory in Appendix A of the supplement [2] clearly separates the roles of s_0, k and m: k is an upper bound on the cardinality of the set (3.9) of covariates highly correlated with the noise, m is an upper bound on $\mathrm{supp}(\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}) \setminus (S \cup B)$, and consequently $\|\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}\|_0 \le s_0 + k + m$.

Stochastically bounded $O_{\mathbb{P}}(\cdot)$ notation. In the following results, we consider an asymptotic regime with growing $\{s_0, m, k, n, p\}$ such that

$$p/k \to +\infty, \qquad s_* \lambda_0^2 \to 0 \tag{3.10}$$

where $s_* = s_0 + m + k$. This means that we consider a sequence of regression problems (1.1) indexed by n and $\{s_0, m, k, p\}$ are functions of n such that (3.10) holds and Assumption 3.1 is satisfied for all n with constants $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ independent of n. For a deterministic sequence a_n, we write $W_n = O_{\mathbb{P}}(a_n)$ if the sequence of random variables (W_n) is such that for any arbitrarily small $\gamma > 0$, there exists constants K, N depending on γ and $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ such that for all $n \ge N$, $\mathbb{P}(W_n > K) \le \gamma$. We also write $W_n = o_{\mathbb{P}}(1)$ if $W_n = O_{\mathbb{P}}(a_n)$ for some $a_n \to 0$. Under the above Assumption 3.1, our main result is the following.

Theorem 3.1. *Let (3.10) and Assumption 3.1 be fulfilled. Let $F_\theta = 1/(\sigma C_0)^2$ be the Fisher information as in (1.7), and $T_n = \sqrt{n F_\theta}\langle z_0, \boldsymbol{\varepsilon}\rangle/\|z_0\|_2^2$ so that T_n has the t-distribution with n degrees of freedom. For any random degrees-of-freedom adjustment $\nu \in [0, n]$ we have*

$$\sqrt{n F_\theta}(1 - \nu/n)(\widehat{\theta}_\nu - \theta) = T_n + \sqrt{F_\theta/n}\left\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathit{lasso})} - \boldsymbol{\beta}\right\rangle \left(|\widehat{S}| - \nu\right) + O_{\mathbb{P}}\left(\lambda_0 \sqrt{s_*}\right).$$

If the condition number $\phi_{\mathrm{cond}}(p;\varnothing,\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}\|_{op}\|\boldsymbol{\Sigma}^{-1}\|_{op}$ *of the population covariance matrix* $\boldsymbol{\Sigma}$ *is bounded, then* $O_{\mathbb{P}}\left(\lambda_0\sqrt{s_*}\right)$ *above can be replaced by* $O_{\mathbb{P}}\left(\lambda_0\sqrt{s_0+k}\right)$ *[by* $O_{\mathbb{P}}\left(\lambda_0\sqrt{s_0}\right)$ *when the penalty is chosen with* $k \lesssim s_0$ *in* (3.4)*]*.

The result is proved in Section 8.4. If $\lambda_0\sqrt{s_*} \to 0$ and $k/p \to 0$, the above result implies that $\sqrt{nF_\theta}(1-\nu/n)(\widehat{\theta}_\nu - \theta)$ is within $o_{\mathbb{P}}(1)$ of T_n of the t-distribution with n degrees of freedom if and only if

$$\sqrt{F_\theta/n}\left\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta}\right\rangle \left(|\widehat{S}| - \nu\right) = o_{\mathbb{P}}(1). \tag{3.11}$$

The left hand side of (3.11) is negligible either because the modified de-biasing scheme (2.2) is correctly adjusted with $\nu = |\widehat{S}|$ (or $\nu \approx |\widehat{S}|$) to account for the degrees of freedom of the initial estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$, or because the estimation error of the initial estimator $\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta}\rangle$ is significantly small.

The choice of degrees-of-freedom adjustment $\nu = |\widehat{S}|$ ensures that the quantity (3.11) is always equal to 0. This leads to the following corollary.

Corollary 3.2. *Let* (3.10) *and Assumption* 3.1 *be fulfilled. With the notation from Theorem* 3.1, *if* $\nu = |\widehat{S}|$ *then*

$$\sqrt{nF_\theta}\left(1 - |\widehat{S}|/n\right)\left(\widehat{\theta}_{\nu=|\widehat{S}|} - \theta\right) = T_n + O_{\mathbb{P}}\left(\lambda_0\sqrt{s_*}\right). \tag{3.12}$$

Hence if $\lambda_0\sqrt{s_*} \to 0$ and $k/p \to 0$, the de-biasing scheme (2.2) correctly adjusted with $\nu = |\widehat{S}|$ enjoys asymptotic efficiency. To highlight this fact and give an example of typical values for m, k and $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ in Assumption 3.1, let us explain how Corollary 3.2 implies (2.3) of Theorem 2.1. Set $\eta_2^{-1} = \sqrt{1.01}$, $\eta_3 = \sqrt{1.01} - 1$ and $k = s_0$, so that the tuning parameter (3.4) is equal to λ defined in Theorem 2.1. Set also $\epsilon_1 = \epsilon_2 = 1/4$ so that $\tau_* = 1/4, \tau^* = 9/4$. Under the assumptions of Theorem 2.1, the spectrum of $\boldsymbol{\Sigma}$ is bounded away from 0 and ∞ (e.g. a subset of $[1/2, 2]$) and the sparse condition number appearing in (3.5) is bounded (e.g. at most 4 respectively). Next, set $m = Cs_0$ for some large enough absolute constant $C > 0$ chosen so that (3.5) holds; this gives $s_* = s_0 + m + k = (C+2)s_0$. The conditions in Assumption 3.1 are satisfied thanks to $\lambda_0\sqrt{s_*} \to 0$ and $k/p \to 0$. By Lemma 8.1 we get $|\widehat{S}| = O_{\mathbb{P}}(s_0)$. Then (2.3) is a direct consequence of (3.12).

By Theorem 3.1, the unadjusted de-biasing scheme (1.8) enjoys asymptotic efficiency for all fixed $\boldsymbol{a}_0$ and $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_0 \le s_0$ if and only if (3.11) holds with $\nu = 0$, i.e., if

$$\sqrt{F_\theta/n}\left\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta}\right\rangle |\widehat{S}| = o_{\mathbb{P}}(1). \tag{3.13}$$

By the Cauchy-Schwarz inequality, $|\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta}\rangle| \le C_0\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta})\|_2$. Under Assumption 3.1 or other typical conditions on the restricted eigenvalues of $\boldsymbol{\Sigma}$ and the sample size, the population risk $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta})\|_2$ is of order $O_{\mathbb{P}}(\sigma\lambda_0\sqrt{s_*})$ which grants (3.13) if $\lambda_0\sqrt{s_*}s_*/\sqrt{n} \to 0$. This is the content of the following corollary which is formally proved in Section 8.5.

Corollary 3.3 (Unadjusted LDPE). *Let* (3.10) *and Assumption* 3.1 *be fulfilled. With the notation from Theorem* 3.1, *if* $\nu = 0$ *then*

$$\sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) = T_n + O_{\mathbb{P}}\left(\lambda_0\sqrt{s_*}\left(1 + \frac{s_*}{\sqrt{n}}\right)\right). \tag{3.14}$$

If $\lambda_0^2(s_*)^3/n \to 0$ then the right hand side of (3.14) converges in probability to T_n. In this asymptotic regime, the degrees-of-freedom adjustment is not necessary and the unadjusted (1.8) enjoys asymptotic efficiency. Note that although the adjustment $\nu = |\widehat{S}|$ that leads to the efficiency of $\widehat{\theta}_\nu$ in Corollary 3.2 is not necessary in this particular asymptotic regime, such adjustment does not harm either. Since the practitioner cannot establish whether the asymptotic regime $\lambda_0^2(s_*)^3/n \to 0$ actually occurs because s_0 and s_* are unknown, it is still recommended to use the adjustment $\nu = |\widehat{S}|$ as in Corollary 3.2 to ensure efficiency for the whole range of sparsity.

An outcome of Theorem 2.3 is that the unadjusted de-biasing scheme (1.8) cannot be efficient in the regime (2.15). By Theorem 2.3 and the discussion surrounding (2.15) on the one hand, and Corollary 3.3 and the discussion of the previous paragraph on the other hand, we have established the following phase transition:

- If $\lambda_0^2(s_*)^3/n \lll 1$, the unadjusted de-biasing scheme (1.8) is efficient for every $\boldsymbol{a}_0$, by Corollary 3.3.
- If $\lambda_0^2 s_0^3/n \ggg 1$, the unadjusted de-biasing scheme (1.8) cannot be efficient for certain specific $\boldsymbol{a}_0$.

In other words, there is a phase transition at $s_* \asymp n^{2/3}$ (up to a logarithmic factor) where degrees-of-freedom adjustment becomes necessary to achieve asymptotic efficiency for all preconceived directions $\boldsymbol{a}_0$. Condition $s_* \lll n^{2/3}$ is a weaker requirement than the assumption $s_* \lll \sqrt{n}$ commonly made in the literature on de-biasing.

4. De-biasing without degrees of freedom adjustment under additional assumptions on $\boldsymbol{\Sigma}$

The left hand side of (3.13) quantifies the remaining bias of the unadjusted de-biasing scheme (1.8). Under an additional assumption on $\boldsymbol{\Sigma}$, namely a bound on $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1$, the initial bias of the Lasso $\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\rangle$ is small enough to grant asymptotic efficiency to the unadjusted de-biasing scheme (1.8). The following theorem makes this precise.

Theorem 4.1. *Let* (3.10) *and Assumption* 3.1 *be fulfilled. Suppose*

$$\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 \le K_{0,n,p} = K_{1,n,p}\sqrt{n/s_*} \tag{4.1}$$

for some quantities $K_{0,n,p}$ and $K_{1,n,p}$. Then, $\sqrt{F_\theta}|\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\rangle| = O_{\mathbb{P}}(\lambda_0 K_{0,n,p})$ and

$$\sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) = T_n + O_{\mathbb{P}}\big((1 + K_{1,n,p})\lambda_0\sqrt{s_*} + s_*/n\big).$$

This implies that $\sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) = T_n + o_{\mathbb{P}}(1)$ when $K_{1,n,p} = O(1)$.

The proof is given in Appendix B of the supplement [2]. In other words, the unadjusted de-biasing scheme (1.8) is efficient and degrees-of-freedom adjustment is not needed for efficiency if the ℓ_1 norm of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ is bounded from above as in

$$\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = O(\sqrt{n/s_*})$$

with $s_*/p \to 0$ and $(s_*/n)\log(p/s_*) \to 0$. This improves by a logarithmic factor the condition $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = O(1)$ required for efficiency in [15].

The above result explains why the necessity of degrees-of-freedom adjustment did not appear in previous analysis such as [15]; $\sqrt{F_\theta}|\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\rangle| = O_{\mathbb{P}}(\lambda_0)$ when $K_{0,n,p} = O(1)$ in (4.1), and the

unadjusted de-biasing scheme (1.8) is efficient when $K_{1,n,p} = O(1)$ in (4.1). However, by Theorem 2.3 and the discussion surrounding (2.15), there exist certain $\boldsymbol{a}_0$ with large $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2$ such that the unadjusted de-biasing scheme cannot be efficient. For such $\boldsymbol{a}_0$, degrees-of-freedom adjustments are necessary to achieve efficiency.

5. An ℓ_∞ error bound for the lasso

The idea of the previous section can be applied to $\boldsymbol{a}_0 = \boldsymbol{e}_j$ simultaneously for all vectors $\boldsymbol{e}_j$ of the canonical basis $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$. This yields the following ℓ_∞ bound on the error of the Lasso.

Theorem 5.1. *Let Assumption 3.1 be fulfilled, and further assume that* $\log p < n$. *Then the Lasso satisfies simultaneously for all* $j = 1, \ldots, p$

$$\left|\widehat{\beta}_j^{(lasso)} - \beta_j\right| \le \frac{M_5^2\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1\lambda + \sigma\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2\sqrt{\log p/n}\left(2M_5 + 3\bar{M}\lambda_0\sqrt{s_*}\right)}{1 - s_*/n} \tag{5.1}$$

on an event Ω_{ℓ_∞} *such that* $\mathbb{P}(\Omega_{\ell_\infty}^c) \to 0$ *when* (3.10) *holds, where* $s_* = s_0 + m + k$, $M_5 = 1/(1-\eta_3)$ *and* $\bar{M}$ *is a constant that depends on* $\{\rho_*, \eta_2, \eta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ *only. Consequently, since* $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2 \le \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1$, *on the same event we have*

$$\|\hat{\boldsymbol{\beta}}^{(lasso)} - \boldsymbol{\beta}\|_\infty \le \rho(\boldsymbol{\Sigma})\left(\frac{M_5^2 + 2M_5 + 4\bar{M}\lambda_0\sqrt{s_*}}{1 - s_*/n}\right)\max\left(\lambda, \sigma\sqrt{\frac{\log p}{n}}\right)$$

where $\rho(\boldsymbol{\Sigma}) = \max_{j=1,\ldots,p}\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1$.

The proof is given in Appendix B of the supplement [2]. The above result asserts that if the ℓ_1-norms of the columns of $\boldsymbol{\Sigma}^{-1}$ are bounded from above by some constant $\rho(\boldsymbol{\Sigma}) > 0$ then

$$\|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_\infty \le C(\boldsymbol{\Sigma})\max(\lambda, \sigma\sqrt{\log(p)/n})$$

holds with overwhelming probability for some constant $C(\boldsymbol{\Sigma}) \lesssim \rho(\boldsymbol{\Sigma})$.

Although some ℓ_∞ bounds for the lasso have appeared previously in the literature, we are not aware of previous results comparable to Theorem 5.1 for $s_0 \gg \sqrt{n}$. The result of [16] and [3], Theorem 2(2), requires incoherence conditions on the design, i.e., that non-diagonal elements of $\boldsymbol{X}^\top\boldsymbol{X}/n$ are smaller than $1/s_0$ up to a constant. This assumption is strong and cannot be satisfied in the regime $s_0 \gg \sqrt{n}$, even for the favorable $\boldsymbol{\Sigma} = \boldsymbol{I}_p$: for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ the standard deviation of the i,j-th entry is $\mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X}/n)_{ij}^2]^{1/2} = 1/\sqrt{n}$. In a random design setting comparable to ours, Section 4.4 of [22] explains that $\|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_\infty \lesssim \max_j\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1\sigma\sqrt{\log(p)/n}(1 + \|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_1/\sigma)$. This bound is only comparable to our ℓ_∞ bound in the regime $\|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_1 = O_P(1)$, i.e., in the regime $s_0 \lesssim \sqrt{n}$ (up to logarithmic factors) since $\|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_1 \approx \lambda s_0 \approx \sigma s_0\sqrt{\log(p)/n}$. Again this result is not applicable (or substantially worse than Theorem 5.1) in the more challenging regime $s_0 \gg \sqrt{n}$ of interest here.

6. Regularity and asymptotic efficiency

Theorem 2.1(i) shows that the test statistic $\sqrt{nF_\theta}(1 - |\widehat{S}|/n)(\hat{\theta}_\nu - \boldsymbol{a}_0^\top\boldsymbol{\beta})$, properly adjusted with $\nu = |\widehat{S}|$, converges in distribution to $N(0,1)$, where $F_\theta = 1/\{\sigma^2 C_0^2\}$ and $C_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|$. This

holds under any sequence of distributions $\{\mathbb{P}_0^n\}_{n\geq 1}$ defined by $\|\boldsymbol{\beta}\| = s_0$, $s_0 \log(p/s_0)/n \to 0$, $\max(\|\boldsymbol{\Sigma}\|_{op}, \|\boldsymbol{\Sigma}^{-1}\|_{op}) \leq K$ for some constant K independent of n, p, and

$$\boldsymbol{X} \text{ has iid rows } N(\boldsymbol{0}, \boldsymbol{\Sigma}), \qquad \boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

Here, we denote the unknown parameter $\boldsymbol{a}_0^\top \boldsymbol{\beta}$ by $\theta(\boldsymbol{P}_0^n)$ to avoid confusion with the probability measures defined in the next paragraph. By Slutsky's theorem, since $|\widehat{S}|/n$ converges to 0 in probability by Theorem 2.1(i), we have

$$\mathcal{L}\left(\sqrt{nF_\theta}(\hat{\theta}_v - \theta(\mathbb{P}_0^n)) \,;\, \mathbb{P}_0^n\right) \to N(0,1). \tag{6.1}$$

Given $\boldsymbol{a}_0 \in \mathbb{R}^p$, a positive-definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\mathcal{B}_n \subset \mathbb{R}^p$ as a parameter space, let F_θ be as in (1.7),

$$\mathcal{U}_n \subseteq \left\{\boldsymbol{u} \subset \mathbb{R}^p : \boldsymbol{u}^\top \boldsymbol{a}_0 = 1,\ \boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta} \in \mathcal{B}_n\ \forall t \in [0, t_{\boldsymbol{u}}],\ t_{\boldsymbol{u}} \to \infty\right\}$$

as a collection of directions of univariate sub-models $\{\boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta} : 0 \leq t \leq t_{\boldsymbol{u}}\}$. For $t > 0$ and $\boldsymbol{u} \in \mathcal{U}_n$ let $\mathbb{P}_{t,\boldsymbol{u}}^n$ be probabilities under which

$$\boldsymbol{y}|\boldsymbol{X} \sim N\left(\boldsymbol{X}(\boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta}), \sigma^2 \boldsymbol{I}_n\right) \tag{6.2}$$

(for either deterministic or possibly non-Gaussian random $\boldsymbol{X}$) and

$$\theta(\mathbb{P}_{t,\boldsymbol{u}}^n) = \left\langle \boldsymbol{a}_0, \boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta}\right\rangle = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle + t/\sqrt{nF_\theta}.$$

That is, under $\mathbb{P}_{t,\boldsymbol{u}}^n$ the vector $\boldsymbol{\beta}$ is perturbed with the additive term $t\boldsymbol{u}/\sqrt{nF_\theta}$, resulting a perturbation of the parameter of interest with $t/\sqrt{nF_\theta}$. In the above framework, an estimator $\widetilde{\theta}$ is regular (in the directions $\boldsymbol{u} \in \mathcal{U}_n$) if

$$\mathcal{L}\left(\sqrt{nF_\theta}(\widetilde{\theta} - \theta(\mathbb{P}_{t,\boldsymbol{u}}^n)) \,;\, \mathbb{P}_{t,\boldsymbol{u}}^n\right) \to G \tag{6.3}$$

for all fixed $t > 0$ and $\boldsymbol{u} \in \mathcal{U}_n$ and some distribution G not depending on t and $\boldsymbol{u}$. That is, the limiting distribution is stable under the small perturbation as defined above.

Our first task is to show that $\widehat{\theta}_v$ is regular in all directions with the same limiting distribution as in (6.1), i.e. (6.3) holds with $\mathcal{U}_n = \mathbb{R}^p$ and $G \sim N(0,1)$. For $t = 0$, (6.1) is implied by Theorem 2.1(i). However Theorem 2.1(i) does not directly imply (6.3) for $t \neq 0$ because $\boldsymbol{u} \in \mathcal{U}_n$, as well as the unknown regression vector $\boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta}$ under $\mathbb{P}_{t,\boldsymbol{u}}^n$, may not be sparse. The following device due to Le Cam shows that (6.3) still holds with the perturbation $t\boldsymbol{u}/\sqrt{nF_\theta}$ for any fixed $t \neq 0$ independent of n, p.

The likelihood-ratio L_n between $\mathbb{P}_{t,\boldsymbol{u}}^n$ and $\mathbb{P}_0^n$ is given by

$$\begin{aligned}\log L_n &= \{-\|\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta})\|^2 + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\}/(2\sigma^2).\\ &= -t^2 C_0^2 \|\boldsymbol{X}\boldsymbol{u}\|^2/(2n) - \langle \boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{u}\rangle t C_0/(\sigma\sqrt{n}).\end{aligned}$$

Under $\mathbb{P}_0^n$, the random variable $\sqrt{nF_\theta}(\widehat{\theta}_v - \theta(\mathbb{P}_0^n))$ can be written as $\langle \boldsymbol{\varepsilon}, \boldsymbol{z}_0\rangle C_0/(\sqrt{n}\sigma) + o_{\mathbb{P}}(1)$ so that the vector $(\sqrt{nF_\theta}(\widehat{\theta}_v - \theta(\mathbb{P}_0^n)), \log L_n)^\top$ converges in distribution under $\mathbb{P}_0^n$ to a bivariate normal vector with mean $(0, -t^2 C_0^2 \langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle/2)^\top$ and covariance

$$\begin{pmatrix} 1 & tC_0^2\langle \boldsymbol{u}_0, \boldsymbol{\Sigma}\boldsymbol{u}\rangle \\ tC_0^2\langle \boldsymbol{u}_0, \boldsymbol{\Sigma}\boldsymbol{u}\rangle & t^2 C_0^2 \langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle \end{pmatrix} = \begin{pmatrix} 1 & t \\ t & t^2 C_0^2 \langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle \end{pmatrix},$$

where the equality is due to $\boldsymbol{u}_0 = C_0^{-2}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ and $\langle \boldsymbol{a}_0, \boldsymbol{u}\rangle = 1$. It directly follows by Le Cam's third lemma (see, for instance, [25], Example 6.7) that $\sqrt{nF_\theta}(\widehat{\theta}_\nu - \theta(\mathbb{P}_0^n))$ converges to $N(t,1)$ under $\{\mathbb{P}_{t,\boldsymbol{u}}^n\}_{n\ge 1}$ and that (6.3) holds. For more details, see also [25], Section 7.5, about situations where the log-likelihood ratio converges to normal distributions of the form $N(-a^2/2, a^2)$.

Hence, properly adjusted with $\nu = |\widehat{S}|$, the estimator $\hat{\theta}_\nu$ is regular and asymptotic normality still holds if the sparse coefficient vector $\boldsymbol{\beta}$ is replaced by $\boldsymbol{\beta} + t\boldsymbol{u}/\sqrt{nF_\theta}$ for constant $t \in \mathbb{R}$, even if the perturbation $\boldsymbol{u}$ is non-sparse. By the Le Cam-Hayek convolution theorem (see, for instance, [25], Theorem 8.8), the asymptotic variance of $\sqrt{n}(\hat{\theta}_\nu - \theta)$ must be at least $1/F_\theta$ and our estimator $\widehat{\theta}_\nu$ is efficient, i.e., it achieves the smallest possible asymptotic variance among regular estimators.

Note that the above reasoning does not inherently rely on the Gaussian design assumption. As soon as the second moment of the row of $\boldsymbol{X}$ exists, $\|\boldsymbol{X}\boldsymbol{u}\|^2/(n\langle \boldsymbol{u}, \boldsymbol{\Sigma}\boldsymbol{u}\rangle) \to 1$ and $\langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{X}\boldsymbol{u}_0\rangle/(n\langle \boldsymbol{u}_0, \boldsymbol{\Sigma}\boldsymbol{u}\rangle) \to 1$ almost surely by the law of large numbers. If additionally $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{X}$ is such that $\sqrt{nF_\theta}(\hat{\theta}_\nu - \boldsymbol{a}_0^\top\boldsymbol{\beta}) = \langle \boldsymbol{\varepsilon}, z_0\rangle C_0/(\sigma\sqrt{n}) + o_{\mathbb{P}_0^n}(1)$ for sparse $\boldsymbol{\beta}$, the argument of the previous paragraph is applicable and $\hat{\theta}_\nu$ is regular in the sense of (6.3). For instance, if $\boldsymbol{a}_0$ is a canonical basis vector, equation $\sqrt{nF_\theta}(\hat{\theta}_{\nu=0} - \boldsymbol{a}_0^\top\boldsymbol{\beta}) = \langle \boldsymbol{\varepsilon}, z_0\rangle C_0/(\sigma\sqrt{n}) + o_{\mathbb{P}_0^n}(1)$ can be obtained for sub-gaussian design and $s_0 \lll \sqrt{n}$ using an ℓ_1/ℓ_∞ duality inequality, cf. [13,24,31]. In such asymptotic regime, the argument of the previous paragraph shows that $\hat{\theta}_{\nu=0}$ is stable for non-sparse perturbations of the form $t\boldsymbol{u}/\sqrt{nF_\theta}$.

We formally state the above analysis and existing lower bounds,

Proposition 6.1. *Let $\mathcal{V}_n$ be the linear span of $\mathcal{U}_n$ as a tangent space. Suppose*

$$\mathbb{P}_0^n\left\{\left|\|\boldsymbol{X}\boldsymbol{u}\|_2^2/(n\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u}) - 1\right| > \epsilon\right\} = o(1), \qquad \boldsymbol{u} \in \mathcal{U}_n,$$

and $\dim(\mathcal{V}_n) = O(1)$. Let $\boldsymbol{u}_0$ be as in (1.6) and $\tau = \tau(\mathcal{V}_n) = \widetilde{\boldsymbol{u}}_0^\top\boldsymbol{\Sigma}\widetilde{\boldsymbol{u}}_0/F_\theta$ with

$$\widetilde{\boldsymbol{u}}_0 = \arg\min\left\{\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u} : \boldsymbol{u} \in \mathcal{V}_n, \langle \boldsymbol{a}_0, \boldsymbol{u}\rangle = 1\right\}.$$

(i) Let $\widetilde{\theta}$ be a regular estimator in the sense of (6.3) with a limiting distribution G. Let $\xi \sim G$. Then, (a) $\mathrm{Var}(\xi) \ge 1/\tau$; (b) If $\mathrm{Var}(\xi) = 1/\tau$, then $\xi \sim N(0, 1/\tau)$; (c) If $\widetilde{\boldsymbol{u}}_0 = a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2$ for two $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathcal{U}_n$ and $\{a\boldsymbol{u}_1 + (1-a)\boldsymbol{u}_2 : 0 \le a \le 1\} \subseteq \mathcal{U}_n$, then $\xi = \xi_1 + \xi_2$ where $\xi_1 \sim N(0, 1/\tau)$ and ξ_2 is independent of ξ_1.

(ii) If $\boldsymbol{u}_0 \in \mathcal{V}_n$, then $\widetilde{\boldsymbol{u}}_0 = \boldsymbol{u}_0$ and $\tau = \tau(\mathcal{V}_n) = 1$.

(iii) If (6.1) holds, then $\widehat{\theta}_\nu$ is regular and locally asymptotically efficient in the sense of (6.3) with $\mathcal{B}_n = \mathcal{U}_n = \mathbb{R}^p$.

The above statement is somewhat more general than the usual version as we wish to accommodate general parameter space $\mathcal{B}_n$, cf. [25], Theorem 8.8, for $\mathcal{U}_n = \{\boldsymbol{u} \in \mathcal{V}_n : \langle \boldsymbol{a}_0, \boldsymbol{u}\rangle = 1\}$ and [18] and [27], Theorem 6.1, for general $\mathcal{U}_n$. We note that the condition on $\widetilde{\boldsymbol{u}}_0$ in Proposition 6.1(i)(c), known as the convolution theorem, is equivalent to the convexity of $\mathcal{U}_n$ and $\widetilde{\boldsymbol{u}}_0 \in \mathcal{V}_n$. The minimum Fisher information is sometimes defined as $\min\{\sigma^{-2}\boldsymbol{u}^\top\boldsymbol{\Sigma}\boldsymbol{u} : \langle \boldsymbol{a}_0, \boldsymbol{u}\rangle = 1, \boldsymbol{u} \in \mathcal{U}_n\}$. However, when this minimum over $\mathcal{U}_n$ is strictly larger than the minimum over its linear span $\mathcal{V}_n$, the larger minimum information is not attainable by estimators regular with respect to $\mathcal{U}_n$ in virtue of (i)(a) above.

In Proposition 6.1, the parameter $\tau = \tau(\mathcal{V}_n)$ can be viewed as the relative efficiency for the tangent space $\mathcal{V}_n$ generated by the collection $\mathcal{U}_n$ of directions of univariate sub-models. As the minimization for $\widetilde{\boldsymbol{u}}_0$ is taken over no greater a space compared with (1.6), $\tau \ge 1$ always holds. When the parameter space $\mathcal{B}_n$ is strictly smaller than $\mathbb{R}^p$ or the regularity (stability of the limiting distribution) is required only for deviations from the true $\boldsymbol{\beta}$ in a small collection of directions, $\tau > 1$ may materialize and

an estimator regular and efficient relative to $\mathcal{U}_n$ would become super-efficient in the full model with $\mathcal{B}_n = \mathcal{V}_n = \mathbb{R}^p$. According to Le Cam's local asymptotic minimax theorem, in the full model, such a super-efficient estimator would perform strictly worse than a regular efficient estimator when the true $\boldsymbol{\beta}$ is slightly perturbed in a certain direction.

The super-efficiency was observed in [23] where an estimator, also based on the de-biased lasso, achieves asymptotic variance strictly smaller than $1/F_\theta$. The construction of [23], Theorem 2.1, goes as follows: Consider a sequence $\lambda_n^\sharp$ and a sequence of sub-regions $\mathcal{B}_n \subset \mathbb{R}^p$ of the parameter space such that the Lasso satisfies uniformly over all $\boldsymbol{\beta} \in \mathcal{B}_n$ both

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta})\|_2 = o_{\mathbb{P}}(1), \qquad \sqrt{n}\lambda_n^\sharp \|\hat{\boldsymbol{\beta}}^{(\text{lasso})} - \boldsymbol{\beta}\|_1 = o_{\mathbb{P}}(1).$$

Then [23] constructs an asymptotically normal estimator of the first component β_1 of $\boldsymbol{\beta}$. However, this estimator depends on a fixed sub-region $\mathcal{B}_n$ that achieves a particular ℓ_1 convergence rate given by $\lambda_n^\sharp$, and the estimator would need to be changed to satisfy asymptotic normality on a superset of $\mathcal{B}_n$. Hence this construction is a super-efficiency phenomenon: it is possible to achieve a strictly smaller variance than the Fisher information lower bound with the F_θ in (1.7) as the estimators are only required to perform well on a specific parameter space $\mathcal{B}_n$. Additionally, the estimator from [23] cannot be regular on perturbations of the form $\boldsymbol{\beta} + t\boldsymbol{u}_0/\sqrt{nF_\theta}$ for non-sparse $\boldsymbol{u}_0$, otherwise that estimator would not be able to achieve an asymptotic variance smaller than $1/F_\theta$ according to Proposition 6.1.

7. Necessity of the degrees-of-freedom adjustment in a more general setting

This section extends Theorem 2.3 to subgaussian designs. It shows that the degrees-of-freedom adjustment is necessary when the Lasso is sign-consistent.

Theorem 7.1. *Let S be a support of size $s_0 = o(n)$ and assume that $\boldsymbol{X}_S \boldsymbol{\Sigma}_{S,S}^{-1/2}$ has iid entries from a mean-zero, variance one and subgaussian distribution. Assume that $(\boldsymbol{\beta}, \boldsymbol{a}_0)$ follows a prior independent of $(\boldsymbol{X}, \boldsymbol{\varepsilon})$ with $\operatorname{supp}(\boldsymbol{\beta}) = S$, $\boldsymbol{\beta}$ has iid random signs on S and fixed amplitudes $\{|\beta_j|,\ j \in S\}$, and set $\boldsymbol{a}_0 = \boldsymbol{\Sigma}\operatorname{sgn}(\boldsymbol{\beta})_S/\sqrt{s_0}$. Then on the selection event $\{\widehat{S} = S, \operatorname{sgn}(\hat{\boldsymbol{\beta}}^{(\text{lasso})}) = \operatorname{sgn}(\boldsymbol{\beta})\}$, the de-biased estimate $\widehat{\theta}_\nu$ in (2.8) with adjustment ν satisfies*

$$\begin{aligned}
&\sqrt{n}(1-\nu/n)(\widehat{\theta}_\nu - \theta) - \sqrt{n}(1-\nu/n)\langle \boldsymbol{a}_0, (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top \boldsymbol{\varepsilon}\rangle \\
&= -(s_0 - \nu)\left(\lambda\sqrt{n}\boldsymbol{a}_0^\top (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\operatorname{sgn}(\boldsymbol{\beta})_S\right) \\
&\quad + O_{\mathbb{P}}\left(\lambda\sqrt{s_0 \log s_0} + \phi_{\text{cond}}(\boldsymbol{\Sigma}_{S,S})^{1/2}\lambda\sqrt{s_0}\right).
\end{aligned}$$

Furthermore, $\lambda\sqrt{n}\boldsymbol{a}_0^\top\left[(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\right]\operatorname{sgn}(\boldsymbol{\beta})_S = \lambda\sqrt{s_0/n}(1 - o_P(1))$ when $\phi_{\text{cond}}(\boldsymbol{\Sigma}_{S,S}) \le C$ for some constant $C > 0$ independent of n, p, s_0. Consequently, if $\nu = 0$ and $s_0^{3/2} \ge n$, the right-hand side above is unbounded.

The proof is given in Appendix I of the supplement [2]. In conclusion, for designs with subgaussian independent entries and under sign-consistency for the Lasso, the unadjusted $\widehat{\theta}_\nu$ with $\nu = 0$ is not asymptotically normal as soon as $s_0 \gg n^{2/3}$, similarly to the Gaussian design case and the conclusion of Theorem 2.3.

8. Outline of the proof

8.1. The interpolation path

Throughout the sequel, let $\boldsymbol{h}^{(\mathrm{lasso})} = \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \boldsymbol{\beta}$. It follows from the definition of $\widehat{\theta}_\nu$ in (2.2) that

$$(1-\nu/n)\big(\widehat{\theta}_\nu - \theta\big) = \frac{\langle z_0, \boldsymbol{\varepsilon}\rangle}{\|z_0\|_2^2} - (\nu/n)\big\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\big\rangle - \frac{\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}\big\rangle}{\|z_0\|_2^2}$$

with $z_0 = \boldsymbol{X}\boldsymbol{u}_0$ and $\boldsymbol{Q}_0 = \boldsymbol{I}_{p\times p} - \boldsymbol{u}_0\boldsymbol{a}_0^\top$, where $\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle$.

In the above expression, z_0 is independent of $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ but not of $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$. If z_0 were independent of $\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}$, we would have

$$\begin{aligned}\mathscr{L}\Big(\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}\big\rangle \Big| \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}\Big) &\sim N\Big(0, C_0^{-2}\|\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}\|_2^2\Big)\\ &= O_{\mathbb{P}}(1/C_0)\|\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}\|_2, \end{aligned} \tag{8.1}$$

where $\mathscr{L}(\xi|\zeta)$ denotes the conditional distribution of ξ given ζ and $C_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2$. Our idea is to decouple z_0 and $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$ by replacing z_0 with an almost independent copy of itself in the definition of $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$.

We proceed as follows. Let $\boldsymbol{g} \sim N(\boldsymbol{0}, \mathbb{E}[z_0 z_0^\top])$ be a random vector independent of $(\boldsymbol{\varepsilon}, z_0, \boldsymbol{X})$ such that $\boldsymbol{g}$ and z_0 have the same distribution. Next, define the random vector

$$\widetilde{z}_0 = \boldsymbol{P}_{\boldsymbol{\varepsilon}} z_0 + \boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp \boldsymbol{g}, \quad \text{where } \boldsymbol{P}_{\boldsymbol{\varepsilon}} = \|\boldsymbol{\varepsilon}\|^{-2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \text{ and } \boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\varepsilon}}.$$

Conditionally on $\boldsymbol{\varepsilon}$, the random vectors z_0 and $\tilde{z}_0$ are identically distributed, so that $\widetilde{z}_0$ is independent of $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$.

Next, let $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{Q}_0 + \widetilde{z}_0\boldsymbol{a}_0^\top$ and let $\tilde{\boldsymbol{\beta}}^{(\mathrm{lasso})}$ be the Lasso solution with $(\boldsymbol{X}, \boldsymbol{y})$ replaced by $(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$. Conditionally on $\boldsymbol{\varepsilon}$, the random vector $\boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp z_0$ is normally distributed and independent of $\boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}$ by construction, so that

$$\begin{aligned}\big|\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\big\rangle\big| &\le \Big|\big\langle \boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp z_0, \boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\big\rangle\Big| + \|\boldsymbol{P}_{\boldsymbol{\varepsilon}} z_0\|\ \|\boldsymbol{P}_{\boldsymbol{\varepsilon}}\boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\|,\\ &\le O_{\mathbb{P}}(1/C_0)\Big(\|\boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp\boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\| + \|\boldsymbol{P}_{\boldsymbol{\varepsilon}}\boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\|\Big),\end{aligned}$$

where the last inequality is a consequence of $\mathbb{E}\|\boldsymbol{P}_{\boldsymbol{\varepsilon}} z_0\|_2^2 = \mathbb{E}\|z_0\|_2^2/n = 1/C_0^2$. The above inequalities are formally proved in Lemma 8.9. Although $\tilde{z}_0$ and z_0 are not independent, conditionally on $\boldsymbol{\varepsilon}$, their $(n-1)$-dimensional projections $\boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp z_0$ and $\boldsymbol{P}_{\boldsymbol{\varepsilon}}^\perp\tilde{z}_0$ are independent and the quantity $\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\big\rangle$ is of the same order as in (8.1) where $\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}^{(\mathrm{lasso})}$ and z_0 were assumed independent.

This motivates the expansion

$$(1-\nu/n)\big(\widehat{\theta}_\nu - \theta\big) = \frac{\langle z_0, \boldsymbol{\varepsilon}\rangle}{\|z_0\|_2^2} - \frac{\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(\mathrm{lasso})}\big\rangle}{\|z_0\|_2^2} + \mathrm{Rem}_\nu, \tag{8.2}$$

with $\mathrm{Rem}_\nu = \|z_0\|_2^{-2}\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0(\tilde{\boldsymbol{\beta}}^{(\mathrm{lasso})} - \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})})\big\rangle - (\nu/n)\big\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\big\rangle$.

The key to our analysis is to bound Rem_ν by differentiating a continuous solution path of the Lasso from $\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}$ to $\tilde{\boldsymbol{\beta}}^{(\mathrm{lasso})}$. To this end, define for any $t \in \mathbb{R}$

$$\begin{aligned} z_0(t) &= \boldsymbol{P}_{\boldsymbol{\varepsilon}} z_0 + \boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\big[(\cos t) z_0 + (\sin t)\boldsymbol{g}\big], \\ \boldsymbol{X}(t) &= \boldsymbol{X}\boldsymbol{Q}_0 + z_0(t)\boldsymbol{a}_0^{\top}, \end{aligned} \tag{8.3}$$

and the Lasso solution corresponding to the design $\boldsymbol{X}(t)$ and noise $\boldsymbol{\varepsilon}$,

$$\hat{\boldsymbol{\beta}}(t) = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \Big\{ \|\boldsymbol{\varepsilon} + \boldsymbol{X}(t)\boldsymbol{\beta} - \boldsymbol{X}(t)\boldsymbol{b}\|_2^2/(2n) + \lambda\|\boldsymbol{b}\|_1 \Big\}. \tag{8.4}$$

For each t, by construction, $(z_0(t), \boldsymbol{X}(t), \hat{\boldsymbol{\beta}}(t))$ has the same distribution as $(z_0, \boldsymbol{X}, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})})$. The above construction defines a continuous path of Lasso solutions along which the distribution of $(z_0(t), \boldsymbol{X}(t), \hat{\boldsymbol{\beta}}(t))$ is invariant. Furthermore,

$$\begin{aligned} \text{at } t = 0, &\qquad z_0(0) = z_0 \text{ and } \hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}, \\ \text{while at } t = \tfrac{\pi}{2}, &\qquad z_0(\tfrac{\pi}{2}) = \tilde{z}_0 \text{ and } \hat{\boldsymbol{\beta}}(\tfrac{\pi}{2}) = \tilde{\boldsymbol{\beta}}^{(\mathrm{lasso})}. \end{aligned}$$

Thus, with $\dot{z}_0(t) = (\partial/\partial t) z_0(t) = \boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}[(-\sin t) z_0 + (\cos t)\boldsymbol{g}]$ and $\boldsymbol{D}(t) = (\partial/\partial z_0(t))\hat{\boldsymbol{\beta}}(t)^{\top} \in \mathbb{R}^{n\times p}$, an application of the chain rule yields

$$\mathrm{Rem}_\nu = \int_0^{\pi/2} \frac{\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{D}^{\top}(t)\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\dot{z}_0(t)\big\rangle}{\|z_0\|_2^2} dt - (\nu/n)\big\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\big\rangle. \tag{8.5}$$

We will prove in Lemma 8.5 below that the above calculus is legitimate with

$$\begin{aligned} &\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{D}^{\top}(t)\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp} \\ &= -\Big\{ \boldsymbol{w}_0(t) - z_0(t)\|\boldsymbol{w}_0(t)\|_2^2 \Big\}\big(\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\boldsymbol{X}(t)\boldsymbol{h}(t)\big)^{\top} \\ &\quad -\Big\{ \widehat{\boldsymbol{P}}(t) - z_0(t)\big(\boldsymbol{w}_0(t)\big)^{\top} \Big\}\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle, \end{aligned} \tag{8.6}$$

where $\widehat{S}(t) = \mathrm{supp}(\hat{\boldsymbol{\beta}}(t))$, $\widehat{\boldsymbol{P}}(t)$ is the orthogonal projection onto the linear span of $\{\boldsymbol{X}_j(t),\ j \in \widehat{S}(t)\}$, $\boldsymbol{w}_0(t) = \boldsymbol{X}_{\widehat{S}(t)}(t)\big(\boldsymbol{X}_{\widehat{S}(t)}^{\top}(t)\boldsymbol{X}_{\widehat{S}(t)}(t)\big)^{-1}(\boldsymbol{a}_0)_{\widehat{S}(t)}$, and $\boldsymbol{h}(t) = \hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}$. We note that the $n \times n$ matrix in (8.6) is a function of $(\boldsymbol{X}(t), \boldsymbol{\varepsilon})$ and

$$z_0 = \boldsymbol{P}_{\boldsymbol{\varepsilon}} z_0 + \boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\big[(\cos t) z_0(t) - (\sin t)\dot{z}(t)\big]$$

with $z_0(t) = \boldsymbol{X}(t)\boldsymbol{u}_0$. Thus, as $\dot{z}_0(t)$ is a $N(\boldsymbol{0}, \boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}/C_0^2)$ vector given $(\boldsymbol{X}(t), \boldsymbol{\varepsilon})$, the mean and variance of the integrand $\big\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{D}^{\top}(t)\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\dot{z}_0(t)\big\rangle$ in (8.5) can be readily computed conditionally on $(\boldsymbol{X}(t), \boldsymbol{\varepsilon})$ as a quadratic form in $\dot{z}_0(t)$. This would provide an upper bound for the remainder in (8.5) based on the size of $\widehat{S}(t)$ and the prediction error $\boldsymbol{X}(t)\boldsymbol{h}(t)$. For example, the main term in this calculation is

$$\begin{aligned} &\Big(\mathbb{E}\|z_0\|_2^2\Big)^{-1}\int_0^{\pi/2} \mathbb{E}\Big[\big\langle z_0, -\widehat{\boldsymbol{P}}(t)\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\dot{z}_0(t)\big\rangle\big\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\big\rangle\Big|\boldsymbol{X}(t), \boldsymbol{\varepsilon}\Big] dt \\ &= \frac{1}{n}\int_0^{\pi/2} (\sin t)\Big\{ \big|\widehat{S}(t)\big| - \mathrm{trace}\Big(\boldsymbol{P}_{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{P}}(t)\boldsymbol{P}_{\boldsymbol{\varepsilon}}\Big)\Big\}\big\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\big\rangle dt, \end{aligned}$$

which has approximately the same mean as $(\nu/n)\big\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\big\rangle$ when $\nu = \big|\widehat{S}(0)\big| = \big|\widehat{S}\big|$.

Remark 8.1. For a fixed j-th column the leave-one-out technique explained in [15], Section 6.1, studies the modified estimate

$$\widehat{\boldsymbol{\theta}}^{(j)} = \mathop{\arg\min}_{\boldsymbol{b}\in\mathbb{R}^p : b_j=\beta_j} \|\boldsymbol{X}\boldsymbol{b}-\boldsymbol{y}\|_2^2/(2n) + g(\boldsymbol{b}) \tag{8.7}$$

with the constraint $b_j = \beta_j$, so that the design matrix in the quadratic term is replaced by $\boldsymbol{X}_{-j}$. The study of this perturbed $\widehat{\boldsymbol{\theta}}^{(j)}$ allows [15] to prove efficiency under the condition $\max_{j=1,\dots,p} \|\Sigma^{-1}e_j\|_1 \le \rho$. This differs from our construction in at least three major ways:

(i) The $\widehat{\boldsymbol{\theta}}^{(j)}$ of [15] does not have the same distribution as $\hat{\boldsymbol{\beta}}^{(\text{lasso})}$, while with our construction $\tilde{\boldsymbol{\beta}}^{(\text{lasso})}$ as well as $\hat{\boldsymbol{\beta}}(t)$ for each $t \in [0,\pi/2]$ all have the same distribution as the Lasso $\hat{\boldsymbol{\beta}}$ itself;
(ii) In our construction the decomposition $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + z_0\boldsymbol{a}_0^\top$ has two independent terms $\boldsymbol{X}\boldsymbol{Q}_0$ and $z_0\boldsymbol{a}_0^\top$, while in the construction (8.7) above, $\boldsymbol{X} = \boldsymbol{X}_{-j} + \boldsymbol{X}\boldsymbol{e}_j$ but $\boldsymbol{X}_{-j}$ is not independent of the j-th column $\boldsymbol{X}\boldsymbol{e}_j$;
(iii) Our construction allows for general direction $\boldsymbol{a}_0$, while the analogue of (8.7) with constraint $\boldsymbol{a}_0^\top\boldsymbol{b}$ for dense $\boldsymbol{a}_0$, namely $\widehat{\boldsymbol{\theta}}^{(0)} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p : \boldsymbol{a}_0^\top(\boldsymbol{\beta}-\boldsymbol{b})=0} \|\boldsymbol{X}\boldsymbol{b}-\boldsymbol{y}\|_2^2/(2n) + g(\boldsymbol{b})$, leads to an estimator that is *not* a Lasso estimator, and its analysis would not be straightforward.

8.2. The lasso prediction error and model size

Our next task is to show that with high probability, simultaneously for all t along the path, the Lasso solutions $\hat{\boldsymbol{\beta}}(t)$ enjoy guarantees in terms of prediction error and model size similar to the bounds available for a single Lasso problem. Define the event Ω_1 by

$$\Omega_1 = \left\{ 0 < \inf_{t,t'\ge 0} \phi_{\min}\left(\frac{1}{n}\left(\boldsymbol{X}(t)^\top \boldsymbol{X}(t) \right)_{\widehat{S}(t')\cup\widehat{S}(t),\widehat{S}(t')\cup\widehat{S}(t)} \right). \right\} \tag{8.8}$$

Define also $\boldsymbol{h}^{(noiseless)}(t) = \boldsymbol{\beta}^{(noiseless)}(t) - \boldsymbol{\beta}$ where $\boldsymbol{\beta}^{(noiseless)}(t)$ is the Lasso solution for design matrix $\boldsymbol{X}(t)$ in the absence of noise, that is,

$$\boldsymbol{\beta}^{(noiseless)}(t) = \mathop{\arg\min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \|\boldsymbol{X}(t)(\boldsymbol{\beta}-\boldsymbol{b})\|_2^2/(2n) + \lambda\|\boldsymbol{b}\|_1 \right\}. \tag{8.9}$$

Consider the following conditions: For a certain $s_* \in [s_0 \vee 1, n]$ and positive λ_0,

$$\begin{aligned}
&\|\boldsymbol{X}(t)\boldsymbol{h}(t)\|_2 \le M_1\sqrt{ns_*}\sigma\lambda_0,\\
&\|\boldsymbol{X}(t)\boldsymbol{h}^{(noiseless)}(t)\|_2 \le M_1\sqrt{ns_*}\sigma\lambda_0,\\
&\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}(t)\|_2 \le M_2\sqrt{s_*}\sigma\lambda_0,\\
&|\widehat{S}(t)| \le s_* \le M_3(s_0+k),\\
&\left\| \left(\boldsymbol{\Sigma}^{-1/2}_{\widehat{S}(t),\widehat{S}(t)} \boldsymbol{X}^\top_{\widehat{S}(t)}(t) \boldsymbol{X}_{\widehat{S}(t)}(t) \boldsymbol{\Sigma}^{-1/2}_{\widehat{S}(t),\widehat{S}(t)}/n \right)^{-1} \right\|_{op} \le M_4,\\
&(\|\boldsymbol{\varepsilon}\|_2/\sigma) \vee (C_0\|z_0(t)\|_2) \vee (n/(C_0\|z_0(t)\|_2)) \le M_5\sqrt{n},
\end{aligned} \tag{8.10}$$

where $M_1, M_2, M_3, M_4, M_5 > 0$ are constants to be specified. Define the event Ω_2 by

$$\Omega_2(t) = \left\{ \text{ (8.10) holds for } t \right\} \text{ and } \Omega_2 = \cap_{t\ge 0}\Omega_2(t). \tag{8.11}$$

For a single and fixed value of t, the fact that the Lasso enjoys the inequalities (8.10) under conditions on the design $\mathbf{\Sigma}$ can be obtained using known techniques. For instance, the first and third inequalities in (8.10) describe the prediction rate of the Lasso with respect to the empirical covariance matrix and the population covariance matrix when the tuning parameter of the Lasso is proportional to $\sigma\lambda_0$. For the purpose of the present paper, however, we require the above inequalities to hold with high probability simultaneously for all t. The following lemma shows that this is the case: $\Omega_1 \cap \Omega_2$ has overwhelming probability under Assumption 3.1.

Lemma 8.1. *Let the setting and conditions of Assumption 3.1 be fulfilled. Set* $M_1 = (1+\eta_2)\eta_2^{-1}(1+\eta_3)/\sqrt{\rho_*\tau_*}$, $M_2 = M_1/\sqrt{\tau_*}$,

$$M_3 = 1 + \frac{(\tau^*/\tau_*)\phi_{\mathrm{cond}}(p;\varnothing,\mathbf{\Sigma}) - 1}{2(1-\eta_2)^2/(1+\eta_2)^2},$$

$M_4 = 1/\tau_*$, $M_5 = 1/(1-\eta_3)$. *Then the events* Ω_1, Ω_2 *defined in* (8.8) *and* (8.11) *satisfy*

$$\begin{aligned} 1 - \mathbb{P}(\Omega_1 \cap \Omega_2) \leq \quad & 2e^{-n\epsilon_4} + 2e^{-(\eta_3 - \sqrt{2/n})_+^2 n/2} \\ & + e^{-n\eta_3^2/2} + 4(2\pi L_k^2 + 4)^{-1/2} + (L_k + (L_k^2+2)^{-1/2})^{-2}. \end{aligned} \tag{8.12}$$

where $L_k = \sqrt{2\log(p/k)}$.

Lemma 8.1 is proved in Appendix A of the supplement [2]. Equipped with the result that the events Ω_1 and Ω_2 have overwhelming probability, we are now ready to bound Rem_ν in (8.2).

8.3. An intermediate result

Before proving the main result (Theorem 3.1) in the next subsections, we now prove the following intermediate result.

Theorem 8.2. *There exists a constant* $\bar{M} > 0$ *that depends on* M_1, M_2, M_4, M_5 *only such that the following holds. Let* $F_\theta = 1/(\sigma C_0)^2$ *be the Fisher information as in* (1.7), *and* $T_n = \sqrt{nF_\theta}\langle z_0, \boldsymbol{\varepsilon}\rangle/\|z_0\|_2^2$ *so that* T_n *has the* t*-distribution with* n *degrees of freedom. Let* Ω_1 *and* Ω_2 *be the events defined in* (8.8) *and* (8.11). *Define random variables* Rem_I *and* Rem_{II} *by*

$$\mathrm{Rem}_I = \sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) - T_n - \sqrt{F_\theta/n}\int_0^{\pi/2} (\sin t)\left(|\widehat{S}(t)|\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle\right) dt,$$

$$\mathrm{Rem}_{II} = \sqrt{nF_\theta}(\widehat{\theta}_{\nu=0} - \theta) - T_n - \sqrt{F_\theta/n}\left\langle \boldsymbol{a}_0, \boldsymbol{h}^{(lasso)}\right\rangle \int_0^{\pi/2} (\sin t)\left(|\widehat{S}(t)|\right) dt.$$

Then for any $u \in \mathbb{R}$ *such that* $|u| \leq \sqrt{n}/\bar{M}$,

$$\max\left\{\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{u\mathrm{Rem}_I}{\lambda_0\sqrt{s_*}}\right)\right], \mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{u\mathrm{Rem}_{II}}{\lambda_0\sqrt{s_*}}\right)\right]\right\} \leq 2\exp\left(\bar{M}^2u^2\right).$$

We now gather some notation and lemmas to prove Theorem 8.2. Recall that the degrees-of-freedom adjusted LDPE is

$$\widehat{\theta}_\nu = \left\langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}\right\rangle + \frac{\left\langle z_0, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\mathrm{lasso})}\right\rangle}{(1-\nu/n)\|z_0\|_2^2},$$

with $z_0 = \boldsymbol{X}\boldsymbol{u}_0$, where $\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle$ is the direction of the least favorable one-dimensional sub-model for the estimation of $\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$. Recall that the Fisher information for the estimation of $\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ is $F_\theta = \sigma^{-2}/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle$, and that $\mathbb{E}\|z_0\|_2^2/n = \sigma^2 F_\theta = 1/C_0^2$. We note that the estimation of $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ is scale equi-variant under the transformation

$$\left\{\boldsymbol{a}_0, \theta, \widehat{\theta}_v, \boldsymbol{u}_0, z_0, F_\theta\right\} \to \left\{c\boldsymbol{a}_0, c\theta, c\widehat{\theta}_v, \boldsymbol{u}_0/c, z_0/c, F_\theta/c^2\right\}. \tag{8.13}$$

Thus, without loss of generality, we may take the scale $\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\rangle = 1$ in which

$$\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0, \qquad z_0 = \boldsymbol{X}\boldsymbol{u}_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_n), \qquad F_\theta = \sigma^{-2}, \qquad C_0 = 1. \tag{8.14}$$

Furthermore, for any subset $A \subset \{1, \ldots, p\}$ we have

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}_{A,A}^{-1/2}(\boldsymbol{a}_0)_A\right\|_2^2 &= \left\|\boldsymbol{\Sigma}_{A,A}^{-1/2}(\boldsymbol{\Sigma}^{1/2})_{A,*}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\right\|_2^2 \\
&\le C_0^2\phi_{\max}\left(\boldsymbol{\Sigma}_{A,A}^{-1/2}(\boldsymbol{\Sigma}^{1/2})_{A,*}(\boldsymbol{\Sigma}^{1/2})_{*,A}\boldsymbol{\Sigma}_{A,A}^{-1/2}\right) \\
&\le C_0^2\phi_{\max}\left(\boldsymbol{\Sigma}_{A,A}^{-1/2}\boldsymbol{\Sigma}_{A,A}\boldsymbol{\Sigma}_{A,A}^{-1/2}\right) \\
&= C_0^2.
\end{aligned} \tag{8.15}$$

Let $\dot{f}(t) = (\partial/\partial t) f(t)$ for all functions of t. By construction of the interpolation path (8.3), we have

$$\dot{z}_0(t) = \boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\left[(-\sin t)z_0 + (\cos t)\boldsymbol{g}\right], \tag{8.16}$$

so that $\langle \boldsymbol{\varepsilon}, \dot{z}_0(t)\rangle = 0$ holds for every t. Conditionally on $\boldsymbol{\varepsilon}$, the random vector $(\boldsymbol{X}(t), \dot{z}_0(t))$ is jointly normal and $\dot{z}_0(t)$ is independent of $\boldsymbol{X}(t)$, so that the conditional distribution of $\dot{z}_0(t)$ given $(\boldsymbol{X}(t), \boldsymbol{\varepsilon})$ is

$$\mathscr{L}\left(\dot{z}_0(t)\middle|\boldsymbol{X}(t), \boldsymbol{\varepsilon}\right) = N\left(\boldsymbol{0}, (1/C_0)^2\boldsymbol{P}_{\boldsymbol{\varepsilon}}^{\perp}\right). \tag{8.17}$$

Here is an outline of the proof of Theorem 8.2.

(i) Starting from the expansion (8.2), the key to our analysis is to bound the remainder in (8.2) by differentiating the continuous solution path (8.3)-(8.4) from $\hat{\boldsymbol{\beta}}^{(\text{lasso})}$ to $\tilde{\boldsymbol{\beta}}^{(\text{lasso})}$.

(ii) Lemma 8.3 shows that the function $t \to \hat{\boldsymbol{\beta}}(t)$ is Lipschitz in t, hence differentiable almost everywhere along the path.

(iii) Next, Lemma 8.5 computes the gradient of $t \to \hat{\boldsymbol{\beta}}(t)$ along the path. To compute the gradient, we make use of Lemma 8.4 which shows that the KKT conditions of the Lasso hold strictly almost everywhere.

(iv) Finally, we write $\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0(\tilde{\boldsymbol{\beta}}^{(\text{lasso})} - \hat{\boldsymbol{\beta}}^{(\text{lasso})})\rangle$ as an integral from 0 to $\pi/2$ of the derivative of the function $t \to \langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\hat{\boldsymbol{\beta}}(t)\rangle$ and the Lemmas 8.6, 8.7 and 8.8 bound from above this derivative on the event $\Omega_1 \cap \Omega_2$, thanks to the conditional distribution (8.17) of $\dot{z}_0(t)$ given $(\boldsymbol{X}(t), \boldsymbol{\varepsilon})$.

Lemma 8.3 (Lipschitzness of regularized least-squares with respect to the design). *Let $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Let $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ be two design matrices of size $n \times p$ in a compact convex set $\tilde{K}$. Let h be a norm in $\mathbb{R}^p$. Let $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ be the minimizers*

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p}\{L(\boldsymbol{X}, \boldsymbol{b}) + h(\boldsymbol{b})\}, \qquad \tilde{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p}\left\{L(\tilde{\boldsymbol{X}}, \boldsymbol{b}) + h(\boldsymbol{b})\right\}$$

where $L(\boldsymbol{M},\boldsymbol{b}) = \|\boldsymbol{\varepsilon} + \boldsymbol{M}\boldsymbol{\beta} - \boldsymbol{M}\boldsymbol{b}\|_2^2/(2n)$ *for all* $\boldsymbol{M} \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{b} \in \mathbb{R}^p$. *Then*

$$\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\|^2 + \|\tilde{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\|^2 \le C(\tilde{K}, h, \boldsymbol{\varepsilon}, \boldsymbol{\beta})\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_{op}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2,$$

where $C(\tilde{K}, h, \boldsymbol{\varepsilon}, \boldsymbol{\beta})$ *is a quantity that depends on* $\tilde{K}, h, \boldsymbol{\varepsilon}, \boldsymbol{\beta}$ *only.*

Lemma 8.4. *Consider a random design matrix* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *and independent random noise* $\boldsymbol{\varepsilon}$ *such that both* $\boldsymbol{X}$ *and* $\boldsymbol{\varepsilon}$ *admit a density with respect to the Lebesgue measure. Then with probability one, the KKT conditions of the Lasso hold strictly, that is,* $\mathbb{P}(\forall j \in \widehat{S}, \quad |\boldsymbol{x}_j^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(lasso)})| < 1) = 1.$

Proof. Since the distribution of $\boldsymbol{X}$ is continuous, the assumption of [1], Proposition 4.1, is satisfied almost surely with respect to $\boldsymbol{X}$ and the result follows by conditionaning on $\boldsymbol{X}$. □

Lemma 8.5. *Let* $\boldsymbol{h}(t) = \hat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}$. *In the event* Ω_1 *defined by* (8.8),

$$\tilde{\boldsymbol{\beta}}^{(lasso)} - \hat{\boldsymbol{\beta}}^{(lasso)} = \int_0^{\pi/2} \boldsymbol{D}^\top(t)\dot{\boldsymbol{z}}_0(t)dt \tag{8.18}$$

almost surely, where $\boldsymbol{D}(t)$ *is an* $n \times p$ *matrix given by* $\boldsymbol{D}_{\widehat{S}^c(t)}(t) = 0$ *and*

$$\boldsymbol{D}^\top_{\widehat{S}(t)}(t) = \Big(\boldsymbol{X}^\top(t)\boldsymbol{X}(t)\Big)^{-1}_{\widehat{S}(t),\widehat{S}(t)}\Big((\boldsymbol{a}_0)_{\widehat{S}(t)}\big(\boldsymbol{\varepsilon} - \boldsymbol{X}(t)\boldsymbol{h}(t)\big)^\top - \boldsymbol{X}^\top_{\widehat{S}(t)}(t)\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle\Big).$$

It follows from (8.3) and (8.16) that conditionally on $\boldsymbol{\varepsilon}$, the random vector $\dot{\boldsymbol{z}}_0(t)$ is independent of $(\boldsymbol{X}(t), \boldsymbol{h}(t), \boldsymbol{D}(t), I_{\Omega_2(t)})$ and the conditional distribution of $\dot{\boldsymbol{z}}_0(t)$ given $(\boldsymbol{\varepsilon}, \boldsymbol{X}(t))$ is given by (8.17). Furthermore, by (8.16) we always have $\langle \dot{\boldsymbol{z}}_0(t), \boldsymbol{\varepsilon}\rangle = 0$ so that $(\boldsymbol{\varepsilon} - \boldsymbol{X}(t)\boldsymbol{h}(t))^\top \dot{\boldsymbol{z}}_0(t) = -(\boldsymbol{X}(t)\boldsymbol{h}(t))^\top \dot{\boldsymbol{z}}_0(t)$ which simplifies the expression $\boldsymbol{D}^\top_{\widehat{S}(t)}(t)\dot{\boldsymbol{z}}_0(t)$. Furthermore on $\Omega_2(t)$ defined in (8.11), by the Cauchy-Schwarz inequality,

$$\begin{aligned} |\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle| &\le C_0\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}(t)\|_2 \le C_0 M_1 \sigma\lambda_0\sqrt{s_*}, \\ \|\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{h}(t)\|_2/\sqrt{n} &\le (M_1 + M_5M_2)\sigma\lambda_0\sqrt{s_*}, \\ \|\boldsymbol{w}_0(t)\|_2^2 &\le (M_4/n)\|\boldsymbol{\Sigma}^{-1/2}_{\widehat{S}(t),\widehat{S}(t)}(\boldsymbol{a}_0)_{\widehat{S}(t)}\|_2^2 \le (M_4/n)C_0^2 \end{aligned} \tag{8.19}$$

with $\boldsymbol{w}_0(t) = \boldsymbol{X}_{\widehat{S}(t)}(t)\big(\boldsymbol{X}^\top_{\widehat{S}(t)}(t)\boldsymbol{X}_{\widehat{S}(t)}(t)\big)^{-1}(\boldsymbol{a}_0)_{\widehat{S}(t)}$, thanks to (8.10) and (8.15). We will use these properties several times in the following lemmas in order to bound Rem_v in (8.2).

Lemma 8.6. *The quantity*

$$W = C_0\sqrt{n}\left(\frac{\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0(\tilde{\boldsymbol{\beta}}^{(lasso)} - \hat{\boldsymbol{\beta}}^{(lasso)})\rangle}{C_0^2\|z_0\|_2^2} - \frac{\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0(\tilde{\boldsymbol{\beta}}^{(lasso)} - \hat{\boldsymbol{\beta}}^{(lasso)})\rangle}{n}\right) \tag{8.20}$$

satisfies for any $u \in \mathbb{R}$

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{uW}{\sigma\lambda_0\sqrt{s_*}}\right)\right] \le \exp(C|u| + Cu^2) \tag{8.21}$$

for some constant $C = C(M_1, M_2, M_5) > 0$ *that depends on* M_1, M_2, M_5 *only.*

Lemma 8.7. *The quantity*

$$W' = \frac{C_0\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0(\tilde{\boldsymbol{\beta}}^{(lasso)} - \hat{\boldsymbol{\beta}}^{(lasso)})\rangle}{\sqrt{n}} - \int_0^{\pi/2} (\sin t)\frac{|\widehat{S}(t)|\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle}{C_0\sqrt{n}}dt \tag{8.22}$$

satisfies

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{uW'}{\sigma\lambda_0\sqrt{s_*}}\right)\right] \le \exp\left(|u|C'/\sqrt{n} + \frac{u^2C'}{1-|u|C'/\sqrt{n}}\right) \tag{8.23}$$

for any $u\in\mathbb{R}$ such that $|u| < \sqrt{n}/C'$, for some constant $C' = C'(M_1, M_2, M_4, M_5) > 0$ that depends on M_1, M_2, M_4, M_5 only.

Lemma 8.8. *The quantity*

$$W'' = \frac{1}{C_0\sqrt{n}}\int_0^{\pi/2}(\sin t)|\widehat{S}(t)|\langle \boldsymbol{a}_0, \boldsymbol{h}(t)\rangle dt - \frac{\langle \boldsymbol{a}_0, \boldsymbol{h}^{(lasso)}\rangle}{C_0\sqrt{n}}\int_0^{\pi/2}(\sin t)|\widehat{S}(t)|dt \tag{8.24}$$

satisfies for all $u\in R$

$$\mathbb{E}\left[\exp\left(\frac{uW''}{\sigma\lambda_0\sqrt{s_*}}\right)\right] \le 2\exp(C''u^2) \tag{8.25}$$

for some constant $C'' = C''(M_1, M_2, M_4, M_5) > 0$ that depends on M_1, M_2, M_4, M_5 only.

Lemma 8.9. *The quantity*

$$W''' = -\frac{\sqrt{n}C_0\langle z_0, \boldsymbol{X}\boldsymbol{Q}_0\widetilde{\boldsymbol{h}}^{(lasso)}\rangle}{C_0^2\|z_0\|_2^2} \tag{8.26}$$

satisfies for all $u\in\mathbb{R}$

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{uW'''}{\sigma\lambda_0\sqrt{s_*}}\right)\right] \le 2\exp(C'''u^2) \tag{8.27}$$

for some constant $C''' = C'''(M_1, M_2, M_5)$ that depends on M_1, M_2, M_5 only.

We are now ready to combine the above lemmas to prove Theorem 8.2.

Proof of Theorem 8.2. The random variables Rem_I and Rem_{II} in Theorem 8.2 satisfy

$$\sigma\mathrm{Rem}_I = W''' + W + W', \qquad \sigma\mathrm{Rem}_{II} = \sigma\mathrm{Rem}_I + W'' = W''' + W + W' + W''.$$

where W, W', W'' and W''' are defined in (8.20), (8.22), (8.24) and (8.26). By Lemmas 8.6 to 8.9 , there exists a constant $\bar{M} > 0$ that depends only on M_1, M_2, M_4, M_5 such that for all $u\in\mathbb{R}$ with $|u| < \sqrt{n}/\bar{M}$,

$$\max_{V\in\{W, W', W'', W'''\}}\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp\left(\frac{uV}{\sigma\lambda_0\sqrt{s_*}}\right)\right] \le 2\exp\left(\bar{M}^2u^2\right) \tag{8.28}$$

because one can always increase $\bar{M}$ so that the right hand side of the previous display is larger than the right hand side of (8.21), (8.23) (8.25) and (8.27). By Jensen's inequality,

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\exp(\frac{u\mathrm{Rem}_I}{\lambda_0\sqrt{s_*}})\right] \le \frac{1}{3}\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\left(e^{\frac{3uW'''}{\sigma\lambda_0\sqrt{s_*}}}+e^{\frac{3uW}{\sigma\lambda_0\sqrt{s_*}}}+e^{\frac{3uW'}{\sigma\lambda_0\sqrt{s_*}}}\right)\right].$$

The right hand side is bounded from above thanks to (8.28). We apply the same technique to obtain the desired bound on Rem_{II}, using Lemma 8.8 for W''. □

8.4. Proof of Theorem 3.1

From Theorem 8.2, in order to complete prove Theorem 3.1 we will need the following additional lemma.

Lemma 8.10. *The upper bound*

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\left(\int_0^{\pi/2}(\sin t)(|\widehat{S}(t)|-|\widehat{S}(0)|)dt\right)^2\right] \qquad \le n\left(\lambda_0^2 s_* C'''' + 6(3+2M_1^2\lambda_0^2 s_*)\right)$$

holds, where $C'''' = 3(M_5M_1 + M_2M_5M_4)^2$.

Proof of Theorem 3.1. Thanks to the scale equivariance (8.13), we take the scale $C_0 = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = 1$ without loss of generality, so that (8.14) holds. Let Rem_{II} be defined in Theorem 8.2. Then for any degrees-of-freedom adjustment ν we have

$$\begin{aligned}&\sqrt{F_\theta n}(1-\nu/n)(\widehat{\theta}_\nu-\theta) - T_n + \sqrt{F_\theta/n}\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\rangle(\nu-|\widehat{S}|)\\ &= \sqrt{F_\theta/n}\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\rangle\int_0^{\pi/2}(\sin t)(|\widehat{S}(t)|-|\widehat{S}|)dt + \mathrm{Rem}_{II}.\end{aligned}$$

Denote by Rem_{final} the above quantity. Then

$$\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\left|\frac{\mathrm{Rem}_{final}}{\lambda_0\sqrt{s_*}}\right|^2\right] \le \begin{cases}2M_2^2\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\left(\int_0^{\pi/2}(\sin t)(|\widehat{S}(t)|-|\widehat{S}|)n^{-1/2}dt\right)^2\right]\\ +2\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\left\{\mathrm{Rem}_{II}/(\lambda\sqrt{s_*})\right\}^2\right].\end{cases}$$

By Theorem 8.2, $\mathbb{E}\left[I_{\Omega_1\cap\Omega_2}\mathrm{Rem}_{II}{}^2\right]$ is bounded by a constant that depends on M_1, M_2, M_4, M_5 only. By Lemma E.1 of the supplement [2] and the assumption $\lambda_0\sqrt{s_*}\le 1$ in Assumption 3.1, the same holds for the first term. Observe that since $\mathbb{P}(\Omega_1\cap\Omega_2)\to 1$, any random variable Y such that $\mathbb{E}[I_{\Omega_1\cap\Omega_2}Y^2]\le C\lambda_0^2 s_*$ for some constant C satisfies $Y = O_{\mathbb{P}}(\sqrt{s_*}\lambda_0)$ by Markov's inequality. This shows that $\mathrm{Rem}_{final} = O_{\mathbb{P}}(\lambda_0\sqrt{s_*})$ and the proof is complete. □

8.5. Proof of Corollary 3.3

On Ω_2 we have $|\widehat{S}|\le s_*$ and $|\langle \boldsymbol{a}_0, \boldsymbol{h}^{(\mathrm{lasso})}\rangle| \le M_2\sigma\lambda_0\sqrt{s_*}$ so the claim of Corollary 3.3 follows from the same argument as the previous subsection.

Funding

P.C.B. was partially supported supported by the NSF Grants DMS-1811976 and DMS-1945428.

C-H.Z. was partially supported by the NSF Grants DMS-1513378, IIS-1407939, DMS-1721495, IIS-1741390 and CCF-1934924.

Supplementary Material

Proofs of the results (DOI: 10.3150/21-BEJ1348SUPP; .pdf). The supplement [2] contains all proofs of the results stated in the paper. DOI to be added by the typesetter.

References

[1] Bellec, P.C. and Zhang, C.-H. (2021). Second order Stein: Sure for sure and other applications in high-dimensional inference. *Ann. Statist.* To appear.

[2] Bellec, P.C. and Zhang, C.-H. (2022). Supplement to "De-biasing the lasso with degrees-of-freedom adjustment." https://doi.org/10.3150/21-BEJ1348SUPP

[3] Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163 https://doi.org/10.3150/11-BEJ410

[4] Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

[5] Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.* **42** 757–788. MR3210986 https://doi.org/10.1214/14-AOS1204

[6] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[7] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[8] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

[9] Cai, T., Cai, T. and Guo, Z. (2019). Individualized treatment selection: An optimal hypothesis testing approach in high-dimensional models. ArXiv preprint. Available at arXiv:1904.12891.

[10] Cai, T.T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

[11] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[12] Candes, E.J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[13] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[14] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629

[15] Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630

[16] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087 https://doi.org/10.1214/08-EJS177

[17] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[18] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055
[19] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166 https://doi.org/10.1093/biomet/ass043
[20] Tibshirani, R.J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948 https://doi.org/10.1214/12-AOS1003
[21] Tropp, J.A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory* **52** 1030–1051. MR2238069 https://doi.org/10.1109/TIT.2005.864420
[22] van de Geer, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math.* **2159**. Cham: Springer. MR3526202 https://doi.org/10.1007/978-3-319-32774-7
[23] van de Geer, S. (2017). On the efficiency of the de-biased lasso. arXiv preprint. Available at arXiv:1708.07986.
[24] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
[25] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256
[26] Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018
[27] Zhang, C.-H. (2005). Estimation of sums of random variables: Examples and information bounds. *Ann. Statist.* **33** 2022–2041. MR2211078 https://doi.org/10.1214/009053605000000390
[28] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729
[29] Zhang, C.-H. (2011). *Statistical inference for high-dimensional data*, Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report 48, 28–31.
[30] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520
[31] Zhang, C.-H. and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026
[32] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135 https://doi.org/10.1214/12-STS399
[33] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449
[34] Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electron. J. Stat.* **12** 3312–3364. MR3861831 https://doi.org/10.1214/18-EJS1443
[35] Zhu, Y. and Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319
[36] Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electron. J. Stat.* **12** 3312–3364. MR3861831 https://doi.org/10.1214/18-EJS1443
[37] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127

Received May 2020 and revised March 2021