*Submitted to the Annals of Statistics*

# DEBIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS*

By Adel Javanmard[§] and Andrea Montanari[¶]

*University of Southern California*[§] *and Stanford University*[¶]

Performing statistical inference in high-dimensional models is challenging because of the lack of precise information on the distribution of high-dimensional regularized estimators.

Here, we consider linear regression in the high-dimensional regime $p \gg n$ and the Lasso estimator: we would like to perform inference on the parameter vector $\theta^* \in \mathbb{R}^p$. Important progress has been achieved in computing confidence intervals and p-values for single coordinates $\theta^*_i$, $i \in \{1, \dots, p\}$. A key role in these new inferential methods is played by a certain debiased estimator $\widehat{\theta}^{\rm d}$. Earlier work establishes that, under suitable assumptions on the design matrix, the coordinates of $\widehat{\theta}^{\rm d}$ are asymptotically Gaussian provided the true parameters vector $\theta^*$ is $s_0$-sparse with $s_0 = o(\sqrt{n}/\log p)$.

The condition $s_0 = o(\sqrt{n}/\log p)$ is considerably stronger than the one for consistent estimation, namely $s_0 = o(n/\log p)$. In this paper, we consider Gaussian designs with known or unknown population covariance. When the covariance is known, we prove that the debiased estimator is asymptotically Gaussian under the nearly optimal condition $s_0 = o(n/(\log p)^2)$.

The same conclusion holds if the population covariance is unknown but can be estimated sufficiently well. For intermediate regimes, we describe the trade-off between sparsity in the coefficients $\theta^*$, and sparsity in the inverse covariance of the design. We further discuss several applications of our results beyond high-dimensional inference. In particular, we propose a thresholded Lasso estimator that is minimax optimal up to a factor $1 + o_n(1)$ for i.i.d. Gaussian designs.

## 1. Introduction.

1.1. *Background.* Consider a random design model where we are given $n$ i.i.d. pairs $(y_1, x_1)$, $(y_2, x_2)$, $\cdots$, $(y_n, x_n)$ with $y_i \in \mathbb{R}$, and $x_i \in \mathbb{R}^p$. The

* A. Javanmard was partially supported by a Google Faculty Research Award. A. Javanmard would also like to acknowledge the financial support of the Office of the Provost at the University of Southern California through the Zumberge Fund Individual Grant Program. A. Montanari was supported in part by NSF grants CCF-1319979 and DMS-1106627, and the AFOSR grant FA9550-13-1-0036.

*MSC 2010 subject classifications:* Primary 62J05, 62J07; secondary 62F12

*Keywords and phrases:* Lasso, high-dimensional regression, confidence intervals, hypothesis testing, bias and variance, sample size.

response variable $y_i$ is a linear function of $x_i$, contaminated by noise $w_i$ independent of $x_i$

$$y_i = \langle \theta^*, x_i \rangle + w_i \,, \qquad w_i \sim \mathsf{N}(0, \sigma^2)\,. \tag{1}$$

Here $\theta^* \in \mathbb{R}^p$ is a vector of parameters to be estimated and $\langle \cdot , \cdot \rangle$ is the standard scalar product.

In matrix form, letting $y = (y_1, \dots, y_n)^{\mathsf{T}}$ and denoting by $X$ the matrix with rows $x_1^{\mathsf{T}}, \cdots, x_n^{\mathsf{T}}$ we have

$$y = X\,\theta^* + w \,, \qquad w \sim \mathsf{N}(0, \sigma^2 \mathrm{I}_{n\times n})\,. \tag{2}$$

We are interested in the high-dimensional regime wherein the number of parameters $p$ exceeds the sample size $n$. Over the last 20 years, impressive progress has been made in developing and understanding highly effective estimators in this regime [16, 9, 11]. A prominent approach is the Lasso [57, 19] defined through the following convex optimization problem

$$\widehat{\theta}^{\text{Lasso}}(y, X; \lambda) \equiv \arg\max_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\} . \tag{3}$$

(We will omit the arguments of $\widehat{\theta}^{\text{Lasso}}(y, X; \lambda)$ whenever clear from the context.)

A far less understood question is how to perform statistical inference in the high-dimensional setting, for instance computing confidence intervals and p-values for quantities of interest. Progress in this direction was achieved only over the last couple of years. In particular, several papers [10, 64, 41, 58, 42] develop methods to compute confidence intervals for single coordinates of the parameters vector $\theta^*$. More precisely, these methods compute intervals $J_i(\alpha)$ depending on $y, X$, of nearly minimal size, with the coverage guarantee

$$\mathbb{P}\big(\theta_i^* \in J_i(\alpha)\big) \geq 1 - \alpha - o_n(1)\,. \tag{4}$$

The $o_n(1)$ term is explicitly characterized, and vanishes along sequence of instances of increasing dimensions under suitable condition on the design matrix $X$.

The fundamental idea developed in [64, 41, 58, 42] is to construct a debiased (or de-sparsified) estimator that takes the form

$$\widehat{\theta}^{\mathrm{d}} = \widehat{\theta}^{\text{Lasso}} + \frac{1}{n} M X^{\mathsf{T}} (y - X\widehat{\theta}^{\text{Lasso}})\,, \tag{5}$$

where $M \in \mathbb{R}^{p\times p}$ is a matrix that is a function of $X$, but not of $y$. While the construction of $M$ varies across different papers, the basic intuition is that $M$ should be a good estimate of the precision matrix $\Omega = \Sigma^{-1}$, where $\Sigma = \mathbb{E}\{x_1 x_1^{\mathsf{T}}\}$ is the population covariance.

Assume $\theta^*$ is $s_0$-sparse, i.e. it has only $s_0$ non-zero entries. The key result that allows the construction of confidence intervals in [64, 58, 42] is the following (holding under suitable conditions on the design matrix). If $M$ is 'sufficiently close' to $\Omega$, and the sparsity level is

$$s_0 \ll \frac{\sqrt{n}}{\log p}\,, \tag{6}$$

then $\widehat{\theta}^{\rm d}_i$ is approximately Gaussian with mean $\theta^*_i$ and variance of order $\sigma^2/n$.

The condition (6) comes as a surprise, and is somewhat disappointing. Indeed, consistent estimation using –for instance– the Lasso can be achieved under the much weaker condition $s_0 \ll n/\log p$. More specifically, in this regime, with high probability [16, 63, 9, 61, 11]

$$\big\|\widehat{\theta}^{\rm Lasso} - \theta^*\|_2^2 \le \frac{Cs_0\sigma^2}{n}\,\log p\,. \tag{7}$$

This naturally leads to the following question:

> *Does the debiased estimator have a Gaussian limit under the weaker condition $s_0 \ll n/\log p$?*

Let us emphasize that the key technical challenge here does not lie in the fact that $M$ is not a good estimate of the precision matrix $\Omega$. Of course, if $M$ is not close to $\Omega$, then $\widehat{\theta}^{\rm d}$ will not have a Gaussian limit. However *earlier proofs [64, 58, 42] cannot establish the Gaussian limit for $s_0 \gtrsim \sqrt{n}/\log p$, even if $\Omega$ is known and we set $M = \Omega$.* Even the idealized case where the columns of $X$ are known to be independent and identically distributed (i.e. $\Omega = {\rm I}$) is only understood in the asymptotic limit $s_0, n, p \to \infty$ with $s_0/p$, $n/p$ having constant limits in $(0, 1)$ [41].

In order to describe the challenge, let us set $M = \Omega$, and recall the common step of the proofs in [64, 58, 42]. Using the definitions (2), (5), we get

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) &= \sqrt{n}(\widehat{\theta}^{\rm Lasso} - \theta^*) + \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}}(X\theta^* + w - X\widehat{\theta}^{\rm Lasso})\\ &= \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}} w + \sqrt{n}(\Omega\widehat{\Sigma} - {\rm I})(\theta^* - \widehat{\theta}^{\rm Lasso})\,,\end{aligned} \tag{8}$$

where $\widehat{\Sigma} = X^{\mathsf{T}}X/n \in \mathbb{R}^{p\times p}$ is the empirical design covariance. Since $w \sim {\sf N}(0, \sigma^2{\rm I}_n)$, it is easy to see that vector $\Omega X^{\mathsf{T}} w/\sqrt{n}$ has Gaussian entries of variance of order one. In order for $\widehat{\theta}^{\rm d}$ to be approximately Gaussian, we need the second term (which can be interpreted as a bias) to vanish. Earlier papers [64, 58, 42] address this by a simple $\ell_1$-$\ell_\infty$ bound. Namely (denoting by $|Q|_\infty$ the maximum absolute value of any entry of matrix $Q$):

$$
\begin{aligned}
\Big\|\sqrt{n}(\Omega\widehat{\Sigma}-\mathrm{I})(\theta^*-\widehat{\theta}^{\mathrm{Lasso}})\Big\|_\infty &\le \sqrt{n}|\Omega\widehat{\Sigma}-\mathrm{I}|_\infty\|\theta^*-\widehat{\theta}^{\mathrm{Lasso}}\|_1 \\
&\le \sqrt{n}\times C\sqrt{\frac{\log p}{n}}\times Cs_0\sigma\sqrt{\frac{\log p}{n}} \\
&\le C^2\sigma\frac{s_0\log p}{\sqrt{n}}\,,
\end{aligned}
\tag{9}
$$

where the bound $|\Omega\widehat{\Sigma}-\mathrm{I}|_\infty \le C\sqrt{(\log p)/n}$ follows from standard concentration arguments, and the bound on $\|\theta^*-\widehat{\theta}^{\mathrm{Lasso}}\|_1$ is order-optimal and is proved, for instance, in [9, 11].

This simple argument implies that the debiased estimator is approximately Gaussian if the upper bound in Eq. (9) is negligible, i.e. if $s_0 = o(\sqrt{n}/\log p)$. We see therefore that this requirement is not imposed as to control the error in estimating $\Omega$. It instead follows from the simple $\ell_1$-$\ell_\infty$ bound *even if* $\Omega$ *is known.*

1.2. *Main results.* The above exposition should clarify that the $\ell_1-\ell_\infty$ bound is quite conservative. Considering the $i$-th entry in the bias vector $\mathsf{bias} = (\Omega\widehat{\Sigma}-\mathrm{I})(\theta^*-\widehat{\theta}^{\mathrm{Lasso}})$, the $\ell_1$-$\ell_\infty$ bound controls it as $|\mathsf{bias}_i| \le \|(\Omega\widehat{\Sigma}-\mathrm{I})_{i,\cdot}\|_\infty\|\theta^*-\widehat{\theta}^{\mathrm{Lasso}}\|_1$. This bound would be accurate only if the signs of the entries $(\theta^*_j-\widehat{\theta}^{\mathrm{Lasso}}_j)$ were aligned to the signs $(\Omega\widehat{\Sigma}-\mathrm{I})_{i,j}$, $j\in\{1,\dots,p\}$. While intuitively this is quite unlikely, it is difficult to formalize this intuition; Note that in a random design setting, the terms $(\Omega\widehat{\Sigma}-\mathrm{I})_{i,\cdot}$ and $\theta^*-\widehat{\theta}^{\mathrm{Lasso}}$ are highly dependent: $\widehat{\theta}^{\mathrm{Lasso}}$ is a deterministic function of the random pair $(X,w)$, while $(\Omega\widehat{\Sigma}-\mathrm{I}) = (\Omega XX^{\mathsf{T}}/n-\mathrm{I})$ is a function of $X$.

Our main result overcomes this technical hurdle via a careful analysis of such dependencies. We follow a leave-one-out proof technique. Roughly speaking, in order to understand the distribution of the $i$-th coordinate of the debiased estimator $\widehat{\theta}^{\mathrm{d}}_i$, we consider a modified problem in which column $i$ is removed from the design matrix $X$. We then study the consequences of adding back this column, and bound the effect of this perturbation. An outline of this proof strategy is provided in Section 6.1.

We state below a simplified version of our main result, referring to Theorem 3.8 below for a full statement, including technical conditions.

THEOREM 1.1 (Known covariance). *Consider the linear model (2) where $X$ has independent Gaussian rows, with zero mean and covariance $\Sigma =$*

$\Omega^{-1}$. *Assume that* $\Sigma$ *satisfies the technical conditions stated in Theorem 3.8. Define the debiased estimator* $\widehat{\theta}^{\rm d}$ *via Eq. (5) with* $M = \Omega$ *and* $\widehat{\theta}^{\rm Lasso} = \widehat{\theta}^{\rm Lasso}(y, X; \lambda)$ *with* $\lambda = 8\sigma\sqrt{(\log p)/n}$.

*If* $n, p \to \infty$ *with* $s_0 = o(n/(\log p)^2)$, *then we have*

$$\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) = Z + o_P(1)\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,. \tag{10}$$

*Here* $o_P(1)$ *is a (random) vector satisfying* $\|o_P(1)\|_\infty \to 0$ *in probability as* $n, p \to \infty$, *and* $Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)$ *means that the conditional distribution of* $Z$ *given* $X$ *is centered Gaussian, with the stated covariance.*

REMARK 1.2. The more complete statement of this result, Theorem 3.8 provides explicit non-asymptotic bounds on the error term $o_P(1)$, In particular $\|o_P(1)\|_\infty$ turns out to be of order $\sqrt{s_0/n}\,(\log p)$ with probability converging to one as $n, p \to \infty$.

Theorem 1.1 raises an important question: *Does the Gaussian limit hold even if* $M$ *is an imperfect estimate of* $\Omega$*?*

If the precision matrix $\Omega$ is sufficiently structured, then it can be reliably estimated from the design matrix $X$. Both [64] and [58] assume that $\Omega$ is sparse, and use the node-wise Lasso to construct an estimate $\widehat{\Omega}$ [48]. They then set $M = \widehat{\Omega}$.

We followed the same procedure and hence generalized Theorem 1.1 to the setting of unknown, sparse precision matrix. We state here a simplified version of this result, deferring to Theorem 3.13 for a more technical statement including non-asymptotic probability bounds.

THEOREM 1.3 (Unknown covariance). *Consider the linear model (2) where* $X$ *has independent Gaussian rows with precision matrix* $\Omega$, *satisfying the technical conditions of Theorem 1.1 (stated in Theorem 3.8). Define the debiased estimator* $\widehat{\theta}^{\rm d}$ *via Eq. (5) with* $\widehat{\theta}^{\rm Lasso} = \widehat{\theta}^{\rm Lasso}(y, X; \lambda)$, $\lambda = 8\sigma\sqrt{(\log p)/n}$, *and* $M = \widehat{\Omega}$ *computed through node-wise Lasso (see Section 3.3).*

*Let* $s_\Omega$ *the maximum number of non-zero entries in any row of* $\Omega$. *If* $n, p \to \infty$ *with* $s_0 = o(n/(\log p)^2)$ *and* $\min(s_\Omega, s_0) = o(\sqrt{n}/\log p)$, *then we have*

$$\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) = Z + o_P(1)\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,, \tag{11}$$

*where* $o_P(1)$ *is a (random) vector satisfying* $\|o_P(1)\|_\infty \to 0$ *in probability as* $n, p \to \infty$.

REMARK 1.4. As mentioned above, this version of the debiased estimator can be constructed entirely from data. The only unspecified steps are the choice of the regularization parameter $\lambda$, and the estimation of the noise level $\sigma$. These can be addressed as in [64, 58, 42] without changes in the sparsity condition. we will further discuss these points below.

REMARK 1.5. The sparsity condition $\min(s_0, s_\Omega) = o(\sqrt{n}/\log p)$ nicely illustrates the practical improvement implied by our more refined analysis. If the sparsity of the precision matrix is larger than the sparsity of $\theta^*$, we recover the condition $s_0 = o(\sqrt{n}/\log p)$ which is assumed in the results of [64, 58]. (Note that [42] obtain the same condition without sparsity assumption on $\Omega$.) In this regime, our improved analysis does not bring any advantage, since the bottleneck is due to the inaccurate estimation of $\Omega$.

On the other hand, if the precision matrix is sparser, we obtain a much weaker condition on the coefficients $\theta^*$. In particular, if $s_\Omega = o(\sqrt{n}/\log p)$, then the condition on $s_0$ is relaxed into a nearly optimal condition $s_0 = o(n/(\log p)^2)$.

It is instructive to compare this with the past progress in sparse estimation and compressed sensing. In that context, earlier work based on incoherence conditions [26, 23] implied accurate reconstruction from a number of random samples scaling quadratically in the number of non-zero coefficients. Subsequent progress was based on the restricted isometry property [15, 16], and established accurate reconstruction from a linear number of measurements.

1.3. *Extensions and applications.* In this Section, we discuss a few directions for extending this result along with potential applications.

**Sample splitting.** An alternative approach to avoid the $\ell_1$-$\ell_\infty$ bound in Eq. (9) is to modify the definition of debiased estimator in Eq. (5), using sample-splitting. Roughly speaking, we can split the same in two batches of size $n/2$. One batch is then used to estimate $\widehat{\theta}^{\rm Lasso}$ and the other batch for $y$ and $X$ appearing in Eq. (5) (and possibly for computing $M$).

In the Supplementary Material [43], we discuss this method in greater detail. This approach is subject to variations due to the random splitting, and does not make use of part of half of the response variables. While it provides a viable alternative, it is not the focus of the present work.

**Confidence intervals.** Theorem 1.3 (and its formal version, Theorem 3.13) allows the construction of confidence intervals using the same general procedure as in [64, 58, 42]. Namely, we construct the debiasing matrix $M$ from the design matrix $X$, and an estimate $\widehat{\sigma}$ of the noise variance. Then, for a significance level $\alpha \in (0, 1)$, we form the following confidence interval for parameter $\theta_i$:

$$J_i(\alpha) \equiv [\widehat{\theta}_i^{\mathrm{d}} - \delta(\alpha, n), \widehat{\theta}_i^{\mathrm{d}} + \delta(\alpha, n)] \tag{12}$$

$$\delta(\alpha, n) \equiv \Phi^{-1}(1-\alpha/2)\frac{\widehat{\sigma}}{\sqrt{n}}(M\widehat{\Sigma}M^{\mathsf{T}})_{i,i}^{1/2}, \tag{13}$$

where $\Phi(x) \equiv \int_{-\infty}^{x} e^{-t^2/2}\mathrm{d}t/\sqrt{2\pi}$ is the Gaussian distribution. Section 3.3 presents a formal analysis of this procedure. A straightforward generalization also allows to compute p-values for the null hypothesis $H_{0,i}:\ \theta_i^* = 0$.

**Noise level and regularization.** The construction of the confidence interval $J_i(\alpha)$ in Eqs. (12), (13) requires a suitable choice of the regularization parameter $\lambda$, and an estimate of the noise level $\widehat{\sigma}$. The same difficulty was present in [64, 58, 42]. The approaches used there (for instance, using the scaled Lasso [55]) can be followed in the present case as well. Under the assumptions of Theorem 1.1, the same proofs of [42] show that the additional error due to the choice of $\lambda$ and $\widehat{\sigma}$ are negligible.

**Semi-supervised learning.** In some applications, the precision matrix $\Omega$ can be estimated more accurately thanks to additional information. For instance, in semi-supervised learning, the statistician is given additional samples $\overline{x}_1, \overline{x}_2, \dots, \overline{x}_N \in \mathbb{R}^p$ with the same distribution as the $\{x_i\}_{1\le i\le n}$. For these 'unlabeled' samples, the response variable is unknown. There are indeed many applications in which acquiring the response variable is much more challenging than capturing the covariates [17], and therefore $N \gg n$ or even $N \gg p$. In this setting, we can estimate $\Omega$ more accurately from $\{\overline{x}_i\}_{1\le i\le N}$, then use this estimate to construct $M$.

**Non-Gaussian designs.** We expect that generalization of Theorem 1.1 and Theorem 1.3 should hold for a broad class of random designs with independent sub-Gaussian rows, although new proof ideas are required. The main technical challenge in extending the present approach is to generalize the leave-one-out construction. As discussed in Section 6.1, when studying the effect of modifying column $i$, we need to account for dependencies between columns. For Gaussian designs, these dependencies are fully captured by the design covariance $\Sigma$.

Note that the Gaussian assumption holds in the context of estimating Gaussian graphical models. This is itself a broad topic that attracted significant interest, since the seminal work of [48]. Remarkably, recent contributions have shown the utility of debiasing methods in this context [36, 18, 35].

1.4. *Organization and contributions.* The rest of the paper presents the following contributions:

1. Section 3. We state formally our Gaussian limit theorems, and use them to construct valid confidence intervals, of nearly optimal size. In particular, our results subsume (and improve) all previously known results on the debiased estimator for Gaussian designs.
2. Section 4. We establish a minimax lower bound on the $\ell_\infty$ norm of the non-Gaussian component in $\widehat{\theta}^{\rm d}$. This implies that our Gaussian limit theorems cannot be substantially improved.
3. Section 5. Apart from the construction of confidence intervals, our Gaussian limit theorems have several fundamental implications. We discuss a a few examples, that we consider particularly interesting. In particular, we construct a thresholded Lasso estimator that is minimax optimal up to a factor $(1+o_n(1))$ (an alternative approach to the same problem was recently proposed in [54]).

Section 2 discusses relations with earlier work in this area. Outlines of the proofs of the main theorems are given in Section 6 with most of the technical work deferred to the Supplementary Material [43].

**2. Related work.** A parallel line of research develops methods for performing valid inference after a low-dimensional model is selected for fitting high-dimensional data [45, 34, 56, 20]. The resulting significance statements are typically conditional on the selected model. In contrast, here we are interested in classical (unconditional) significance statements: the two approaches are broadly complementary.

The focus of the present paper is assessing statistical significance, such as confidence intervals, for single coordinates in the parameters vector $\theta^*$ and more generally for small groups of coordinates. Other inference tasks are also interesting and challenging in high-dimension, and were the object of recent investigations [4, 3, 37, 38, 39]. In particular, [39] uses the idea of debiased estimator to construct an $\ell_\infty$ projection statistic for testing null hypothesis of form $H_0 : \theta_0 \in \Omega_0$ versus alternative $H_A : \theta_0 \notin \Omega_0$, for a general set $\Omega_0 \subset \mathbb{R}^p$. This framework encompasses testing whether the parameter lies in a convex cone, testing the signal strength, testing arbitrary functionals of the parameter, and testing adaptive hypothesis, among many other hypotheses.

Sample splitting provides a general methodology for inference in high dimension [60, 49]. As mentioned above, sample splitting can also be used to define a modified debiased estimator, see Supplementary Material [43]. However sample splitting techniques typically use only part of the data for inference, and are therefore sub-optimal. Also, the result depend on the random split of the data.

A method for inference without assumptions on the design matrix was developed in [47]. The resulting confidence intervals are typically quite conservative.

The debiasing method was developed independently from several points of view [10, 64, 41, 58, 42]. The present authors were motivated by the AMP analysis of the Lasso [29, 6, 7, 5], and by the Gaussian limits that this analysis implies. In particular [41] used those techniques to analyze standard Gaussian designs (i.e. the case $\Sigma = \mathrm{I}$) in the asymptotic limit $n, p, s_0 \to \infty$ with $s_0/p$, $n/p$ constant. In this limit, the debiased estimator was proven to be asymptotically Gaussian provided $s_0 \le C\, n/\log(p/s_0)$ (for a universal constant $C$). This sparsity condition is even weaker than the one of Theorem 1.1 (or Theorem 3.8), but the result of [41] only holds asymptotically. Also [41] proved Gaussian convergence in a weaker sense than the one established here, implying coverage of the constructed confidence intervals only 'on average' over the coordinates $i \in \{1, \dots, p\}$.

A non-asymptotic result under weaker sparsity conditions, and for designs with dependent columns, was proved in [40]. However, this only establishes gaussianity of $\widehat{\theta}_i^{\rm d}$ for most of the coordinates $i \in \{1, \dots, p\}$. Here we prove a significantly stronger result holding uniformly over $i \in \{1, \dots, p\}$.

Most of the work on statistical inference in high-dimensional models has been focused so far on linear regression. The debiasing method admits a natural extension to generalized linear models that was analyzed in [58]. Robustness to model misspecification was studied in [12]. An R-package for inference in high-dimension that uses the node-wise Lasso is available [21]. An R implementation of the method [42] (which does not make sparsity assumptions on $\Omega$) is also available[1].

## 3. Main results: Gaussian limit theorems.

3.1. *General notations.* We use $e_i$ to refer to the $i$-th standard basis element, e.g., $e_1 = (1, 0, \dots, 0)$. For a vector $v$, $\mathrm{supp}(v)$ represents the positions of nonzero entries of $v$. Further, $\mathrm{sign}(v)$ is the vector with entries $\mathrm{sign}(v)_i = +1$ if $v_i > 0$, $\mathrm{sign}(v)_i = -1$ if $v_i < 0$, and $\mathrm{sign}(v)_i = 0$ otherwise. For a matrix $M \in \mathbb{R}^{n \times p}$ and sets of indices $I, \subseteq \{1, \dots, p\}$ we use $M_{I,J}$ to denote the submatrix formed by rows in $I$ and columns in $J$, and we write $M_J$ to refer to the submatrix formed by columns in $J$. Likewise, for a vector $\theta$ and a subset $S$, $\theta_S$ is the restriction of $\theta$ to indices in $S$. For an integer $p \ge 1$, we use the notation $[p] = \{1, \cdots, p\}$ and the shorthand $\sim i$ for the set $[p]\backslash i$. We write $\|v\|_p$ for the standard $\ell_p$ norm of a

[1]See http://web.stanford.edu/ montanar/sslasso/.

vector $v$, i.e., $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$ and $\|v\|_0$ for the number of nonzero entries of $v$. For a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_p$ denotes it $\ell_p$ operator norm; in particular, $\|A\|_\infty = \max_{1\le i\le m}\sum_{j=1}^n |A_{ij}|$. This is to be contrasted with the maximum absolute value of any entry of $A$ that, as mentioned above, we denote by $|A|_\infty \equiv \max_{i\le m, j\le n}|A_{ij}|$. For a matrix $A$, we denote its maximum and minimum singular values by $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, respectively. If $A$ is symmetric, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are its maximum and minimum eigenvalues. Finally, for two functions $f(n)$ and $g(n)$, the notation $f(n) \gg g(n)$ means that $f$ 'dominates' $g$ asymptotically, namely, for every fixed positive $C$, there exists $n(C)$ such that $f(n) \ge Cg(n)$ for $n > n(C)$. We also use $f(n) \lesssim g(n)$ to indicate that $f$ is 'bounded' above by $g$ asymptotically, i.e., $f(n) \le Cg(n)$ for some positive constant $C$. The notations $f(n) \ll g(n)$ and $f(n) = o(g(n))$ are defined analogously, and we use $o_P(\,\cdot\,)$ to indicate asymptotic behavior in probability as the sample size $n$ tends to infinity.

We will use $c, C, \dots$ to denote generic constants that can vary from one position to the other of the paper.

3.2. *Preliminaries.* This section includes some preliminary results that are repeatedly used in our proofs. We start by some well-known results about the Lasso estimator. For the sake of simplicity, we will often use $\widehat{\theta} = \widehat{\theta}(y, X; \lambda)$ instead of $\widehat{\theta}^{\text{Lasso}}$ to denote the Lasso estimator.

We denote the rows of the design matrix $X$ by $x_1, \dots, x_n \in \mathbb{R}^p$ and its columns by $\tilde{x}_1, \dots, \tilde{x}_p \in \mathbb{R}^n$. The empirical covariance of the design $X$ is defined as $\widehat{\Sigma} \equiv (X^{\mathsf{T}}X)/n$. The population covariance will be denoted by $\Sigma$, and we let $\Omega \equiv \Sigma^{-1}$ be the precision matrix.

DEFINITION 3.1. *Given a symmetric matrix $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ and a set $S \subseteq [p]$, the corresponding* compatibility constant *is defined as*

$$\phi^2(\widehat{\Sigma}, S) \equiv \min\left\{\frac{|S|\,\langle\theta, \widehat{\Sigma}\,\theta\rangle}{\|\theta_S\|_1^2} :\ \theta \in \mathbb{R}^p,\ \|\theta_{S^c}\|_1 \le 3\|\theta_S\|_1\right\}. \tag{14}$$

*We say that $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ satisfies the* compatibility condition *for the set $S \subseteq [p]$, with constant $\phi$ if $\phi(\widehat{\Sigma}, S) \ge \phi$. We say that it holds for the design matrix $X$, if it holds for $\widehat{\Sigma} = X^{\mathsf{T}}X/n$.*

It is also useful to recall some notation for the restricted eigenvalue condition, introduced by Bickel, Ritov and Tsybakov [9]. For an integer $0 < s_0 < p$ and a positive number $L$, define $\mathcal{C}(s_0, L) \in \mathbb{R}^p$ by the following cone constraints:

$$\mathcal{C}(s_0, L) \equiv \{\theta \in \mathbb{R}^p :\ \exists S \subseteq [p],\ |S| = s_0,\ \|\theta_{S^c}\|_1 \le L\|\theta_S\|_1\}. \tag{15}$$

In high-dimension, the empirical covariance $\widehat{\Sigma}$ is singular. However, we can ask for non-singularity of $\widehat{\Sigma}$ for vectors in $\mathcal{C}(s_0, L)$. Rudelson and Zhou [52] prove a reduction principle that bounds the restricted eigenvalues of the empirical covariance in terms of those of the population covariance. We will use their result specified to the case of Gaussian matrices.

LEMMA 3.2 ([52], Theorem 16). *Suppose that $\sigma_{\min}(\Sigma) > C_{\min} > 0$ and $\sigma_{\max}(\Sigma) < C_{\max} < \infty$. Let $X \in \mathbb{R}^{n\times p}$ have independent rows drawn from $\mathsf{N}(0, \Sigma)$. Set $0 < \delta < 1$, $0 < s_0 < p$, and $L > 0$. Define the following event*

$$\mathcal{B}_\delta(n, s_0, L) \equiv$$
$$\Big\{X \in \mathbb{R}^{n\times p} : (1-\delta)\sqrt{C_{\min}} \le \frac{\|Xv\|_2}{\sqrt{n}\|v\|_2} \le (1+\delta)\sqrt{C_{\max}}\,,\ \forall v \in \mathcal{C}(s_0, L)\ s.t.\ v \neq 0\Big\}\,.$$

*Then, there exists a constant $c_1 = c_1(L)$ such that, for sample size $n \ge c_1 s_0 \log(p/s_0)$, we have*

$$\mathbb{P}(\mathcal{B}_\delta(n, s_0, L)) \ge 1 - 2e^{-\delta^2 n}\,. \tag{16}$$

REMARK 3.3. Fix $S \subseteq [p]$ with $|S| = s_0$. Under the event $\mathcal{B}_\delta(n, s_0, 3)$, we have

$$\phi^2(\widehat{\Sigma}, S) \ge \min_{\theta\in\mathcal{C}(s_0,3)} \frac{s_0\langle\theta, \widehat{\Sigma}\theta\rangle}{\|\theta_S\|_1^2} \ge \min_{\theta\in\mathcal{C}(s_0,3)} \frac{\langle\theta, \widehat{\Sigma}\theta\rangle}{\|\theta_S\|_2^2} \ge (1-\delta)^2 C_{\min}\,,$$

where the second inequality follows from Cauchy-Schwartz inequality.

We next introduce the event

$$\tilde{\mathcal{B}}(n, p) \equiv \Big\{w \in \mathbb{R}^n :\ \frac{1}{n}\|X^{\mathsf{T}} w\|_\infty \le \sigma\sqrt{\frac{6\log p}{n}}\Big\}\,. \tag{17}$$

On $\tilde{\mathcal{B}}(n, p)$ we can control the randomness due to the measurement noise. A well-known union bound argument shows that $\tilde{\mathcal{B}}(n, p)$ has large probability (see, for instance, [11]).

LEMMA 3.4 ([11], Lemma 6.2). *Suppose that $\widehat{\Sigma}_{ii} \le 1$ for $i \in [p]$. Then we have*

$$\mathbb{P}(\tilde{\mathcal{B}}(n, p)) \ge 1 - 2p^{-2}\,.$$

The following Lemma states that the Lasso estimator is sparse. Its proof is given in the Supplementary Material [43].

LEMMA 3.5. *Consider the Lasso selector $\widehat{\theta}$ with $\lambda = \kappa\sigma\sqrt{\log p/n}$, for a constant $\kappa \geq 8$. On the event $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, s_0, 3)$, the following holds:*

$$|\widehat{S}| < C_* s_0\,, \tag{18}$$

*with*

$$C_* \equiv \frac{16 C_{\max}}{(1-\delta)^2 C_{\min}}\,. \tag{19}$$

Our next Lemma states a property of Gaussian design matrices which will be used repeatedly in our analysis. Its proof is very short and is given here for the reader's convenience.

LEMMA 3.6. *Let $v_i = X\Omega e_i$. Then $v_i$ and $X_{\sim i}$ are independent.*

PROOF. Define $u = \Omega e_i$ and fix $j \neq i$. Recall that $\tilde{x}_\ell$ denotes the $\ell$-th column of $X$. We write $v_i = \sum_{\ell=1}^p \tilde{x}_\ell u_\ell$ and

$$\begin{aligned}\mathbb{E}(v_i \tilde{x}_j^{\mathsf{T}}) &= \sum_{\ell=1}^p u_\ell \mathbb{E}(\tilde{x}_\ell \tilde{x}_j^{\mathsf{T}}) = \sum_{\ell=1}^p u_\ell \Sigma_{\ell j} \mathrm{I}_{n\times n} \\ &= \sum_{\ell=1}^p \Omega_{\ell i}\Sigma_{\ell j}\mathrm{I}_{n\times n} = (\Omega\Sigma)_{ij}\mathrm{I}_{n\times n} = 0\,,\end{aligned}$$

where the last step holds since $i \neq j$. Since $v_i$ and $\tilde{x}_j$ are jointly Gaussian, this implies that they are independent. □

We finally introduce some parameters that are used in stating our main theorems. For an integer $k$ and an invertible matrix $A \in \mathbb{R}^{p\times p}$, we define $\rho(A,k)$ as follows:

$$\rho(A,k) \equiv \max_{T\subseteq[p], |T|\leq k} \|A_{T,T}^{-1}\|_\infty\,, \tag{20}$$

where we adopt the convention $A_{T,T}^{-1} = (A_{T,T})^{-1}$ and recall that $\|\cdot\|_\infty$ denotes the $\ell_\infty$ operator norm (maximum $\ell_1$ norm of the rows). It is clear that $\rho(A,k)$ is non-decreasing in $k$.

LEMMA 3.7. *Assume an invertible matrix $A$. For every $1 \leq k \leq p$, we have*

$$\rho(A,k) \leq \min\left(\|A^{-1}\|_\infty, \frac{\sqrt{k}}{\sigma_{\min}(A)}\right). \tag{21}$$

Lemma 3.7 is proved in the Supplementary Material [43].

3.3. *Statement of main theorems.* In our first theorem, we assume that the precision matrix $\Omega \equiv \Sigma^{-1}$ is available and we set $M = \Omega$. We prove the corresponding debiased estimator is asymptotically unbiased provided that $n \gg s_0(\log p)^2$.

3.3.1. *Known covariance.*

THEOREM 3.8 (Known covariance). *Consider the linear model* (2) *where* $X$ *has independent Gaussian rows, with zero mean and covariance* $\Sigma$ *and* $\theta^*$ *is* $s_0$*-sparse. Suppose that* $\Sigma$ *satisfies the following conditions:*

- $(i)$ *For* $i \in [p]$, *we have* $\Sigma_{ii} \le 1$.
- $(ii)$ *We have* $\sigma_{\min}(\Sigma) > C_{\min} > 0$ *and* $\sigma_{\max}(\Sigma) < C_{\max}$ *for some constants* $C_{\min}$ *and* $C_{\max}$.
- $(iii)$ *Define* $C_0 \equiv (32C_{\max}/C_{\min}) + 1$. *We have* $\rho(\Sigma, C_0 s_0) \le \rho$, *for some constant* $\rho > 0$.

*Let* $\widehat{\theta}$ *be the Lasso estimator defined by* (3) *with* $\lambda = \kappa\sigma\sqrt{(\log p)/n}$, *for* $\kappa \in [8, \kappa_{\max}]$. *Further, let* $\widehat{\theta}^{\rm d}$ *be defined as per equation* (5), *with* $M = \Omega \equiv \Sigma^{-1}$. *Then, there exist constants* $c, C$ *depending solely on* $C_{\min}, C_{\max}$, *and* $\kappa_{\max}$, *such that, for* $n \ge \max(25 \log p, cs_0 \log(p/s_0))$ *the following holds true:*

$$\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) = Z + R\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)\,, \tag{22}$$

$$\mathbb{P}\Big(\|R\|_\infty \ge C\rho\sigma\sqrt{\frac{s_0}{n}}\log p\Big) \le 2pe^{-c_* n/s_0} + pe^{-n/1000} + 6p^{-2}\,, \tag{23}$$

*with* $c_* \equiv C_{\min}/16$.

The proof of this theorem is presented in Section 6.

This theorem states that if the sample size satisfies $n = \Omega(s_0 \log p)$, then the maximum size of the 'bias' $R_i$ over $i \in [p]$ is bounded by

$$\|R\|_\infty = O_P\Big(\sqrt{\frac{s_0}{n}}\log p\Big)\,.$$

On the other hand, each entry of the 'noise term' $Z_i$ has variance $\sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}$. Applying Lemma 7.2 in [40], we have $|\Omega\widehat{\Sigma}\Omega - \Omega|_\infty = o_P(1)$. Therefore, $\min_{i\in[p]}(\Omega\widehat{\Sigma}\Omega)_{ii} \ge \min_{ii}\Omega_{ii} - o_P(1)$ is of order one because $\Omega_{ii} \ge C_{\max}^{-1}$. Hence, $|R_i|$ is much smaller than $Z_i$ for $n \gg s_0(\log p)^2$. We summarize this observation in the remark below.

REMARK 3.9. (*Discussion of the assumptions on* $\Sigma$.) Assumption $(i)$ sets the normalization of the design matrix. Assumptions $(ii)$ on the eigenvalues of $\Sigma$ is common in high-dimensional models. Further, note that by Assumption $(ii)$ and invoking Lemma (3.7), we have $\rho(\Sigma, C_0 s_0) \le \sqrt{C_0 s_0}/C_{\min}$. Using this bound for $\rho$ in Eq. (23), we recover the bound $\|R\|_\infty \lesssim s_0 \log p/\sqrt{n}$ which is established in previous work [64, 58, 42]. Note that this bound on the bias does not require Assumption $(iii)$ (namely, that $\rho$ is a bounded constant). However, Theorem 3.8 asserts that, if $\rho$ is a constant (Assumption $(iii)$), we have a sharper bound on the bias, namely $\|R\|_\infty \lesssim \sqrt{s_0/n}\,\log p$.

A large family of covariance matrices satisfy conditions of Theorem 3.8. Examples include block diagonal matrices where the size of blocks are bounded, and circulant matrices, where $\Sigma_{i,j} = r^{|i-j|}$, for some $r \in (0,1)$.

COROLLARY 3.10. *Under the assumptions of Theorem 3.8, if $s_0 \ll n/(\log p)^2$, then $\widehat{\theta}^{\rm d}$ is normally distributed. More precisely, let $\widehat{\sigma} = \widehat{\sigma}(y, X)$ be an estimator of the noise level satisfying, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{\theta^*\in\mathbb{R}^p;\, \|\theta^*\|_0\le s_0} \mathbb{P}\Big(\Big|\frac{\widehat{\sigma}}{\sigma} - 1\Big| \ge \varepsilon\Big) = 0\,. \tag{24}$$

*If $s_0 \ll n/(\log p)^2$ and $p/n$ then, for all $x \in \mathbb{R}$, we have the following almost surely*

$$\lim_{n\to\infty} \sup_{\theta_0\in\mathbb{R}^p;\, \|\theta^*\|_0\le s_0} \left|\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\rm d}_i - \theta^*_i)}{\widehat{\sigma}[\Omega\widehat{\Sigma}\Omega^{\sf T}]^{1/2}_{i,i}} \le x\right\} - \Phi(x)\right| = 0\,. \tag{25}$$

Proof of Corollary 3.10 is given in the Supplementary Material [43].

There are several proposals for a consistent estimator of $\sigma$. A non-exhaustive list includes [31, 32, 53, 62, 55, 8, 51, 22, 33, 4]. For concreteness, we use the the scaled Lasso [55] given by

$$\{\widehat{\theta}, \widehat{\sigma}\} \equiv \arg\min_{\theta\in\mathbb{R}^p, \sigma>0} \Big\{\frac{1}{2\sigma n}\|Y - X\theta\|_2^2 + \frac{\sigma}{2} + \bar{\lambda}\|\theta\|_1\Big\}\,. \tag{26}$$

The following proposition shows that the scaled Lasso estimate $\widehat{\sigma}$ satisfies the consistency criterion (24).

LEMMA 3.11. *Under the assumptions of Theorem 3.8, let $\widehat{\sigma}$ be the scaled Lasso estimator of the noise level, see Equation (26), with $\bar{\lambda} = 10\sqrt{(2\log p)/n}$. Then $\widehat{\sigma}$ satisfies Equation (24).*

We refer to our earlier work [42, Appendix C] for the proof of Lemma 3.11.

Armed with the distributional characterization of $\widehat{\theta}^{\rm d}$, given by (25), we can construct asymptotically valid confidence intervals for each parameter

$\theta_i^*$. Indeed, for the confidence interval $J_i(\alpha)$ described by Equations. (12), (13), we have the following coverage guarantee

$$\lim_{n\to\infty} \mathbb{P}\big(\theta_i^* \in J_i(\alpha)\big) = 1 - \alpha\,. \tag{27}$$

Let us emphasize that the coverage probability is taken with respect to the random noise vector $w$ as well as the design matrix $X$. It would be interesting (and important) to derive similar guarantees *conditional* on $X$.

Furthermore, in the context of hypothesis testing, we can test the null hypothesis $H_{0,i} : \theta_i^* = 0$ versus the alternative $H_{A,i} : \theta_i^* \neq 0$. We construct the two sided $p$-values

$$P_i = 2\bigg(1 - \Phi\Big(\frac{\sqrt{n}|\widehat{\theta}_i^{\rm d}|}{\widehat{\sigma}(\Omega\widehat{\Sigma}\Omega^{\sf T})_{i,i}^{1/2}}\Big)\bigg)\,. \tag{28}$$

The decision rule follows immediately: we reject $H_{0,i}$ if $P_i \le \alpha$.

REMARK 3.12. It is worth noting that the sample splitting approach, discussed in the Supplementary Material [43], does not require Assumption *(iii)* in Theorem 3.8. However as pointed in the introduction, this approach suffers from variability due to the random splitting and does not make use of half of the response variables.

3.3.2. *Unknown covariance.* We next generalize our result to the case of unknown covariance, where following [64, 58] we construct the debiasing matrix $M$ using node-wise Lasso on matrix $X$. For reader's convenience, we first describe this construction.

For $i \in [p]$, we define the vector $\hat{\gamma}_i = (\hat{\gamma}_{i,j})_{j\in[p]\setminus i} \in \mathbb{R}^{p-1}$ by performing sparse regression of the $i$-th column of $X$ against all the other columns. Formally

$$\hat{\gamma}_i(\widetilde{\lambda}) = \arg\min_{\gamma\in\mathbb{R}^p}\Big\{\frac{1}{2n}\|\tilde{x}_i - X_{\sim i}\gamma\|_2^2 + \widetilde{\lambda}\|\gamma\|_1\Big\}\,, \tag{29}$$

where $X_{\sim i}$ is the sub-matrix obtained by removing the $i$-th column (and columns indexed by $[p]\setminus i$). Also define

$$\widehat{C} = \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{bmatrix}\,, \tag{30}$$

and let

$$\widehat{T}^2 = \mathrm{diag}(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2), \qquad \hat{\tau}_i^2 = \frac{1}{n}(\tilde{x}_i - X_{\sim i}\hat{\gamma}_i)^{\mathsf{T}}\tilde{x}_i\,. \tag{31}$$

Finally, define $M = M(\widetilde{\lambda})$ by

$$M = \widehat{T}^{-2}\widehat{C}\,. \tag{32}$$

THEOREM 3.13 (Unknown covariance). *Consider the linear model* (2) *where $X$ has independent Gaussian rows, with zero mean and covariance $\Sigma$. Suppose that Assumptions $(i), (ii), (iii)$ in Theorem 3.8 hold true for $\Sigma$. We further let $s_\Omega$ be the maximum sparsity of the rows of $\Omega \equiv \Sigma^{-1}$, i.e.*

$$s_\Omega \equiv \max_{i\in[p]} |\{j \neq i, \Omega_{i,j} \neq 0\}|\,. \tag{33}$$

*Let $\widehat{\theta}$ be the Lasso estimator defined by* (3) *with $\lambda = \kappa\sigma\sqrt{(\log p)/n}$, for $\kappa \in [8, \kappa_{\max}]$ and let $\widehat{\theta}^{\mathrm{d}}$ be debiased estimator with $M$ given by* (32) *and $\widetilde{\lambda} = K\sqrt{\log p/n}$ (with $K$ a suitably large universal constant).*

*Then, there exist constants $c, C$ depending solely on $C_{\min}, C_{\max}, \kappa_{\max}, K$ such that, for $n \geq c\max(s_0, s_\Omega)\log p$, the following holds true:*

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + R\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})\,, \tag{34}$$

$$\|R\|_\infty \leq C\rho\sigma\sqrt{\frac{s_0}{n}}\log p + C\sigma\min(s_0, s_\Omega)\frac{\log p}{\sqrt{n}}\,, \tag{35}$$

*with probability at least $1 - 2pe^{-c_* n/s_0} - pe^{-cn} - 6p^{-2}$, for some constants $c_*, c', c'' > 0$.*

The proof of Theorem 3.13 is deferred to Section C.

A result similar to Corollary 3.10 holds true for the case of unknown covariance. The proof is completely analogous to the one of Corollary 3.10, and hence omitted.

COROLLARY 3.14. *Let $\widehat{\sigma} = \widehat{\sigma}(y, X)$ be an estimator of the noise level satisfying Eq. (24) for any $\varepsilon > 0$.*

*Under the assumptions of Theorem 3.13, if $\min(s_0, s_\Omega) \ll \sqrt{n}/\log p$ and $s_0 \ll n/(\rho(\log p)^2)$, then for all $x \in \mathbb{R}$ we have*

$$\lim_{n\to\infty} \sup_{\theta_0\in\mathbb{R}^p;\, \|\theta^*\|_0\leq s_0} \left|\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathrm{d}}_i - \theta^*_i)}{\widehat{\sigma}[M\widehat{\Sigma}M^{\mathsf{T}}]^{1/2}_{i,i}} \leq x\right\} - \Phi(x)\right| = 0\,, \tag{36}$$

*where $M$ is given by equation* (32).

Using the above distributional characterization, we can construct confidence intervals for the individual model parameters $\theta_i^*$ as in (12), (13) with $M$ given by (32) and $\widehat{\sigma}$ given by the scaled Lasso as per (26). As mentioned above, the resulting coverage probability includes expectation with respect to both the noise and the random design, cf. Eq. (27). For hypothesis testing task, two sided $p$-values can be built similar to (28), where we replace $\Omega\widehat{\Sigma}\Omega$ with $M\widehat{\Sigma}M^{\mathsf{T}}$.

## 4. Minimax lower bound on the residual $R$.

In case that the design covariance matrix is unknown, Theorem 3.13 establishes the following high probability bound on the residual term $R$:

$$\|R\|_\infty \le C\rho\sigma\sqrt{\frac{s_0}{n}}\log p + C\sigma\,\min(s_0, s_\Omega)\frac{\log p}{\sqrt{n}}\,. \tag{37}$$

For sparse precision matrices, such that $s_\Omega \ll \sqrt{n}/(\log p)$, the residual term $\|R\|_\infty$ vanishes asymptotically under the near optimal condition $s_0 \ll n/(\log p)^2$. The question we will study in this section is whether such condition on $s_\Omega$ is necessary. To answer this question, we develop a lower bound for $\|R\|_\infty$ based on a minimax theorem for the estimation of coefficient $\theta_1$. This also clarifies the connection between our results and the ones of [14], whose general approach we build on here.

Before presenting our results we need to introduce some notations and definitions.

Consider the linear model (2) and define parameters of the form $\gamma = (\theta, \Omega, \sigma^2)$, which consists of the signal $\theta$, precision matrix $\Omega = \Sigma^{-1}$, and the noise standard deviation $\sigma$.

For $\alpha \in (0, 1)$ and a given parameter space $\Gamma$, denote by $\mathcal{I}_\alpha(\Gamma)$ the set of all $(1-\alpha)$-confidence intervals for $\theta_1$ over the entire space $\Gamma$,

$$\mathcal{I}_\alpha(\Gamma) \equiv \left\{J_\alpha(y, X) :\, \inf_{\gamma\in\Gamma}\mathbb{P}_\gamma(\theta_1 \in J_\alpha(y, X)) \ge 1-\alpha\right\}, \tag{38}$$

where $\mathbb{P}_\gamma$ is the induced probability distribution on $(y, X)$ for random gaussian design $X$ and noise realization $w$, given the fixed signal $\theta$. Here and below we focus on the first coordinate $\theta_1$ without loss of generality. For a given interval $J_\alpha(y, X) \in \mathcal{I}_\alpha(\Gamma)$, we let $\ell(J_\alpha(y, X))$ be the length of interval $J_\alpha(y, X)$ and denote by $\ell(J_\alpha(\,\cdot\,), \Gamma)$ the maximum expected length over a parameter space $\Gamma$,

$$\ell(J_\alpha(\,\cdot\,), \Gamma) = \sup_{\gamma\in\Gamma}\mathbb{E}_\gamma\{\ell(J_\alpha(y, X))\}\,, \tag{39}$$

with $\mathbb{E}_\gamma$ expectation with respect to $\mathbb{P}_\gamma$. We further define the minimax rate for the expected length of confidence intervals over $\Gamma$ as follows:

$$\ell^*_\alpha(\Gamma) = \inf_{J_\alpha(\,\cdot\,)\in\mathcal{I}_\alpha(\Gamma)} \ell(J_\alpha(\,\cdot\,), \Gamma)\,. \tag{40}$$

We next define parameter space $\Gamma(s_0, s_\Omega, \rho)$ as follows. Applying Lemma 3.7, we strengthen Condition *(iii)* as $\|\Omega\|_\infty \le \rho$ and write

$$\Gamma(s_0, s_\Omega, \rho) \equiv \Big\{\gamma = (\theta, \Omega, \sigma^2) :\, \|\theta\|_0 \le s_0, \sigma^2 \in (0, c],$$
$$(\Omega^{-1})_{ii} \le 1,\, \frac{1}{C_{\max}} < \sigma_{\min}(\Omega) \le \sigma_{\max}(\Omega) < \frac{1}{C_{\min}},\, \|\Omega\|_\infty \le \rho,$$
$$\max_{i\in[p]} |\{j \neq i, \Omega_{i,j} \neq 0\}| \le s_\Omega\Big\}\,. \tag{41}$$

Quantities $c$, $C_{\min}$ and $C_{\max} \ge 1$ are constant which do not effect the minimax rate and therefore we have not made them explicit in our notation $\Gamma(s_0, s_\Omega, \rho)$.

PROPOSITION 4.1. *Consider a debiased estimator of form* (5) *with $M$ being a function of $X$ and $\widehat\theta$ the Lasso estimator at regularization parameter $\lambda$. Further, let $R = \sqrt{n}(M\widehat\Sigma - \mathrm{I})(\widehat\theta - \theta^*)$ be the bias term and $Q = \mathrm{diag}(M\widehat\Sigma M^{\mathsf T})$ be the variance term. Suppose that there exist a choice of $M$ and $\lambda$ such that*

$$\lim_{n\to\infty} \mathbb{P}\bigg(\sup\Big\{\|R\|_\infty :\ (\theta^*, \Omega, \sigma^2) \in \Gamma(s_0, s_\Omega, \rho)\Big\} \le \Delta_n\bigg) = 1\,, \tag{42}$$

$$\lim_{n\to\infty} \mathbb{P}\bigg(\sup\Big\{\|Q\|_\infty :\ (\theta^*, \Omega, \sigma^2) \in \Gamma(s_0, s_\Omega, \rho)\Big\} \le C\bigg) = 1\,, \tag{43}$$

*for some known $\Delta_n$ and for some known constant $C$. Then, we have*

$$\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho)) \lesssim \frac{(1+\Delta_n)}{\sqrt{n}}\,. \tag{44}$$

Note that since $Q$ is a function of only $X$, the arguments $\theta^*$ and $\sigma^2$ in Equation (43) are superfluous. To establish the above upper bound, we construct a confidence interval $J^{\mathrm{d}}_\alpha$ using a debiased estimator, such that $J^{\mathrm{d}}_\alpha \in \mathcal{I}(\Gamma(s_0, s_\Omega, \rho))$. We refer to the Supplementary Material [43] for the proof of Proposition 4.1.

The next proposition provides a lower bound on $\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho))$.

PROPOSITION 4.2. *Suppose that* $\alpha \in (0, 1/2)$ *and* $s_0 \lesssim \min(p^\eta, n/\log p)$ *for some constant* $0 \le \eta < 1/2$. *Further, assume* $\rho \ge 1.02$. *The minimax expected length for* $(1-\alpha)$*-confidence intervals of* $\theta_1$ *over* $\Gamma(s_0, s_\Omega, \rho)$ *satisfies*

$$\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho)) \gtrsim \frac{1}{\sqrt{n}} + \min\left(s_0 \frac{\log p}{n}, s_\Omega \frac{\log p}{n}, \rho\sqrt{\frac{\log p}{n}}\right). \tag{45}$$

Proposition 4.2 generalizes the result of [14, Theorem 2] which shows that without the sparsity constraint on $\Omega$ and the constraint $\|\Omega\|_\infty \le \rho$, the minimax rate for expected confidence interval length is lower bounded as $\ell^*_\alpha(\Gamma(s_0, p)) \ge (1/\sqrt{n} + s_0 \log p/n)$. Proposition 4.2 provides a more refined lower bound that takes into account the sparsity structure of the precision matrix. We refer to the Supplementary Material [43]. for its proof.

By comparing the upper and lower bounds on $\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho))$, we conclude that the condition $\min(s_0, s_\Omega) \log p \lesssim \sqrt{n}$ is necessary for having $\|R\|_\infty \le \Delta_n \to 0$. If this is not the case then $\Delta_n \gtrsim \min(s_0, s_\Omega) \log p/\sqrt{n}$.

In particular, in order to get $\Delta_n = o(1)$ at a nearly optimal condition $s_0 \ll n/(\log p)^2$, we need the precision matrix to be sparse with $s_\Omega \lesssim \sqrt{n}/(\log p)$.

REMARK 4.3. By using the bound (37) for $\Delta_n$ in Proposition 4.1, we obtain the following upper bound on $\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho))$:

$$\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho)) \lesssim \frac{1}{\sqrt{n}} + \rho \frac{\sqrt{s_0}}{n} \log p + \min(s_0, s_\Omega) \frac{\log p}{n}. \tag{46}$$

By comparing the above bound with the lower bound established in Proposition 4.2, we see that the proposed upper and lower bounds do not exactly match. It is worth noting that we derive the lower bound on the parameter space $\Gamma(s_0, s_\Omega, \rho)$, while in deriving the upper bound (Theorem 3.13), we assume that $\rho(\Sigma, C_0 s_0) \le \rho$, and by Lemma 3.7 this assumption is implied if $\|\Omega\|_\infty \le \rho$. Therefore, the upper bound is obtained for a larger class than $\Gamma(s_0, s_\Omega, \rho)$ and this might be a contributing factor to the mismatch of the proposed upper and lower bounds on $\ell^*(\Gamma(s_0, s_\Omega, \rho))$.

## 5. Other applications.

Our main results, Theorem 3.8 and Theorem 3.13 establish a Gaussian limit for the debiased Lasso estimator. While our main motivation was the construction of confidence intervals for single coordinates of the parameter vector, we want to emphasize that the Gaussian limit has other important applications. We illustrate this point using three examples: $(i)$ We establish a characterization of the Lasso estimator in terms of a certain denoising problem. $(ii)$ We develop a new thresholded Lasso estimator and provide a tight characterization of its $\ell_2$ risk. In the case of standard Gaussian designs this approach is minimax optimal up to a factor $1+o_n(1)$. $(iii)$ We prove that the celebrated Stein's Unbiased Estimate of the prediction risk [30] is consistent in high dimension an unbiased estimator, for standard Gaussian designs.

5.1. *A probabilistic approximation result for the Lasso.* As a first consequence of our main theorem, we obtain a precise approximation result for the Lasso estimator. In order to state this result, let $\eta_\Sigma : \mathbb{R}^p \to \mathbb{R}^p$ be the proximal operator defined by

$$\eta_\Sigma(z) \equiv \arg\min_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{2}\big\|\Sigma^{1/2}(\theta - z)\big\|_2^2 + \lambda\|\theta\|_1 \right\}. \tag{47}$$

Note that the minimizer is always unique because $\Sigma$ is strictly positive definite. In the case $\Sigma = \mathrm{I}$, $\eta_\Sigma$ coincides with component-wise soft thresholding at level $\lambda$. More generally, $\eta_\Sigma(\,\cdot\,)$ can be viewed as a denoising operator associated to the problem of estimating $\theta^*$ from the noisy observation $z = \theta^* + \tilde{w}$, where $\tilde{w}$ has covariance $\Sigma$. Our next theorem connects the Lasso to this denoising problem and its proof is given in the Supplementary Material [43].

THEOREM 5.1. *Consider the linear model (2) where $X$ has independent Gaussian rows, with zero mean and covariance $\Sigma$, satisfying the assumptions of Theorem 3.8. Further assume the following condition:*

- $(iv)$ *Letting $C_* \equiv 32C_{\max}/C_{\min}$, we assume $\|\Sigma_{T,T^c}\|_\infty \le \tilde{\rho}$ for some constant $\tilde{\rho}$ and all $T \subseteq [p]$ satisfying $|T| \le 2C_* s_0$.*

*Let $\widehat{\theta}^{\text{Lasso}} = \widehat{\theta}^{\text{Lasso}}(y, X; \lambda)$ be the Lasso estimator with $\lambda = \kappa\sigma\sqrt{(\log p)/n}$, for $\kappa \in [8, \kappa_{\max}]$. Then, there exist constants $c, \widetilde{C}$ (depending on $C_{\min}$, $C_{\max}$, $\rho$, $\tilde{\rho}$, $\kappa_{\max}$), such that for $n \ge \max(25\log p, cs_0\log(p/s_0))$, the following holds true with high probability.*

$$\Big\|\widehat{\theta}^{\text{Lasso}} - \eta_\Sigma\Big(\theta^* + \frac{1}{n}\Omega X^{\mathsf{T}} w\Big)\Big\|_2^2 \le \widetilde{C}\sigma^2\Big(\frac{s_0\log p}{n}\Big)^2. \tag{48}$$

Under the hypothesis of this theorem, the Lasso $\ell_2$ error is known to be bounded as $\|\widehat{\theta}^{\text{Lasso}} - \theta^*\|_2^2 \le C(s_0\log p)/n$ [9]. Hence, Theorem 5.1 provides a characterization of the Lasso estimator that is one order of magnitude more accurate than what available in the literature.

This characterization is particularly convenient if the population covariance has a simple structure. For instance we obtain the following immediate corollary that characterizes the $\ell_2$ error for standard designs. We defer its proof to the Supplementary Material [43].

COROLLARY 5.2. *Consider the linear model (2) where $X$ has independent Gaussian rows, with zero mean and covariance $\Sigma = \mathrm{I}$. Let $\widehat{\theta}^{\text{Lasso}} = \widehat{\theta}^{\text{Lasso}}(y, X; \lambda)$ be the Lasso estimator with $\lambda = \kappa\sigma\sqrt{(\log p)/n}$, for a constant $\kappa \ge 8$. Then, for $n \ge \max(25\log p, cs_0\log(p/s_0))$ we have*

$$\|\widehat{\theta}^{\text{Lasso}} - \theta^*\|_2^2 =$$

$$\sum_{i\in\text{supp}(\theta^*)} \mathbb{E}_Z\big\{\big[\eta(\theta_i^* + n^{-1/2}Z_i; \lambda) - \theta_i^*\big]^2\big\} + O_P\left(\sigma^2\frac{\sqrt{s_0\log p}}{n} \vee \sigma^2\Big(\frac{s_0\log p}{n}\Big)^{3/2}\right)$$

*where expectation is taken with respect to $Z_i \sim \mathsf{N}(0,1)$, and the $O_P(\,\cdot\,)$ is uniform for $\kappa \in [8, \kappa_{\max}]$.*

Let us emphasize that this is not an upper bound, but an equality up to higher order terms. It provides a connection between the Lasso mean square error and the mean square error of soft-thresholding denoising in the classical sequence model. A similar connection was anticipated –for instance– in [25, 27]. An asymptotic characterizations of the Lasso mean square error for standard Gaussian designs was first obtained in [7]. However, in the present case we recover this as a corollary of a result for general Gaussian designs, and in a non-asymptotic form.

5.2. *Minimax optimal estimation.* The analysis in the last section suggests that it is possible to reduce the estimation error through a two step procedure. For the sake of simplicity, we shall assume here that $\Sigma$ is known. Our approach can be extended to imperfectly known covariance by using Theorem 3.13, but we leave this for future work. The suggested procedure is:

(*i*) Compute the Lasso estimator $\widehat{\theta}^{\text{Lasso}} = \widehat{\theta}^{\text{Lasso}}(y, X; \lambda)$ with $\lambda = 8\sigma\sqrt{(\log p)/n}$.
(*ii*) Compute the debiased estimator $\widehat{\theta}^{\text{d}} = \widehat{\theta}^{\text{Lasso}} + n^{-1}\Omega X^{\mathsf{T}}(y - X\widehat{\theta}^{\text{Lasso}})$.
(*iii*) Compute a new estimator $\widehat{\theta}^{(2)}$ by soft thresholding $\widehat{\theta}^{\text{d}}$ component-wise, namely

$$\widehat{\theta}_i^{(2)} = \eta(\widehat{\theta}_i^{\text{d}}; \tau_i)\,, \qquad \tau_i = \sqrt{\frac{2\sigma^2\Omega_{ii}\log(p/s_0)}{n}}\,. \tag{49}$$

Here $\eta(x;\tau) \equiv (|x|-\tau)_+\text{sign}(x)$ is the scalar soft-thresholding function.

Let us emphasize that in the last step we soft-threshold at a level that is smaller than the regularization used in the Lasso. Indeed, since $\Omega_{ii} \le C_{\min}^{-1}$, we have $\tau_i = O(\sqrt{\log(p/s_0)/n})$, while $\lambda$ is of order $\sqrt{(\log p)/n}$.

THEOREM 5.3. *Consider the linear model* (2) *where* $X$ *has independent Gaussian rows, with zero mean and covariance* $\Sigma$, *satisfying the assumptions of Theorem 3.8. Further assume* $s_0 \to \infty$, $s_0/p \to 0$ *and* $(s_0(\log p)^3)/n \to 0$. *Let* $\widehat{\theta}^{(2)}$ *be the two-step estimator defined above. Then*

$$\|\widehat{\theta}^{(2)} - \theta^*\|_2^2 \le \frac{2s_0\sigma^2}{n}\log(p/s_0)\left(\frac{1}{s_0}\sum_{i\in\text{supp}(\theta^*)}\Omega_{ii}\right)\big(1 + o_P(1)\big)\,. \tag{50}$$

We refer to the Supplementary Material [43] for the proof of Theorem 5.3. Note that, in the case $\Sigma = \mathrm{I}$, the right-hand side of (50) is *minimax optimal* risk, up to a factor going to one as $n, s_0, p \to \infty$ [54]. Candés and Su [54] recently proved that SLOPE achieves the same guarantee for Gaussian designs with $\Sigma = \mathrm{I}$. On one hand, the approach of [54] has the advantage of being adaptive to unknown sparsity level $s_0$. On the other, Theorem 5.3 establishes this result as a special case of a guarantee holding for more general Gaussian designs.

5.3. *SURE estimate of the prediction error.* Define the Lasso prediction error as

$$\mathsf{R}(y, X, \theta^*) \equiv \frac{1}{n}\big\|X(\widehat{\theta}^{\text{Lasso}} - \theta^*)\big\|_2^2 + \frac{1}{n}\|w\|_2^2\,. \tag{51}$$

Notice that the first term is the standard prediction error, for given design matrix $X$. The second term is the residual error that would be present even for the perfect estimator $\widehat{\theta} = \theta^*$. We include this contribution for mathematical convenience, but it is just a fixed random variable, independent of the estimator.

The naive empirical estimate for the prediction error is

$$\widehat{\mathsf{R}}(y, X) \equiv \frac{1}{n}\big\|y - X\widehat{\theta}^{\text{Lasso}}\big\|_2^2\,. \tag{52}$$

Of course we expect the empirical risk to under-estimate the actual risk. Stein's Unbiased Risk Estimate (SURE) provides a corrected estimate

$$\widehat{\mathsf{R}}_{\text{SURE}}(y, X) \equiv \frac{1}{n}\big\|y - X\widehat{\theta}^{\text{Lasso}}\big\|_2^2 + \frac{2\sigma^2}{n}\,\|\widehat{\theta}^{\text{Lasso}}\|_0\,. \tag{53}$$

This approach has a rich history for which we can only provide a few pointers. Donoho and Johnstone used SURE to develop an adaptive denoising procedure via wavelet thresholding. From the perspective of linear regression, this corresponds to $X$ being proportional to an orthogonal matrix.

Efron [30] developed a general formula for estimating the prediction error, based on Stein's ideas, and clarified the connection with classical model selection criteria such as Akaike's information criterion [1], and Mallows $C_p$ [46]. Zou, Hastie and Tibshirani [65] showed that the number of degrees of freedom (which enters Efron's formula) coincides with the number of non-zero parameters $\|\widehat{\theta}^{\rm Lasso}\|_0$. They also proved that $\widehat{\mathsf{R}}_{\rm SURE}(y,X)$ is consistent in the classical low-dimensional regime $n\to\infty$ with $p$ fixed.

To the best of our knowledge, this is the first case in which $\widehat{\mathsf{R}}_{\rm SURE}(y,X)$ is proved to be consistent in high dimension (although in a restricted setting, namely for Gaussian designs).

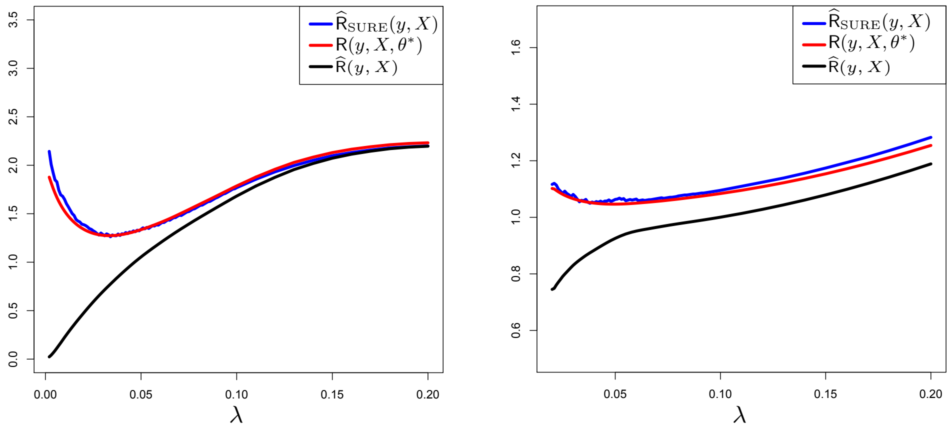
THEOREM 5.4. *Consider the linear model (2) where $X$ has independent Gaussian rows, with zero mean and identity covariance $\Sigma = {\rm I}$. Let $\widehat{\theta}^{\rm Lasso} = \widehat{\theta}^{\rm Lasso}(y,X;\lambda)$ be the Lasso estimator with $\lambda \ge 9\sigma\sqrt{(\log p)/n}$. If $n,p\to\infty$ with $s_0 = o(n/(\log p)^2)$, then there exists $\varepsilon_n \to 0$ as $n\to\infty$, such that the following holds with probability at least $1-e^{-ct^2}-o_n(1)$:*

$$\big|\widehat{\mathsf{R}}_{\rm SURE}(y,X) - \mathsf{R}(y,X,\theta^*)\big| \le \frac{t\sigma^2}{\sqrt{n}} + \frac{s_0\sigma^2\varepsilon_n}{n}\,. \tag{54}$$

Proof of Theorem 5.4 is provided in the Supplementary Material [43]. Let us emphasize a few important points:

- The error bound in Eq. (54) is of smaller order with respect to the correction in (53) which typically is of order $s_0\sigma^2/n$.
- The SURE risk estimate $\widehat{\mathsf{R}}_{\rm SURE}(y,X)$ is perfectly well defined for arbitrary design covariance $\Sigma$.
- While our proof applies to standard designs, $\Sigma = {\rm I}$, we expect the conclusion of Theorem 5.4 to hold more generally. This is also confirmed by the simulations discussed below.

In Figure 1, we present the results of a numerical simulation with $p = 5000$, $n = 1800$. We choose a subset $S\subseteq[p]$ of size $s_0 = |S| = 100$ uniformly at random and set $\theta^*_{0,i} = 0.1$ if $i\in S$ and $\theta^*_{0,i} = 0$, otherwise. The design matrix $X$ has i.i.d random rows $x_i \sim \mathsf{N}(0,\Sigma)$ with $\Sigma_{ij} = r^{|i-j|}$. We set $r = 0.1$ to illustrate a case of low correlation between predictors and $r = 0.9$ for a case of high correlation. In our simulations, we replace the noise level $\sigma$ appearing in Eq. (53) with an estimate $\widehat{\sigma}$, obtained as follows. We first run scaled Lasso and then perform least square after model selection to mitigate the estimation bias. More precisely, we use the R-package scalreg with the default value for the regularization parameter in the scaled Lasso cost function. This selects a model $\widehat{S}$. We then perform least square on $\widehat{S}$




(a) $r = 0.1$ (b) $r = 0.9$

Fig 1: Lasso prediction error $\mathsf{R}(y, X, \theta^*)$, empirical prediction error $\widehat{\mathsf{R}}(y, X)$, and SURE estimator $\widehat{\mathsf{R}}_{\rm SURE}$ curves versus $\lambda$ for the simulation setting described in Section 5.3.

to obtain an estimate $\widehat{\theta}^{\rm LS}$. The noise variance is computed as $\widehat{\sigma} = \|y - X\widehat{\theta}^{\rm LS}\|_2/\sqrt{n}$.

The agreement between $\widehat{\mathsf{R}}_{\rm SURE}(y, X)$ and $\mathsf{R}(y, X, \theta^*)$ is excellent.

Let us mention that [4] also studied estimators similar to $\widehat{\mathsf{R}}_{\rm SURE}(y, X)$, and related ideas were developed in [50] on the basis of non-rigorous but insightful statistical mechanics techniques. Other approaches to the risk estimation, e.g. [13], are based on sample-splitting, which has complementary shortcomings.

## 6. Proof of Theorem 3.8 (known covariance).

6.1. *Outline of the proof.* Fix arbitrary integer $i \in [p]$. In our analysis, we focus on the $i$-th coordinate $\theta^*_i$, and then discuss how the argument can be adjusted to apply to all the coordinates simultaneously. Our argument relies on a perturbation analysis. We let $\widehat{\theta}^{\rm p}$ be the Lasso estimator when one forces $\widehat{\theta}^{\rm p}_i = \theta^*_i$. With a slight abuse of notation, we use the representation $\theta = (\theta_i, \theta_{\sim i})$.[2] Adopting this convention, we have $\widehat{\theta}^{\rm p} = (\theta^*_i, \widehat{\theta}^{\rm p}_{\sim i})$ where

$$\widehat{\theta}^{\rm p}_{\sim i} = \arg\min_{\theta} \mathcal{L}_{y,X}(\theta^*_i, \theta)\,. \tag{55}$$

[2]Or without loss of generality one can assume $i = 1$.

Throughout, we make the convention that $\mathcal{L}_{y,X}(\theta_i^*, \theta) \equiv \mathcal{L}_{y,X}((\theta_i^*, \theta))$.

We observe that $\widehat{\theta}^{\rm p}_{\sim i}$ can be written as a Lasso estimator. Specifically, by definition of Lasso cost function we have

$$\mathcal{L}_{y,X}(\theta_i^*, \theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta_i^* - X_{\sim i}\theta\|_2^2 + \lambda|\theta_i^*| + \lambda\|\theta\|_1 \,.$$

Letting $\tilde{y} \equiv y - \tilde{x}_i\theta_i^* = w + X_{\sim i}\theta^*_{\sim i}$, we obtain

$$\widehat{\theta}^{\rm p}_{\sim i} = \arg\min_{\theta} \mathcal{L}_{\tilde{y}, X_{\sim i}}(\theta)\,. \tag{56}$$

Let $v_i = X\Omega e_i$ and expand $\widehat{\theta}^{\rm d}_i - \theta_i^*$ as follows:

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}^{\rm d}_i - \theta_i^*) &\equiv \sqrt{n}\widehat{\theta}_i + \frac{1}{\sqrt{n}} e_i^{\mathsf{T}}\Omega X^{\mathsf{T}}(y - X\widehat{\theta}) - \sqrt{n}\theta_i^* \\
&= \sqrt{n}\widehat{\theta}_i + \frac{v_i^{\mathsf{T}}}{\sqrt{n}}\Big[w + \tilde{x}_i(\theta_i^* - \widehat{\theta}_i) + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}_{\sim i})\Big] - \sqrt{n}\theta_i^* \\
&= \sqrt{n}\Big(1 - \frac{1}{n}\langle v_i, \tilde{x}_i\rangle\Big)(\widehat{\theta}_i - \theta_i^*) + \frac{v_i^{\mathsf{T}}}{\sqrt{n}}\Big[w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}_{\sim i})\Big]\,. \qquad (57)
\end{aligned}$$

We decompose the above expression into the following terms:

$$\begin{aligned}
Z_i &\equiv \frac{v_i^{\mathsf{T}} w}{\sqrt{n}}\,, \\
R_i^{(1)} &\equiv \sqrt{n}\Big(1 - \frac{\langle v_i, \tilde{x}_i\rangle}{n}\Big)(\widehat{\theta}_i - \theta_i^*)\,, \\
R_i^{(2)} &\equiv \frac{v_i^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\,, \\
R_i^{(3)} &\equiv \frac{v_i^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\widehat{\theta}^{\rm p}_{\sim i} - \widehat{\theta}_{\sim i})\,.
\end{aligned} \tag{58}$$

The bulk of the proof consists in treating each of the terms above separately. Term $Z_i$ gives the Gaussian component $Z$ in equation (22). For bounding $R_i^{(2)}$, note that $\widehat{\theta}^{\rm p}_{\sim i}$ is a deterministic function of $(\tilde{y}, X_{\sim i})$ (and thus a deterministic function of $(w, X_{\sim i})$) by Equation (56). Further, $v_i$ is independent of $X_{\sim i}$, as per Lemma 3.6, and independent of noise $w$. Hence, $v_i$ is independent of $X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})$.

Bounding $R_i^{(3)}$ relies on a perturbation analysis showing that the solutions of Lasso $\widehat{\theta}$ and its perturbed form $\widehat{\theta}^{\rm p}$, are close to each other. Here is where Condition *(iii)* in the theorem statement comes into picture. The perturbation bound $\|\widehat{\theta}^{\rm p}_{\sim i} - \widehat{\theta}_{\sim i}\|_2$ depends on the correlation of $x_i$ with other columns $x_j$, with $j \in T$, where $T = \mathrm{supp}(\widehat{\theta}^{\mathrm{p}}_{\sim i} - \theta^*)$. For Gaussian designs, we have

$$\tilde{x}_i = X_T(\Sigma_{T,T})^{-1}\Sigma_{T,i} + \Sigma_{i|T}^{1/2} z\,,$$

with $z \sim \mathsf{N}(0, \mathrm{I}_n)$ independent of $X_T$ and the Schur complement $\Sigma_{i|T} \equiv \Sigma_{i,i} - \Sigma_{i,T}(\Sigma_{T,T})^{-1}\Sigma_{T,i}$. It can be shown that $\|\Sigma_{i,T}(\Sigma_{T,T})^{-1}\|_1 \le \rho$.

6.2. *Technical steps.* Let $Z = (Z_i)_{1\le i\le p}$. We rewrite $Z$ as

$$Z = \frac{1}{\sqrt{n}}\Omega X^{\mathsf{T}} w\,.$$

Since $w \sim \mathsf{N}(0, \sigma^2 \mathrm{I})$ is independent of $X$, we get

$$Z|X \sim \mathsf{N}(0, \sigma^2 \Omega\widehat{\Sigma}\Omega)\,.$$

Let $R^{(1)} = (R_i^{(1)})_{i=1}^p, R^{(2)} = (R_i^{(2)})_{i=1}^p, R^{(3)} = (R_i^{(3)})_{i=1}^p \in \mathbb{R}^p$. In the following, we provide a detailed analysis to control the terms $R^{(1)}, R^{(2)}, R^{(3)}$.

- *Bounding term $R^{(1)}$:* Recalling the definition $v_i = X\Omega e_i$, we write

$$R_i^{(1)} = \sqrt{n}\Big(1 - \frac{1}{n} e_i^{\mathsf{T}}\Omega X^{\mathsf{T}} X e_i\Big)(\widehat{\theta}_i - \theta_i^*)\,.$$

Therefore,

$$\|R^{(1)}\|_\infty \le \sqrt{n}|\mathrm{I} - \Omega\widehat{\Sigma}|_\infty \|\widehat{\theta} - \theta^*\|_2\,.$$

For $A > 0$, let $\mathcal{G}_n = \mathcal{G}_n(A)$ be the event that

$$\mathcal{G}_n(A) \equiv \Big\{X \in \mathbb{R}^{n\times p} :\, |\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \le A\sqrt{\frac{\log p}{n}}\Big\}\,. \tag{59}$$

Using the result of [42, Lemma 23] for $n \ge (A^2 C_{\min})/(4e^2 C_{\max})\log p$ we have

$$\mathbb{P}(X \in \mathcal{G}_n(A)) \ge 1 - 2p^{-c}\,, \qquad c = \frac{A^2 C_{\min}}{24 e^2 C_{\max}} - 2\,.$$

By choosing $A \equiv 10e\sqrt{C_{\max}/C_{\min}}$ we get $c \ge 2$. Therefore, provided that $n \ge 25\log p$,

$$\mathbb{P}(X \in \mathcal{G}_n(A)) \ge 1 - 2p^{-2}\,. \tag{60}$$

In addition, on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$ we have [11]

$$\|\widehat{\theta} - \theta^*\|_2 \le \frac{\sqrt{20}}{(1-\delta)^2 C_{\min}} \lambda \sqrt{s_0}\,.$$

Combining the above bounds, we obtain that on event $\mathcal{G}_n(A) \cap \mathcal{B}$,

$$\|R^{(1)}\|_\infty \le \frac{5\kappa A \sigma}{(1-\delta)^2 C_{\min}} \sqrt{\frac{s_0}{n}} \log p\,. \tag{61}$$

- *Bounding term $R^{(2)}$:* To lighten the notation, we define

$$\zeta_i \equiv \frac{1}{\sqrt{n}} X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\,. \tag{62}$$

As discussed $\widehat{\theta}^{\rm p}_{\sim i}$ is a Lasso estimator with design matrix $X_{\sim i}$ and response vector $\tilde{y} = y - \tilde{x}_i \theta^*_i$, as per equation (56). We recall the following results on the prediction error of the Lasso estimator, which bounds $\|\zeta_i\|_2$.

PROPOSITION 6.1 ([11], Theorem 6.1). *Let $S \equiv \mathrm{supp}(\theta^*_{\sim i})$. Then on the event $\tilde{\mathcal{B}}(n,p)$, we have for $\lambda \ge 8\sigma\sqrt{(\log p)/n}$,*

$$\|\zeta_i\|_2^2 \le \frac{4\lambda^2 |S|}{\phi^2(S, \widehat{\Sigma}_{\sim i, \sim i})}\,.$$

From the definition of the compatibility constant (cf. Definition 3.1), it is clear that $\phi^2(S, \widehat{\Sigma}_{\sim i,\sim i}) \ge \phi^2(S, \widehat{\Sigma})$. Therefore, combining Proposition 6.1 and Remark 3.3, we arrive at the following corollary:

COROLLARY 6.2. *On the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$, we have for $\lambda \ge 8\sigma\sqrt{(\log p)/n}$,*

$$\|\zeta_i\|_2^2 \le \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}}\,.$$

Employing Corollary 6.2, we derive a tail bound on $R^{(2)}_i$.

For $i \in [p]$ define the event

$$\mathcal{E}_i \equiv \left\{ \|\zeta_i\|_2^2 \le \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}} \right\}. \tag{63}$$

By Corollary 6.2, we have $\mathcal{B} \subseteq \mathcal{E}_i$ for $i \in [p]$. Hence, for any value $t > 0$

$$
\begin{aligned}
\mathbb{P}\Big(\|R^{(2)}\|_\infty \ge t; \mathcal{B}\Big) &\le \mathbb{P}\Big(\max_{i\in[p]} |v_i^{\mathsf{T}}\zeta_i| \ge t; \mathcal{E}_i\Big) \\
&\le p \max_{i\in[p]} \mathbb{E}\Big\{\mathbb{I}(|v_i^{\mathsf{T}}\zeta_i| \ge t)\cdot \mathbb{I}(\mathcal{E}_i)\Big\} \\
&\le 2p \max_{i\in[p]} \mathbb{E}\Big(\exp\Big[-\frac{t^2}{2\Omega_{ii}\|\zeta_i\|_2^2}\Big]\cdot \mathbb{I}(\mathcal{E}_i)\Big) \\
&\le 2p \exp\Big(-\frac{c_* t^2}{s_0\lambda^2\Omega_{ii}}\Big),
\end{aligned}
$$

with $c_* \equiv (1-\delta)^2 C_{\min}/8$. In the third inequality, we applied Fubini's theorem, and first integrate w.r.t $v_i$ and then w.r.t $\zeta_i$ using the fact that $v_i$ and $\zeta_i$ are independent. Note that $v_i \sim \mathsf{N}(0, \Omega_{ii}\mathrm{I}_{n\times n})$ and thus $v_i^{\mathsf{T}}\zeta_i|\zeta_i \sim \mathsf{N}(0, \Omega_{ii}\|\zeta_i\|_2^2)$. Further, on the event $\mathcal{E}_i$, $\|\zeta_i\|_2^2$ can be bounded as in Equation (63).

Setting $t \equiv \kappa\sigma\sqrt{3s_0/(c_* C_{\min} n)}\log p$, we get

$$
\mathbb{P}\Big(\|R^{(2)}\|_\infty \ge \kappa\sigma\sqrt{\frac{3s_0}{c_* C_{\min} n}}\log p; \mathcal{B}\Big) \le 2p^{-2}\,. \tag{64}
$$

• *Bounding term* $R^{(3)}$*:* In order to bound the last term, we first need to establish the following main lemma that bounds the distance between Lasso estimator and the solution of the perturbed problem. We refer to the Supplementary Material [43], for the proof of Lemma 6.3.

LEMMA 6.3 (Perturbation bound). *Suppose that* $\Sigma_{ii} \le 1$*, for* $i \in [p]$*. Set* $\lambda = 8\sigma\sqrt{(\log p)/n}$ *and let* $\mathcal{B}(C_\delta) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, C_\delta s_0, 3)$*. The following holds true.*

$$
\mathbb{P}\Big(\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i}\|_2 \ge C'\lambda; \mathcal{B}(C_\delta)\Big) \le 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big), \tag{65}
$$

*where,*

$$
C' \equiv \frac{24\rho(1+\delta)\sqrt{C_{\max}}}{(1-\delta)^2 C_{\min}}, \qquad c_* \equiv \frac{1}{8}(1-\delta)^2 C_{\min}\,,
$$
$$
C_\delta \equiv \frac{16 C_{\max}}{(1-\delta)^2 C_{\min}} + 1\,.
$$

We are now ready to bound term $R^{(3)}$.

$$
\begin{aligned}
|R_i^{(3)}| &\le \frac{1}{\sqrt{n}}\|v_i^{\mathsf T} X_{\sim i}\|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_1 \\
&\le \sqrt{\frac{C_\delta s_0}{n}}\|v_i^{\mathsf T} X_{\sim i}\|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_2 \\
&\le \sqrt{C_\delta s_0 n}\,|\Omega\widehat{\Sigma} - \mathrm{I}|_\infty \|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \widehat{\theta}_{\sim i}\|_2\,,
\end{aligned}
$$

where in the first inequality we used Lemma 3.5, which implies that $\|\widehat{\theta}^{\mathrm{p}}_{\sim i} - \theta^*_{\sim i}\|_0 \le C_\delta s_0$, under $\mathcal{B}$. Therefore, by Lemma 6.3 and definition (59) and since $\mathcal{B}(C_\delta) \subseteq \mathcal{B}$, we have

$$
\mathbb{P}\Big(|R_i^{(3)}| \ge C''\sigma\sqrt{\frac{s_0}{n}}\log p; \mathcal{G}_n(A)\cap\mathcal{B}(C_\delta)\Big) \le 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big)\,,
$$

with $C'' \equiv \kappa\sqrt{(C_*+1)}AC'$. Hence, by union bound over the $p$ coordinates, we get

$$
\begin{aligned}
&\mathbb{P}\Big(\|R^{(3)}\|_\infty \ge C''\sigma\sqrt{\frac{s_0}{n}}\log p; \mathcal{G}_n(A)\cap\mathcal{B}(C_\delta)\Big) \\
&\quad\le 2p\exp\Big(-\frac{c_* n}{s_0}\Big) + p\exp\Big(-\frac{n}{1000}\Big)\,.
\end{aligned}
\tag{66}
$$

We are now in position to prove the claim of Theorem 3.8.

Using equations (57) and (58), we have $\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^*) = Z + R$, where $Z|X \sim \mathsf{N}(0, \sigma^2\Omega\widehat{\Sigma}\Omega)$ and $R = R^{(1)} + R^{(2)} + R^{(3)}$. Combining equations (61), (64) and (66), we get

$$
\begin{aligned}
&\mathbb{P}\Big(\|R\|_\infty \ge C\sqrt{\frac{s_0}{n}}\log p; \mathcal{G}_n(A)\cap\mathcal{B}(C_\delta)\Big) \\
&\quad\le 2p\exp\Big(-\frac{c_* n}{s_0}\Big) + p\exp\Big(-\frac{n}{1000}\Big) + 2p^{-2}\,,
\end{aligned}
\tag{67}
$$

where $C$ is given by

$$
C \equiv \kappa\sigma\Big(\frac{5A}{(1-\delta)^2 C_{\min}} + \sqrt{\frac{3}{c_* C_{\min}}} + \sqrt{C_\delta}AC'\Big)\,. \tag{68}
$$

Further, for $n \ge \max(25\log p, c_1 C_\delta s_0\log(p/s_0))$, we have

$$
\begin{aligned}
\mathbb{P}\Big((\mathcal{G}_n(A)\cap\mathcal{B}(C_\delta))^c\Big) &\le \mathbb{P}(\mathcal{G}_n(A)^c) + \mathbb{P}(\tilde{\mathcal{B}}(n,p)^c) + \mathbb{P}(\mathcal{B}_\delta(n, C_\delta s_0, 3)^c) \\
&\le 2p^{-2} + 2p^{-2} + 2e^{-\delta^2 n} = 4p^{-2} + 2e^{-\delta^2 n}\,,
\end{aligned}
\tag{69}
$$

where we used bound (60), Lemma 3.2 and Lemma 3.4.

The result follows from equations (67) and (69), and setting $\delta = 1 - 1/\sqrt{2}$.

## SUPPLEMENTARY MATERIAL

**Supplement to: "Debiasing the Lasso: Optimal Sample Size for Gaussian Designs"**
(doi: 10.1214/00-AOSXXXXSUPP; .pdf). Due to space constraints, proof of theorems and some of the technical details as well as additional numerical studies are provided in the Supplementary Material [43].

## REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19** 716–723.

[2] Anderson, G. W., Guionnet, A. and Zeitouni, O. (2009). *An introduction to random matrices*. Cambridge University Press.

[3] Barber, R. F., Candès, E. J. et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085.

[4] Bayati, M., Erdogdu, M. A. and Montanari, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[5] Bayati, M., Lelarge, M. and Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability* **25** 753–822.

[6] Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inform. Theory* **57** 764-785.

[7] Bayati, M. and Montanari, A. (2012). The LASSO risk for gaussian matrices. *IEEE Trans. on Inform. Theory* **58** 1997–2017.

[8] Belloni, A. and Chernozhukov, V. (2013). Least Squares after Model Selection in High-Dimensional Sparse Models. *Bernoulli* **19** 521–547.

[9] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Amer. J. of Mathematics* **37** 1705-1732.

[10] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242.

[11] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer-Verlag.

[12] Bühlmann, P., van de Geer, S. et al. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9** 1449–1473.

[13] Cai, T. T. and Guo, Z. (2016). Accuracy Assessment for High-dimensional Linear Regression. arXiv:1603.03474.

[14] Cai, T. T., Guo, Z. et al. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* **45** 615–646.

[15] Candes, E., Romberg, J. K. and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics* **59** 1207-1223.

[16] Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313-2351.

[17] Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge.

[18] Chen, M., Ren, Z., Zhao, H. and Zhou, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *Journal of the American Statistical Association* **111** 394–406.

[19] Chen, S. S. and Donoho, D. L. (1995). Examples of basis pursuit. In *Proceedings of Wavelet Applications in Signal and Image Processing III.*

[20] Chernozhukov, V., Hansen, C. and Spindler, M. (2015). Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics* **7** 649–688.

[21] Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N. et al. (2015). High-Dimensional Inference: Confidence Intervals, $p$-Values and R-Software hdi. *Statistical Science* **30** 533–558.

[22] Dicker, L. H. (2012). Residual variance and the signal-to-noise ratio in high-dimensional linear models. arXiv:1209.0012.

[23] Donoho, D., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. on Inform. Theory* **52** 6–18.

[24] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. Amer. Statist. Assn.* **90** 1200-1224.

[25] Donoho, D. L., Maleki, A. and Montanari, A. (2011). The Noise Sensitivity Phase Transition in Compressed Sensing. *IEEE Trans. on Inform. Theory* **57** 6920-6941.

[26] Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. on Inform. Theory* **47** 2845–2862.

[27] Donoho, D. L., Johnstone, I. and Montanari, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *Information Theory, IEEE Transactions on* **59** 3396–3433.

[28] Donoho, D. L. and Johnstone, I. M. (1994). Minimax Risk over $l_p$ balls. *Prob. Th. and Rel. Fields* **99** 277-303.

[29] Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message Passing Algorithms for Compressed Sensing. *Proceedings of the National Academy of Sciences* **106** 18914-18919.

[30] Efron, B. (2012). The estimation of prediction error. *Journal of the American Statistical Association.*

[31] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

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

[33] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research* **10** 2013–2038.

[34] Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. arXiv:1410.2597.

[35] Janková, J. and van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test* **26** 143–162.

[36] Jankova, J., van de Geer, S. et al. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics* **9** 1205–1229.

[37] Janson, L., Barber, R. F. and Candes, E. (2016). EigenPrism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[38] Janson, L., Su, W. et al. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics* **10** 960–975.

[39] Javanmard, A. and Lee, J. D. (2017). A Flexible Framework for Hypothesis Testing in High-dimensions. *arXiv preprint arXiv:1704.07971*.

[40] Javanmard, A. and Montanari, A. (2013). Nearly Optimal Sample Size in Hypothesis Testing for High-Dimensional Regression. In *51st Annual Allerton Conference* 1427–1434.

[41] Javanmard, A. and Montanari, A. (2014). Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory. *IEEE Trans. on Inform. Theory* **60** 6522-6554.

[42] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15** 2869–2909.

[43] Javanmard, A. and Montanari, A. (2017). Supplementary material to "Debiasing the Lasso: Optimal Sample Size for Gaussian Designs". DOI:10.1214/00-AOSXXXXSUPP.

[44] Ledoux, M. (2001). The concentration of measure phenomenon. In *Mathematical Surveys and Monographs*, **89** American Mathematical Society, Providence, RI.

[45] Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* **42** 413.

[46] Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics* **15** 661–675.

[47] Meinshausen, N. (2014). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[48] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34** 1436-1462.

[49] Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417-473.

[50] Obuchi, T. and Kabashima, Y. (2016). Cross validation in LASSO and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment* **2016** 53304–53339.

[51] Reid, S., Tibshirani, R. and Friedman, J. (2013). A Study of Error Variance Estimation in Lasso Regression. *arXiv Preprint* arXiv:1311.5274.

[52] Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. on Inform. Theory* **59** 3434–3447.

[53] Städler, N., Bühlmann, P. and van de Geer, S. (2010). $\ell_1$-penalization for Mixture Regression Models (with discussion). *Test* **19** 209–256.

[54] Su, W., Candes, E. et al. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics* **44** 1038–1068.

[55] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879-898.

[56] Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Exact Post-selection Inference for Forward Stepwise and Least Angle Regression. arXiv:1401.3889.

[57] Tibshirani, R. (1996). Regression shrinkage and selection with the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58** 267-288.

[58] Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202.

[59] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications* (Y. C. Eldar and G. Kutyniok, eds.) 210-268. Cambridge University Press.

[60] Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics* **37** 2178.

[61] Ye, F. and Zhang, C.-H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. *The Journal of Machine Learning Research* **11** 3519–3540.

[62] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

[63] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics* 1567–1594.

[64] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.

[65] Zou, H., Hastie, T., Tibshirani, R. et al. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics* **35** 2173–2192.

# SUPPLEMENTARY MATERIAL TO "DEBIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS"

By Adel Javanmard[§] and Andrea Montanari[¶]

*University of Southern California[§] and Stanford University[¶]*

## APPENDIX A: PROOF OF LEMMA 3.5

This proposition is an improved version of Theorem 7.2 in [9].

We first recall the definition of *restricted eigenvalues* as given by:

$$\phi_{\max}(k) \equiv \max_{1\le \|v\|_0\le k} \frac{\langle v, \widehat{\Sigma} v\rangle}{\|v\|_2^2}\,.$$

Clearly, $\phi_{\max}(k)$ is an increasing function of $k$.

Employing [59, Remark 5.4], for any $1 \le k \le n$ and a fixed subset $J \subset [p]$ with $|J| = k$, we have

$$\mathbb{P}\Big(\sigma_{\max}(\widehat{\Sigma}_{J,J}) \ge C_{\max} + C\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\Big) \le 2e^{-ct^2}\,,$$

for $t \ge 0$, where $C$ and $c$ depend only on $C_{\max}$. Therefore, by union bound over all possible subsets $J \subseteq [p]$ we obtain

$$\mathbb{P}\Big(\phi_{\max}(k) \ge C_{\max} + C\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\Big) \le 2\binom{p}{k} e^{-ct^2} \le 2e^{-ct^2 + k\log p + k}\,, \tag{70}$$

for $t \ge 0$.

Let $\widehat{S} \equiv \text{supp}(\widehat{\theta})$. Recall that the stationarity condition for the Lasso cost function reads $X^{\mathsf{T}}(y - X\widehat{\theta}) = n\lambda\, v(\widehat{\theta})$, where $v(\widehat{\theta}) \in \partial\|\widehat{\theta}\|_1$. Equivalently,

$$\frac{1}{n}X^{\mathsf{T}}X(\theta^* - \widehat{\theta}) = \lambda\, v(\widehat{\theta}) - \frac{1}{n}X^{\mathsf{T}}w\,.$$

On the event $\tilde{\mathcal{B}}(n,p)$, we have $\|X^{\mathsf{T}}w\|_\infty \le n\lambda/4$. Thus for all $i \in \widehat{S}$

$$\Big|\frac{1}{n}[X^{\mathsf{T}}X(\theta^* - \widehat{\theta})]_i\Big| \ge \frac{\lambda}{2}\,.$$

Squaring and summing the last identity over $i \in \widehat{S}$, we obtain that, for $h \equiv n^{-1/2} X(\theta^* - \widehat{\theta})$,

$$\frac{\lambda^2}{4}|\widehat{S}| \le \frac{1}{n}\sum_{i\in\widehat{S}} (e_i^{\mathsf{T}} X^{\mathsf{T}} h)^2 = \langle h, \frac{1}{n} X_{\widehat{S}} X_{\widehat{S}}^{\mathsf{T}} h\rangle$$

$$\le \|\widehat{\Sigma}_{\widehat{S},\widehat{S}}\|_2^2 \|h\|^2 \le \phi_{\max}(|\widehat{S}|)\|h\|_2^2 \,. \tag{71}$$

By a similar argument as in Corollary 6.2, on the event $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}(n, s_0, 3)$ we have

$$\|h\|_2^2 \le \frac{4\lambda^2 s_0}{(1-\delta)^2 C_{\min}} \,.$$

Thus,

$$|\widehat{S}| \le \frac{16\phi_{\max}(\widehat{S})}{(1-\delta)^2 C_{\min}} s_0 \,. \tag{72}$$

Note that $|\widehat{S}| \le n$ by the fact that the columns of $X$ are in generic positions. Using monotonicity property of $\phi_{\max}(\cdot)$, we have $\phi_{\max}(|\widehat{S}|) \le \phi_{\max}(n)$. Invoking equation (70) with $k = n$, we have $\phi_{\max}(n) < c_1\sqrt{\log p}$ with high probability for some constant $c_1$.

Hence, by equation (72)

$$|\widehat{S}| < \widetilde{C} s_0 \sqrt{\log p}\,, \qquad \tilde{C} \equiv \frac{16 c_1}{(1-\delta)^2 C_{\min}} \,. \tag{73}$$

Now, we use this bound on $|\widehat{S}|$ along with equation (72) to get a better bound on $|\widehat{S}|$. Again by using the fact that $\phi_{\max}(k)$ is a non-decreasing function of $k$, we have

$$\phi_{\max}(|\widehat{S}|) < \phi_{\max}(\widetilde{C} s_0 \sqrt{\log p}) \le C_{\max}\,, \tag{74}$$

with high probability where we used the assumption $n \gg s_0(\log p)^2$. Using this bound in equation (72), we get

$$|\widehat{S}| < \frac{16 C_{\max}}{(1-\delta)^2 C_{\min}} s_0 \,.$$

The result follows.

## APPENDIX B: PROOF OF LEMMA 3.7

By definition of $\ell_\infty$ operator norm, for a symmetric invertible matrix $A$ we have

$$\|A^{-1}\|_\infty \equiv \max_{v\neq 0}\frac{\|A^{-1}v\|_\infty}{\|v\|_\infty} = \max_{u\neq 0}\frac{\|u\|_\infty}{\|Au\|_\infty} = \frac{1}{\min_{u\neq 0}\frac{\|Au\|_\infty}{\|u\|_\infty}}\,. \tag{75}$$

Note that for any set $T\subseteq[p]$ we have

$$\min_{u\neq 0}\frac{\|Au\|_\infty}{\|u\|_\infty} \le \min_{\tilde u\neq 0}\frac{\|A_{T,T}\tilde u\|_\infty}{\|\tilde u\|_\infty}\,,$$

whence we obtain

$$\|A^{-1}\|_\infty \ge \frac{1}{\min_{u\neq 0}\frac{\|A_{T,T}\tilde u\|_\infty}{\|\tilde u\|_\infty}} = \|A^{-1}_{T,T}\|_\infty\,. \tag{76}$$

Since the above inequality holds for any $T\subseteq[p]$, we obtain that $\rho(A,p) = \|A^{-1}\|_\infty$.

Therefore, using the non-decreasing property of $\rho(A,k)$, for any $1\le k\le p$ we have

$$\rho(A,k)\le\rho(A,p) = \|A^{-1}\|_\infty\,. \tag{77}$$

To obtain the other upper bound on $\rho(A,k)$, we note that

$$\begin{aligned}\rho(A,k) &\le \max_{T\subseteq[p],|T|\le k}\ \max_{j\in[p]}\ \sqrt{k}\,\|A^{-1}_{T,T}e_j\|_2\\ &\le \max_{T\subseteq[p],|T|\le k}\ \sqrt{k}\,\sigma_{\max}(A^{-1}_{T,T}) \le \frac{\sqrt{k}}{\sigma_{\min}(A)}.\end{aligned} \tag{78}$$

The result follows by combining Equations (77) and (78).

## APPENDIX C: PROOF OF THEOREM 3.13 (UNKNOWN COVARIANCE)

We decompose $\sqrt{n}(\widehat\theta^{\rm d}-\theta^*)$ into three terms:

$$\begin{aligned}\sqrt{n}(\widehat\theta^{\rm d}-\theta^*) &= \sqrt{n}(\widehat\theta-\theta^*) + \frac{1}{\sqrt{n}}MX^{\mathsf T}(y-X\widehat\theta)\\ &= \sqrt{n}({\rm I}-M\widehat\Sigma)(\widehat\theta-\theta^*) + \frac{1}{\sqrt{n}}MX^{\mathsf T}w\\ &= \underbrace{\sqrt{n}({\rm I}-\Omega\widehat\Sigma)(\widehat\theta-\theta^*)}_{I_1} + \underbrace{\sqrt{n}(\Omega-M)\widehat\Sigma(\widehat\theta-\theta^*)}_{I_2} + \underbrace{\frac{1}{\sqrt{n}}MX^{\mathsf T}w}_{I_3}\,.\end{aligned}$$

Note that the term $I_1$ is exactly the bias vector $R$ of the debiased estimator in case of known covariance (with $M = \Omega$). Therefore, by invoking the result of Theorem 3.8, we have

$$\mathbb{P}\Big(\|I_1\|_\infty \ge C\sqrt{\frac{s_0}{n}}\log p\Big) \le 2pe^{-c_* n/s_0} + pe^{-n/1000} + 8p^{-1} + 2e^{-\delta^2 n}\,. \tag{79}$$

We next provide two bounds on $\|I_2\|_\infty$.

In our first bound, we use duality of $\ell_\infty$ norm (on $\widehat\Sigma(\widehat\theta - \theta^*)$) and $\ell_1$ norm on rows of $\Omega - M$ as follows:

$$\|I_2\|_\infty \le \sqrt{n}\|\Omega - M\|_\infty \|\widehat\Sigma(\widehat\theta - \theta^*)\|_\infty\,. \tag{80}$$

By the KKT condition for $\widehat\theta$, there exists a vector $\xi$ in the subgradient of the $\ell_1$ norm at $\widehat\theta$, such that $\widehat\Sigma(\widehat\theta - \theta^*) = X^\mathsf{T} w/n - \lambda\xi$. Therefore,

$$\|\widehat\Sigma(\widehat\theta - \theta^*)\|_\infty \le \frac{1}{n}\|X^\mathsf{T} w\|_\infty + \lambda\|\xi\|_\infty\,. \tag{81}$$

We have $\|\xi\|_\infty \le 1$ and on event $\tilde{\mathcal{B}}(n,p)$,

$$\frac{1}{n}\|X^\mathsf{T} w\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n}} \le \frac{\lambda}{4}\,. \tag{82}$$

Using these bounds in Equation (81), we obtain $\|\widehat\Sigma(\widehat\theta - \theta^*)\|_\infty \le 5\lambda/4$. As proved in [58, Theorem 2.4], we have $\|M - \Omega\|_\infty \lesssim s_\Omega\sqrt{(\log p)/n}$. Combining these bounds in Equation (80) gives our first bound on $I_2$.

$$\|I_2\|_\infty \lesssim \sqrt{n}s_\Omega\sqrt{\frac{\log p}{n}}\sqrt{\frac{\log p}{n}} \lesssim \frac{s_\Omega \log p}{\sqrt{n}}\,. \tag{83}$$

To obtain a second bound on $I_2$, we proceed by writing $I_2$ as

$$I_2 = \sqrt{n}\Big[(\Omega\widehat\Sigma - \mathrm{I}) - (M\widehat\Sigma - \mathrm{I})\Big](\widehat\theta - \theta^*)\,. \tag{84}$$

Therefore,

$$\|I_2\|_\infty \le \sqrt{n}\Big(|\Omega\widehat\Sigma - \mathrm{I}|_\infty + |M\widehat\Sigma - \mathrm{I}|_\infty\Big)\|\widehat\theta - \theta^*\|_1\,. \tag{85}$$

On event $\mathcal{G}_n(A)$ (see Equation (59)), we have $|\Omega\widehat\Sigma - \mathrm{I}|_\infty \lesssim \sqrt{(\log p)/n}$.

We next bound the term $|M\widehat\Sigma - \mathrm{I}|_\infty$. The way that $M$ is constructed as per (32) and definition of $\hat\tau_i^2$ given by (31), it is easy to verify that

$$\begin{aligned} &\frac{1}{n}\tilde{x}_j^\mathsf{T} X M_{j,\cdot} = 1\,,\\ &\frac{1}{n}X_{\sim j}^\mathsf{T} X M_{j,\cdot} = \frac{1}{n\hat\tau_j^2}X_{\sim j}^\mathsf{T}(\tilde{x}_j - X_{\sim j}\hat\gamma_j)\,. \end{aligned} \tag{86}$$

Further, by the KKT condition for optimization (29), we have

$$\frac{1}{n}X_{\sim j}^{\mathsf{T}}(\tilde{x}_j - X_{\sim j}\hat{\gamma}_j) = -\widetilde{\lambda}\partial\|\hat{\gamma}_j\|_1\,. \tag{87}$$

Hence, by combining Equations (86) and (87), we obtain

$$\|\widehat{\Sigma}M_{j,\cdot}^{\mathsf{T}} - e_j\|_\infty \le \frac{\widetilde{\lambda}}{\hat{\tau}_j^2}\,. \tag{88}$$

We proceed by using the following lemma to bound $1/\hat{\tau}_j^2$, uniformly in $j$. This lemma is an improved version of Lemma 5.3 in [58] and can be proved using a similar analysis with a more careful treatment of constants, instead of just working with growth rates and hiding constants in the $O, o$ notations.

LEMMA C.1. *Suppose that the rows of design matrix $X$ are i.i.d. realizations from a Gaussian distribution whose $p$-dimensional inner product matrix $\Sigma$ has strictly positive smallest eigenvalue, that is $\sigma_{\min}(\Sigma) > C_{\min} > 0$. Further assume that $\Sigma_{ii} \le 1$ for $i \in [p]$. Let $s_\Omega$ denote the maximum row sparsity of $\Omega = \Sigma^{-1}$. We also let $M \in \mathbb{R}^{p\times p}$ be constructed as (32) with $\widetilde{\lambda} = K\sqrt{\log p/n}$. If $s_\Omega < \min(1, C_{\min}^3)/(512\widetilde{\lambda}^2)$ then*

$$\max_{j\in[p]} \frac{1}{\hat{\tau}_j^2} \le \frac{2}{C_{\min}}\,. \tag{89}$$

Therefore, using Equation (88) and Lemma C.1 we get

$$|M\widehat{\Sigma} - \mathrm{I}|_\infty \le \frac{2K}{C_{\min}}\sqrt{\frac{\log p}{n}}\,, \tag{90}$$

provided that $n > cs_\Omega \log p$, with $c = 512K^2/\min(1, C_{\min}^3)$.

In addition, on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$ we have $\|\widehat{\theta} - \theta^*\|_1 \lesssim s_0\lambda \approx s_0\sqrt{(\log p)/n}$. (See e.g., [11].)

Combining these bounds, we arrive at

$$\|I_2\|_\infty \lesssim \frac{s_0 \log p}{\sqrt{n}}\,. \tag{91}$$

We summarize bounds given by (83) and (91) as

$$\|I_2\|_\infty \lesssim \min(s_0, s_\Omega)\frac{\log p}{\sqrt{n}}\,. \tag{92}$$

Finally, note that

$$I_3|X \sim \mathsf{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})\,.$$

The result follows by letting $Z \equiv I_3$ and $R \equiv I_1 + I_2$.

## APPENDIX D: SAMPLE SPLITTING TECHNIQUES

In this appendix, we discuss how sample splitting can be used to modify the debiased estimator as to go around the sparsity barrier at $s_0 = o(\sqrt{n}/\log p)$. This provides an alternative to the more careful analysis carried out in the main body of the paper, that we discuss for the sake of simplicity. As mentioned in the introduction, sample splitting has its own drawbacks, most notably the dependence of the results on the random data split, and the sub-optimal use of all the samples.

For the sake of notational simplicity we assume here that the number of samples is $2n$ and is randomly split in two batches of size $n$: $(x_1, y_1)$, ..., $(x_n, y_n)$, and $(\overline{x}_1, \overline{y}_1)$,..., $(\overline{x}_n, \overline{y}_n)$. Note that the change of notation only amounts to a constant multiplicative factor in the sample size, which is of no concern to us. In vector notation, these batches are denoted as $(y, X)$ and $(\overline{y}, \overline{X})$. We then proceed as follows:

1. We use the second batch to compute the Lasso estimator, namely

$$\widehat{\theta}(\overline{y}, \overline{X}; \lambda) \equiv \arg\max_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\overline{y} - \overline{X}\theta\|_2^2 + \lambda \|\theta\|_1 \right\} . \tag{93}$$

2. We use the first batch to compute the debiasing matrix $M$, e.g. using the node-wise Lasso as in Section 3.3.
3. We use the first batch to implement the debiasing, namely

$$\widehat{\theta}^{\text{split}} = \widehat{\theta}(\overline{y}, \overline{X}) + \frac{1}{n} M X^{\mathsf{T}} \big( y - X\widehat{\theta}(\overline{y}, \overline{X}) \big) . \tag{94}$$

The main remark is that, thanks to the splitting, $X$ is statistically independent from $\widehat{\theta}$, which greatly simplifies the analysis. Notice that we did not use the responses in $y$.

For the sake of simplicity, we shall analyze this procedure in the case in which the precision matrix $\Omega$ is known, and we hence set $M = \Omega$. The generalization to $M$ constructed via the node-wise Lasso is straightforward as in the proof of Theorem 3.13.

The next statement implies that, for sparsity level $s_0 = o(n/(\log p)^2)$,the sample splitting debiased estimator is asymptotically Gaussian.

PROPOSITION D.1. *Consider the linear model (2) where $X$ has independent Gaussian rows, with zero mean and covariance $\Sigma$. Suppose that $\Sigma$ satisfies the technical conditions of Theorem 3.8*

*Let $\widehat{\theta}$ be the Lasso estimator defined by (3) with $\lambda = 8\sigma\sqrt{(\log p)/n}$. Further, let $\widehat{\theta}^{\text{split}}$ be the modified (sample-splitting) debiased estimator defined* *in Eq. (94) with $M = \Omega \equiv \Sigma^{-1}$. Then, there exist constants $c, C$ depending solely on $C_{\min}, C_{\max}, \delta$ and $\rho$, such that, for $n \geq c \max(\log p, s_0 \log(p/s_0))$ the following holds true:*

$$\sqrt{n}(\widehat{\theta}^{\rm d} - \theta^*) = Z + R\,, \qquad Z|X \sim \mathsf{N}(0, \sigma^2 \Omega\widehat{\Sigma}\Omega)\,, \tag{95}$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\|R\|_\infty \geq C\sqrt{\frac{s_0}{n}}\log p\Big) = 0\,. \tag{96}$$

PROOF. Proceeding as in the proof of Theorem 3.8, it is sufficient to bound the bias term of $\sqrt{n}(\widehat{\theta}^{\rm split} - \theta^*)$, which is given by (cf. (8))

$$R \equiv \sqrt{n}(\Omega\widehat{\Sigma} - {\rm I})(\theta^* - \widehat{\theta})\,. \tag{97}$$

To lighten the notation, let $u = \theta^* - \widehat{\theta}$. Expanding $R$ we get

$$R = \sqrt{n}(\Omega\widehat{\Sigma} - {\rm I})u = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\Omega x_i x_i^{\mathsf{T}} - {\rm I})u\,. \tag{98}$$

To control $\|R\|_\infty$, we bound each component $R_j$ individually. Let $e_j$ be the $j$-th element of the standard basis with one at the $j$-th position and zero everywhere else. We write

$$R_j = \frac{1}{\sqrt{n}}\sum_{i=1}^n (e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}} u) - u_j\,.$$

Let $Z_i \equiv (e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}} u) - u_j$. Note that conditional on $(\overline{y}, \overline{X})$, $\widehat{\theta}$ and therefore $u$ are deterministic. Furthermore, since the first batch $(y, X)$ is independent of $(\overline{y}, \overline{X})$, the rows $x_i$ are independent conditional on $(\overline{y}, \overline{X})$. Therefore, $Z_i|(\overline{y}, \overline{X})$ are independent with $\mathbb{E}(Z_i|\overline{y}, \overline{X}) = e_j^{\mathsf{T}}\Omega\Sigma u - u_j = 0$. We let $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ respectively denote the sub-exponential and sub-gaussian norms and condition on $(\overline{y}, \overline{X})$ in the sequel. As shown in [59, Remark 5.18],

$$\|Z_i\|_{\psi_1} \leq 2\|(e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}} u)\|_{\psi_1}\,.$$

In addition, for any two random variables $v$ and $w$, we have $\|vw\|_{\psi_1} \leq 2\|v\|_{\psi_2}\|w\|_{\psi_2}$. Hence,

$$\begin{aligned}
\|(e_j^{\mathsf{T}}\Omega x_i)(x_i^{\mathsf{T}} u)\|_{\psi_1} &\leq 2\|e_j^{\mathsf{T}}\Omega x_i\|_{\psi_2}\|x_i^{\mathsf{T}} u\|_{\psi_2}\\
&= 2\|e_j^{\mathsf{T}}\Omega^{1/2}\|_2\|\Omega^{1/2}x_i\|_{\psi_2}^2\|\Omega^{-1/2}u\|_2\\
&\leq 2\sqrt{C_{\max}/C_{\min}}\,\|\Omega^{1/2}x_i\|_{\psi_2}^2\|u\|_2\,.
\end{aligned}$$

Given that $\Omega^{1/2}x_i \sim \mathsf{N}(0, \mathrm{I})$, we get $\|\Omega^{1/2}x_i\|_{\psi_2} = 1$. Hence, $\max_i \|Z_i\|_{\psi_1} \le C\|u\|_2$ with $C \equiv 4\sqrt{C_{\max}/C_{\min}}$. Applying Bernstein-type inequality [59, Proposition 5.16], for every $t \ge 0$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n}\frac{1}{\sqrt{n}}Z_i\Big| \ge t\ \Big|(\overline{y}, \overline{X})\Big\} \le 2\exp\Big[-c\min\Big(\frac{t^2}{C^2\|u\|_2^2}, \frac{t\sqrt{n}}{C\|u\|_2}\Big)\Big]\,, \tag{99}$$

where $c > 0$ is an absolute constant. Observe that on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n,p)$[1], we have

$$\|u\|_2^2 = \|\theta^* - \widehat{\theta}\|_2^2 \lesssim s_0\lambda^2\,.$$

Therefore, by using tail bound (99) and applying union bound over the $p$ entries of $R$, we get (for $n \ge c\log p$ with $c$ a suitable constant)

$$\|R\|_\infty \lesssim \sqrt{\frac{s_0}{n}}\log p\,,$$

with high probability. □

## APPENDIX E: PROOF OF PROPOSITIONS 4.1 AND 4.2

**E.1. Proof of Proposition 4.1.** Fix $M$ and $\lambda$ for which Equations (42)-(43) hold true and let

$$\theta^{\mathrm{d}} = \widehat{\theta} + \frac{1}{n}MX^{\mathsf{T}}(y - X\widehat{\theta})\,.$$

We construct confidence interval $J_\alpha^{\mathrm{d}}$ centered at $\widehat{\theta}^{\mathrm{d}}$ as follows:

$$J_\alpha^{\mathrm{d}} \equiv [\widehat{\theta}_1^{\mathrm{d}} - \delta(\alpha, n), \widehat{\theta}_1^{\mathrm{d}} + \delta(\alpha, n)] \tag{100}$$

$$\delta(\alpha, n) \equiv \Phi^{-1}(1-\alpha/2)\frac{1}{(1-\varepsilon)\sqrt{n}}\min\{\widehat{\sigma}\sqrt{Q_1}, \sqrt{(1+\varepsilon)cC}\} + \frac{\Delta_n}{\sqrt{n}}\,, \tag{101}$$

where $\varepsilon \in (0, 1/2)$ is arbitrary fixed value and $\Phi(x) \equiv \int_{-\infty}^{x} e^{-t^2/}\mathrm{d}t/\sqrt{2\pi}$ is the Gaussian distribution. Further, recall that $c$ is the bound on $\sigma$ in the definition of $\Gamma(s_0, s_\Omega, \rho)$.

We have

$$\ell(J_\alpha^{\mathrm{d}}) \le 2\Phi^{-1}(1-\alpha/2)\frac{1}{(1-\varepsilon)\sqrt{n}}\sqrt{(1+\varepsilon)cC} + \frac{\Delta_n}{\sqrt{n}}\,, \tag{102}$$

and therefore, $\mathbb{E}_\gamma\{\ell(J_\alpha^{\mathrm{d}})\} \lesssim (1+\Delta_n)/\sqrt{n}$.

[1]See Section 3.2 for definition of $\mathcal{B}_\delta(n, s_0, 3)$ and $\tilde{\mathcal{B}}(n,p)$

We next show that $J^{\rm d}_\alpha \in \mathcal{I}_\alpha(\Gamma)$. Define the following events:

$$\begin{aligned}\mathcal{E}_1 &\equiv \{(1-\varepsilon)\sigma \le \widehat{\sigma} \le (1+\varepsilon)\sigma\},\\ \mathcal{E}_2 &\equiv \{\|Q\|_\infty \le C\},\\ \mathcal{E}_3 &\equiv \{\|R\|_\infty \le \Delta_n\}.\end{aligned}$$

We further let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and $Z = e_1^{\sf T} M X^{\sf T} w/\sqrt{n}$. Since $Z/\sqrt{n}|X \sim {\sf N}(0, \sigma^2 Q_1/n)$, we have

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}}|Z| \le \sqrt{\frac{Q_1}{n}}\sigma\Phi^{-1}(1-\alpha/2)\Big|X\Big) = 1-\alpha\,. \tag{103}$$

By integrating w.r.t $X$ we get the same coverage probability unconditionally. Note that on event $\mathcal{E}$, $\widehat{\sigma}\sqrt{Q_1} \le \sqrt{(1+\varepsilon)\sigma^2 C}$ and on $\Gamma(s_0, s_\Omega, \rho)$, we have $\sigma \le \sqrt{c}$. Further, $\sigma \le \widehat{\sigma}/(1-\varepsilon)$. Hence, on event $\mathcal{E}$

$$\begin{aligned}\delta(\alpha, n) &= \Phi^{-1}(1-\alpha/2)\frac{\widehat{\sigma}}{(1-\varepsilon)}\sqrt{\frac{Q_1}{n}} + \frac{\Delta_n}{\sqrt{n}}\\ &\ge \Phi^{-1}(1-\alpha/2)\sigma\sqrt{\frac{Q_1}{n}} + \frac{\Delta_n}{\sqrt{n}} \equiv \delta_0(\alpha, n)\,.\end{aligned} \tag{104}$$

We have the following bound on the coverage probability

$$\begin{aligned}\mathbb{P}(\theta^*_1 \in J^{\rm d}_\alpha) &= \mathbb{P}(|\widehat{\theta}^{\rm d}_1 - \theta^*{}_1| \le \delta(\alpha, n))\\ &\ge \mathbb{P}(\{|\widehat{\theta}^{\rm d}_1 - \theta^*_1| \le \delta_0(\alpha, n)\} \cap \mathcal{E})\\ &\overset{(a)}{\ge} \mathbb{P}\Big(\Big\{\frac{1}{\sqrt{n}}|Z| \le \sqrt{\frac{Q_1}{n}}\sigma\Phi^{-1}(1-\alpha/2)\Big\} \cap \mathcal{E}\Big)\\ &\ge \mathbb{P}\Big(\frac{1}{\sqrt{n}}|Z| \le \sqrt{\frac{Q_1}{n}}\sigma\Phi^{-1}(1-\alpha/2)\Big) - \mathbb{P}(\mathcal{E}^c)\\ &\overset{(b)}{=} 1-\alpha-\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\mathcal{E}) - \alpha\,,\end{aligned}$$

where $(a)$ follows from the decomposition $\widehat{\theta}^{\rm d}_1 = \theta^*{}_1 + Z/\sqrt{n} + R/\sqrt{n}$ and the fact that $\|R\|_\infty \le \Delta_n$ on $\mathcal{E}$; $(b)$ follows from Equation (103). Since $\mathbb{P}(\mathcal{E}) \to 0$, we obtain

$$\liminf_{n\to\infty} \inf_{\gamma\in\Gamma(s_0, s_\Omega, \rho)} \mathbb{P}_\gamma(\theta^*_1 \in J^{\rm d}_\alpha) \ge 1-\alpha\,. \tag{105}$$

Therefore, as claimed,

$$\ell^*_\alpha(\Gamma(s_0, s_\Omega, \rho)) \le \mathbb{E}_\gamma\{\ell(J^{\rm d}_\alpha)\} \lesssim (1+\Delta_n)/\sqrt{n}\,. \tag{106}$$

**E.2. Proof of Proposition 4.2.** The proof of this proposition follows the same lines as [14][Theorem 3]. Under the gaussian design model, the data pairs $(y_i, x_i)$ has a joint gaussian distribution with mean zero and covariance $\tilde{\Sigma}$, where $\tilde{\Sigma}$ admits the following block decomposition:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{yy} & \tilde{\Sigma}_{yx} \\ \tilde{\Sigma}_{xy} & \tilde{\Sigma}_{xx} \end{pmatrix} = \begin{pmatrix} \theta^{\mathsf{T}}\Sigma\theta + \sigma^2 & \theta^{\mathsf{T}}\Sigma \\ \Sigma\theta & \Sigma \end{pmatrix}, \tag{107}$$

where we posit the model $y = X\theta + w$ with $w \sim \mathsf{N}(0, \sigma^2 \mathrm{I})$. (Throughout this section, we simplify our notations by writing $\theta$ instead of $\theta^*$ for the true model parameters.) We also define $\mathsf{PSD}(p) \equiv \{M \in \mathbb{R}^{p\times p} : M \succeq 0\}$, the set of positive semidefinite matrices of size $p$.

Notice that there is a one-to-one map between the parameter space $\Gamma \equiv \{\gamma = (\theta, \Omega, \sigma^2) : \theta \in \mathbb{R}^p, \Omega \in \mathsf{PSD}(p), \sigma^2 \in \mathbb{R}_+\}$ and $\mathsf{PSD}(p+1)$. Specifically, define the function $h : \mathsf{PSD}(p+1) \mapsto \Gamma$ as $h(\tilde{\Sigma}) = ((\tilde{\Sigma}_{xx})^{-1}\tilde{\Sigma}_{xy}, (\tilde{\Sigma}_{xx})^{-1}, \tilde{\Sigma}_{yy} - (\tilde{\Sigma}_{xy})^{\mathsf{T}}(\tilde{\Sigma}_{xx})^{-1}\tilde{\Sigma}_{xy})$. The inverse map $h^{-1}$ is given by

$$h^{-1}((\theta, \Omega, \sigma^2)) = \begin{pmatrix} \theta^{\mathsf{T}}\Omega^{-1}\theta + \sigma^2 & \theta^{\mathsf{T}}\Omega^{-1} \\ \Omega^{-1}\theta & \Omega^{-1} \end{pmatrix}. \tag{108}$$

We next define a null hypothesis $H_0$ and an alternative hypothesis $H_1$ as follows. Let $s_* = \min(s_0 - 1, s_\Omega)$ and $s_1 = s_0 - s_* \ge 1$. The null space is a singleton $H_0 = \{\check{\gamma} = (\check{\theta}, \mathrm{I}, \check{\sigma}^2)\}$ with $\check{\theta}_1 = 0$, $\|\check{\theta}\|_0 = s_1 - 1$ and $\check{\sigma}^2 \in (0, c]$. We further let $S = \mathrm{supp}(\check{\theta})$ and denote by $\pi_{H_0}$ the point mass prior on $H_0$.

Next we construct the alternative parameter space $H_1$. First, we define the following set

$$\mathcal{A}(\nu, k) \equiv \left\{\delta : \delta \in \mathbb{R}^{p_1}, \|\delta\|_0 = k, \delta_i \in \{0, \nu\} \text{ for } 1 \le i \le p_1\right\}, \tag{109}$$

where $p_1 = p - s_1$. We set $k \equiv \min(s_*, (\rho - 1.01)/\nu)$ where $\rho$ comes from the constraint $\|\Omega\|_\infty \le \rho$ in the definition of $\Gamma(s_0, s_\Omega)$ . Later in the proof we enforce some constraints on the value of $\nu$ and in the hindsight, set a suitable value for $\nu$ that complies with those constraints.

For a given $\delta \in \mathbb{R}^{p_1}$, define $\tilde{\Sigma}^\delta$ as follows (here the block decomposition corresponds to decomposition $[p] = \{1\} \cup S \cup (S^c \setminus 1)$):

$$\tilde{\Sigma}^\delta = \begin{pmatrix} \|\check{\theta}\|^2 + \check{\sigma}^2 & 0 & \check{\theta}_S^{\mathsf{T}} & \check{\sigma}\delta^{\mathsf{T}} \\ 0 & 1 & 0_{1\times s_1} & \delta^{\mathsf{T}} \\ \check{\theta}_S & 0_{s_1\times 1} & \mathrm{I}_{s_1\times s_1} & 0_{s_1\times p_1} \\ \check{\sigma}\delta & \delta & 0_{p_1\times s_1} & \mathrm{I}_{p_1\times p_1} \end{pmatrix}. \tag{110}$$

We let $\mathcal{F} \equiv \{\tilde{\Sigma}^\delta : \delta \in \mathcal{A}(\nu, k)\}$ and construct the alternative space

$$H_1 = \Big\{(\theta, \Omega, \sigma^2) : \gamma = h(\tilde{\Sigma}^\delta) \text{ for some } \tilde{\Sigma}^\delta \in \mathcal{F}\Big\}. \tag{111}$$

We need to show that if $\tilde{\Sigma}^\delta \in \mathcal{F}$ then $h(\tilde{\Sigma}^\delta) \in \Gamma(s_0, s_\Omega, \rho)$. Let $(\theta, \Omega, \sigma^2) = h(\tilde{\Sigma}^\delta)$. Then,

$$\theta_1 = \frac{-\check{\sigma}\|\delta\|^2}{1-\|\delta\|^2},\ \theta_S = \check{\theta}_S,\ \theta_{S^c\backslash\{1\}} = (\check{\sigma} - \theta_1)\delta. \tag{112}$$

Therefore, $\|\check{\theta}\|_0 = 1 + |S| + \|\delta\|_0 = 1 + (s_1 - 1) + s_* = s_0$. Further, if $\nu \le 1/\sqrt{s_*}$, then $\|\delta\|_2 \le \sqrt{s_*}\nu < 1$ and we have

$$\sigma^2 = \|\check{\theta}\|^2 + \check{\sigma}^2 - \|\check{\theta}_S\|^2 - \check{\sigma}(\check{\sigma} - \theta_1)\|\delta\|^2 = \check{\sigma}^2 - \frac{\check{\sigma}^2\|\delta\|^2}{1-\|\delta\|^2} \le \check{\sigma}^2 < c\,. \tag{113}$$

Finally we note that

$$\Omega = \frac{1}{1-\|\delta\|^2}\begin{pmatrix} 1 & 0 & -\delta^{\mathsf{T}} \\ 0 & (1-\|\delta\|^2)\mathrm{I}_{s_1\times s_1} & 0_{s_1\times p_1} \\ -\delta & 0_{p_1\times s_1} & (1-\|\delta\|^2)\mathrm{I}_{p_1\times p_1} + \delta\delta^{\mathsf{T}} \end{pmatrix} \tag{114}$$

Hence, $\max_{i\in[p]} |\{j \ne i, \Omega_{i,j} \ne 0\}| = \|\delta\|_0 \le s_* \le s_\Omega$. Further, $(\Omega^{-1})_{ii} = 1$ for all $i \in [p]$. Also by Weyl's inequality, if $\|\delta\|_2 \le \sqrt{s_*}\nu \le \min(C_{\max} - 1, 1 - C_{\min})$, then $C_{\min} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le C_{\max}$. The last condition is on $\|\Omega\|_\infty$. We have

$$\|\Omega\|_\infty \le \frac{1+\|\delta\|_1}{1-\|\delta\|^2} \le \frac{\rho - 0.01}{1-(\rho-1.01)\nu} \le \rho\,, \tag{115}$$

where the second inequality is due to the fact that $\delta \in \mathcal{A}(\nu, k)$ and $k \le (\rho - 1.01)/\nu$. The last inequality holds if we choose $\nu < 0.01/(\rho(\rho - 1.01))$.

Summarizing, $(\theta, \Omega, \sigma^2) \in \Gamma(s_0, s_\Omega, \rho)$ if we choose

$$\nu \le \min\Big\{\frac{1}{\sqrt{s_*}}(C_{\max} - 1), \frac{1}{\sqrt{s_*}}(1 - C_{\min}), \frac{0.01}{\rho(\rho - 1.01)}\Big\}. \tag{116}$$

It is worth noting how our construction of alternative parameters space $H_1$ differs from that in [14][Theorem 3]. Here, in order to have $\|\Omega\|_\infty \le \rho$, we require the sparsity of $\delta$, denoted by $k$, and the magnitude of its nonzero entries, denoted by $\nu$, satisfy $k \le (\rho - 1.01)/\nu$ and $\nu \le 0.01/(\rho(\rho - 1.01))$. Further, in order to ensure the row sparsity assumption of the precision matrix, we require $k \le s_\Omega$. We enforce these additional requirements because we aim at establishing the minimax lower bound on a more restricted set, compared to the one in [14][Theorem 3].

Let $\pi$ be the uniform prior on $\delta$ over $\mathcal{A}(\nu, k)$ for a fixed $\nu$ (whose value is to be determined later) and denote by $\pi_{H_1}$ the induced prior over $H_1$. We define $f_1$ and $f_0$ as the density function of marginal distribution of data $(y, x)$ with priors $\pi_{H_0}$ and $\pi_{H_1}$ respectively. Precisely, for $\gamma = (\theta, \Omega, \sigma^2)$ and $i \in \{0, 1\}$, we have $f_i(y, x) = \int f_\gamma(y, x)\pi_i(\mathrm{d}\gamma)$, where $f_\gamma$ is the induced density on $(y, x)$ for random $x_i \sim \mathsf{N}(0, \Omega^{-1})$ and noise $w \sim \mathsf{N}(0, \sigma^2)$, with $y = \langle x, \theta\rangle + w$ when we fix the signal $\theta$.

Applying [14, Lemma 1], we have (noting that $\theta_1$, $\check{\theta}_1$ are deterministic)

$$\mathbb{E}_{\check{\gamma}}\{\ell(J_\alpha(y, X))\} \geq |\theta_1 - \check{\theta}_1|\Big(1 - 2\alpha - \mathrm{TV}(f_1, f_0)\Big)_+ \,, \tag{117}$$

where for two density functions $\mathrm{TV}(f_1, f_0) \equiv \int |f_1(z) - f_0(z)|\mathrm{d}z$ denotes their total variation distance. Also recall the $\chi^2$ distance between $f_1$ and $f_0$:

$$\chi^2(f_1, f_0) \equiv \int \frac{f_1^2(z)}{f_0(z)}\mathrm{d}z - 1\,.$$

It is well known that $\mathrm{TV}(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}$. Using [14, Lemma 2] we have

$$\chi^2(f_1, f_0) + 1 = \mathbb{E}_{\delta,\tilde{\delta}}(1 - 2\delta^{\mathsf{T}}\tilde{\delta})^{-n} \leq \mathbb{E}_{\delta,\tilde{\delta}} \exp(4n\delta^{\mathsf{T}}\tilde{\delta})\,, \tag{118}$$

for $\delta$ and $\tilde{\delta}$ two independent random draws from prior $\pi$ over $\mathcal{A}(\nu, k)$.

By [14, Lemma 3] we obtain

$$\mathbb{E}_{\delta,\tilde{\delta}} \exp(4n\delta^{\mathsf{T}}\tilde{\delta}) \leq e^{\frac{k^2}{p_1 - k}}\Big(1 - \frac{k}{p_1} + \frac{k}{p_1}e^{4n\nu^2}\Big)^k\,. \tag{119}$$

We set $\nu = c\sqrt{(\log p)/n}$. Since $k \leq s_0 \lesssim p^\eta$ for some constant $\eta \in [0, 1/2)$, by choosing $c$ small enough, we can ensure that $\mathrm{TV}(f_{\pi_{H_1}}, f_{\pi_{H_0}}) \leq 1/2 - \alpha$. Further, given that $s_* \leq s_0 \lesssim n/\log p$ and $\rho$ is a constant, condition (116) holds true for small enough $c$.

Finally, by invoking inequality (117) and substituting for $\theta_1$ from Equation (112) and $\check{\theta}_1 = 0$, we obtain

$$\mathbb{E}_{\check{\gamma}}\{\ell(J_\alpha(y, X))\} \geq \frac{\check{\sigma}\|\delta\|^2}{1 - \|\delta\|^2}\Big(\frac{1}{2} - \alpha\Big) \asymp k\nu^2 = \min(\rho\nu, s_*\nu^2)\,. \tag{120}$$

Note that the inequality (120) implies that $\ell^*_\alpha(\Gamma(s_0, s_\Omega)) \gtrsim \min(\rho\nu, s_*\nu^2)$. Using $\nu \asymp \sqrt{(\log p)/n}$ and $s_* = \min(s_0 - 1, s_\Omega)$, we get $\mathbb{E}_{\check{\gamma}}\{\ell(J_\alpha(y, X))\} \gtrsim$

$\min(\rho\sqrt{(\log p)/n}, s_*(\log p)/n)$. Proof of the lower bound rate $1/\sqrt{n}$ follows along the same lines as the proof in [14, Theorem 3].

It is worth noting that Equation (120) is much stronger than the implied minimax lower bound. Indeed it shows that the expected length of confidence intervals at any given point in a large subset of $\Gamma(s_0, s_\Omega)$, namely $\{(\check{\theta}, \mathrm{I}, \check{\sigma}) : \|\check{\theta}\|_0 = s_1 - 1, \check{\sigma}^2 \in (0, c]\}$, is at least of the provided lower bound rate.

## APPENDIX F: PROOF OF THEOREM 5.1 AND COROLLARY 5.2

**F.1. Proof of Theorem 5.1.** Throughout the proof, we will use $\widehat{\theta} = \widehat{\theta}^{\rm Lasso}(y, X)$ to denote the Lasso estimator. Using the KKT conditions, it is immediate to see that this satisfies

$$\widehat{\theta} = \eta_\Sigma(\widehat{\theta}^{\rm d}) = \eta_\Sigma\Big(\theta^* + \frac{1}{n}\Omega X^{\sf T} w + \frac{1}{\sqrt{n}} R\Big)\,, \tag{121}$$

with $R = \sqrt{n}(\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta})$ defined as in Theorem 3.8. We also define $\widehat{\theta}^0$ by

$$\widehat{\theta}^0 \equiv \eta_\Sigma\Big(\theta^* + \frac{1}{n}\Omega X^{\sf T} w\Big)\,. \tag{122}$$

Recall that $\widehat{S} = {\rm supp}(\widehat{\theta})$ is the support of the Lasso estimator. By Proposition 3.5, we have, with high probability $|\widehat{S}| \le C_* s_0$ for a constant $C_*$. Define $\widehat{S}^0 = {\rm supp}(\widehat{\theta}^0)$. Proceeding as in Proposition 3.5, we obtain, with high probability $|\widehat{S}^0| \le C_* s_0$ as well. Letting $\overline{S} \equiv \widehat{S} \cup \widehat{S}^0$, we have $|\overline{S}| \le 2C_* s_0$.

Write $z^0 \equiv \theta^* + n^{-1}\Omega X^{\sf T} w$, $r \equiv R/\sqrt{n}$. By Eq. (121), and the definition of $\eta_\Sigma(\,\cdot\,)$, cf. Eq. (47), we have

$$\frac{1}{2}\big\|\Sigma^{1/2}(\widehat{\theta} - z^0 - r)\big\|_2^2 + \lambda\|\widehat{\theta}\|_1 \le \frac{1}{2}\big\|\Sigma^{1/2}(\widehat{\theta}^0 - z^0 - r)\big\|_2^2 + \lambda\|\widehat{\theta}^0\|_1\,. \tag{123}$$

Expanding the squares on both sides, this can be rewritten as

$$\frac{1}{2}\big\|\Sigma^{1/2}(\widehat{\theta} - \widehat{\theta}^0)\big\|_2^2 - \langle r, \Sigma(\widehat{\theta} - \widehat{\theta}^0)\rangle \le -\langle(\widehat{\theta} - \widehat{\theta}^0), \Sigma(\widehat{\theta}^0 - z^0)\rangle + \lambda\|\widehat{\theta}^0\|_1 - \lambda\|\widehat{\theta}\|_1\,.$$

By the KKT conditions for $\widehat{\theta}^0$ (which follow from the definition (122), and the definition of $\eta_\Sigma$), there exists a vector $v(\widehat{\theta}^0)$ in the subgradient of the $\ell_1$ norm at $\widehat{\theta}^0$, such that $\Sigma(\widehat{\theta}^0 - z^0) + \lambda\, v(\widehat{\theta}^0) = 0$. Hence, by definition of subgradient

$$\frac{1}{2}\big\|\Sigma^{1/2}(\widehat{\theta} - \widehat{\theta}^0)\big\|_2^2 - \langle r, \Sigma(\widehat{\theta} - \widehat{\theta}^0)\rangle \le -\lambda\big[\|\widehat{\theta}\|_1 - \|\widehat{\theta}^0\|_1 - \langle v(\widehat{\theta}^0), (\widehat{\theta} - \widehat{\theta}^0)\rangle\big] \le 0\,.$$

Using the assumption $\sigma_{\min}(\Sigma) \ge C_{\min}$, we have

$$C_{\min}\|\widehat{\theta}-\widehat{\theta}^0\|_2^2 \le 2\langle \Sigma r, (\widehat{\theta}-\widehat{\theta}^0)\rangle \le 2\|(\Sigma r)_{\overline{S}}\|_2\|\widehat{\theta}-\widehat{\theta}^0\|_2\,.$$

Hence

$$\begin{aligned}
\|\widehat{\theta}-\widehat{\theta}^0\|_2^2 &\le \frac{4}{C_{\min}^2}\|(\Sigma r)_{\overline{S}}\|_2^2\\
&\le \frac{4}{C_{\min}^2}\|\Sigma_{\overline{S},\overline{S}}r_{\overline{S}}+\Sigma_{\overline{S},\overline{S}^c}r_{\overline{S}^c}\|_2^2\\
&\le \frac{8}{C_{\min}^2}\big\{\|\Sigma_{\overline{S},\overline{S}}r_{\overline{S}}\|_2^2+\|\Sigma_{\overline{S},\overline{S}^c}r_{\overline{S}^c}\|_2^2\big\}\\
&\le \frac{8}{C_{\min}^2}\big\{C_{\max}^2|\overline{S}|\|r\|_\infty^2+\tilde{\rho}^2|\overline{S}|\|r_{\overline{S}^c}\|_\infty^2\big\}\\
&\le \frac{32(C_{\max}^2+\tilde{\rho}^2)}{C_{\min}^2}\,C_*\,\frac{s_0}{n}\,\|R\|_\infty^2 \equiv \widetilde{C}^2\frac{s_0}{n}\,\|R\|_\infty^2\,. \qquad (124)
\end{aligned}$$

The proof is completed by using Theorem 3.8.

**F.2. Proof of Corollary 5.2.** As in the previous section, we use $\widehat{\theta} = \widehat{\theta}^{\text{Lasso}}(y,X)$ to denote the Lasso estimator. Recall that $\Sigma = \mathrm{I}$ and therefore the proximal operator $\eta_\Sigma$, given by (47) reduced to the component-wise soft thresholding $\eta(\cdot;\lambda)$. Define $\widehat{\theta}^0$ by

$$\widehat{\theta}^0 \equiv \eta\Big(\theta^* + \frac{1}{n}X^{\mathsf{T}}w;\lambda\Big)\,. \qquad (125)$$

Note that, by Lemma 3.4, we have $\|X^{\mathsf{T}}w/n\|_\infty < \lambda$ with high probability, whence $\widehat{S}^0 \equiv \mathrm{supp}(\widehat{\theta}^0) \subseteq S \equiv \mathrm{supp}(\theta^*)$. By triangular inequality and Theorem 5.1, we get

$$\begin{aligned}
\|\widehat{\theta}-\theta^*\|_2 &= \|\widehat{\theta}^0-\theta^*\|_2 + O_P\Big(\frac{\sigma s_0\log p}{n}\Big)\\
&= \|(\widehat{\theta}^0-\theta^*)_S\|_2 + O_P\Big(\frac{\sigma s_0\log p}{n}\Big)\,. \qquad (126)
\end{aligned}$$

We next show that $\|(\widehat{\theta}^0-\theta^*)_S\|$ concentrates around its expectation. Fixing $X \in \mathbb{R}^{n\times p}$, define

$$F(w;X) = \|\widehat{\theta}^0_S - \theta^*_S\|_2 = \Big\|\eta(\theta^* + \frac{1}{n}\Omega X^{\mathsf{T}}w + \frac{1}{\sqrt{n}}R)_S - \theta^*_S\Big\|_2\,.$$

Noting that the soft-thresholding function $\eta(\cdot;\lambda)$ is 1-Lipschitz continuous, we have

$$\begin{aligned}
F(w;X)-F(w';X) &= \Big\|\eta(\theta^*+\frac{1}{n}\Omega X^{\mathsf T}w+\frac{1}{\sqrt{n}}R)_S-\theta^*_S\Big\|_2\\
&\quad-\Big\|\eta(\theta^*+\frac{1}{n}\Omega X^{\mathsf T}w'+\frac{1}{\sqrt{n}}R)_S-\theta^*_S\Big\|_2\\
&\le \Big\|\eta\Big(\theta^*+\frac{1}{n}X^{\mathsf T}w;\lambda\Big)_S-\eta\Big(\theta^*+\frac{1}{n}X^{\mathsf T}w';\lambda\Big)_S\Big\|_2\\
&\le \frac{1}{n}\big\|(X^{\mathsf T}w-X^{\mathsf T}w')_S\big\|_2\\
&\le \frac{1}{n}\|X_S\|_2\,\|w-w'\|_2\,. \qquad (127)
\end{aligned}$$

Next by the Bai-Yin law [2]), we have $\|X_S\|_2\le 2(\sqrt{s_0}+\sqrt{n})$, with high probability. Therefore, using $s_0\le n$, we obtain $F(w;X)-F(w';X)\le 4\|w-w'\|_2/\sqrt{n}$.

Denote by $\mathbb{P}_w$ and $\mathbb{E}_w$ probability and expectation with respect to $w$. By Gaussian isoperimetry [44], we have $\mathbb{P}(F(w;X)-\mathbb{E}_w\{F(w;X)\}\ge t)\le 2\,e^{-cnt^2/\sigma^2}$, for some universal constant $c>0$ . This implies $\mathbb{E}_w\|(\widehat{\theta}^0-\theta^*)_S\|_2=\mathbb{E}_w\{\|(\widehat{\theta}^0-\theta^*)_S\|_2^2\}^{1/2}+O(\sigma/\sqrt{n})$, and therefore

$$\|(\widehat{\theta}^0-\theta^*)_S\|_2\le \mathbb{E}_w\{\|(\widehat{\theta}^0-\theta^*)_S\|_2^2\}^{1/2}+\frac{t\sigma}{\sqrt{n}}\,, \qquad (128)$$

with probability at least $1-2e^{-ct^2}$. Using this together with Eq. (126), we get

$$\begin{aligned}
\|\widehat{\theta}-\theta^*\|_2 &= \|\widehat{\theta}^0-\theta^*\|_2+O_P\Big(\frac{\sigma s_0\log p}{n}\Big)\\
&= \sqrt{\mathbb{E}_w\big\{\|(\widehat{\theta}^0-\theta^*)_S\|_2^2\big\}}+O_P\Big(\frac{\sigma}{\sqrt{n}}\vee\frac{\sigma s_0\log p}{n}\Big)\\
&= \sqrt{\sum_{i\in\mathrm{supp}(\theta^*)}\mathbb{E}\big\{\big[\eta(\theta^*_i+n^{-1/2}\tilde{Z}_i;\lambda)-\theta^*_i\big]^2\big\}}+O_P\Big(\frac{\sigma}{\sqrt{n}}\vee\frac{\sigma s_0\log p}{n}\Big)\,,
\end{aligned}$$

where in the last equality expectation is with respect to $\tilde{Z}_i\sim\mathsf{N}(0,\|\tilde{x}_i\|_2^2/n)$. The proof is completed by using the fact that, $\max_{i\in[p]}\big|\|\tilde{x}_i\|_2^2/n-1\big|\le C\sqrt{(\log p)/n}$, with high probability, and bounding the resulting error.

## APPENDIX G: PROOF OF THEOREM 5.3

Throughout this proof, we denote by $\mathbb{P}_w$ and $\mathbb{E}_w$, the probability and the expectation with respect to the noise vector $w$ (conditional on $X$). Let

$$\overline{\tau}_i \equiv \sqrt{\frac{2\sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}\log(p/s_0)}{n}}\,, \tag{129}$$

and define the estimators $\widehat{\theta}^{(1)}$, $\overline{\theta}^{(1)}$ by

$$\overline{\theta}^{(1)}_i \equiv \eta\Big(\theta^*_i + \frac{1}{n}(\Omega X^{\mathsf{T}} w)_i; \overline{\tau}_i\Big)\,, \tag{130}$$

$$\widehat{\theta}^{(1)}_i \equiv \eta\Big(\theta^*_i + \frac{1}{n}(\Omega X^{\mathsf{T}} w)_i; \tau_i\Big)\,. \tag{131}$$

Throughout this section, $L_n$ denotes a deterministic sequence with $L_n \to \infty$ arbitrarily slow as $n \to \infty$. First we claim that,, $\|\overline{\theta}^{(1)}\|_0, \|\widehat{\theta}^{(1)}\|_0 \le s_0 L_n$ with high probability for any such sequence $L_n$. In other to prove this, recall that $S \equiv \text{supp}(\theta^*)$, and consider $i \notin S$. Conditional on $X$, we have $(\Omega X^{\mathsf{T}} w)_i/n \sim \mathsf{N}(0, \sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}/n)$. Hence, for $Z \sim \mathsf{N}(0,1)$ independent of $X$ and $W_n$ being a chi-squared random variable with $n$ degrees of freedom, we get

$$\begin{aligned}
\mathbb{P}(\widehat{\theta}^{(1)}_i \neq 0) &= \mathbb{P}\big(|(\Omega X^{\mathsf{T}} w)_i| \ge n\tau_i\big)\\
&= \mathbb{P}\Big(|(\Omega\widehat{\Sigma}\Omega)^{1/2}_{ii} Z| \ge \sqrt{2\Omega_{ii}\log(p/s_0)}\Big)\\
&\le \mathbb{P}\Big(|Z| \ge \sqrt{2(1+\delta)^{-1}\log(p/s_0)}\Big) + \mathbb{P}\big((\Omega\widehat{\Sigma}\Omega)_{ii} \ge (1+\delta)\Omega_{ii}\big)\\
&\le \Big(\frac{s_0}{p}\Big)^{1-\delta} + \mathbb{P}(W_n \ge n(1+\delta))\\
&\le \Big(\frac{s_0}{p}\Big)^{1-\delta} + e^{-cn\delta^2} \le \frac{Cs_0}{p}\,,
\end{aligned}$$

where the last inequality follows with high probability by taking $\delta = C_0\sqrt{\log(p/s_0)/n}$, and using the assumption that $\log(p/s_0)^3/n \to 0$. Hence, by Markov inequality $\|\widehat{\theta}^{(1)}\|_0 \le s_0 L_n$ with high probability. The claim follows by the same argument for $\overline{\theta}^{(1)}$.

We next claim that $\|\widehat{\theta}^{(2)}\|_0 \le s_0 L_n$ with high probability as well. Indeed, by definition

$$\widehat{\theta}^{(2)}_i \equiv \eta\Big(\theta^*_i + \frac{1}{n}(\Omega X^{\mathsf{T}} w)_i + \frac{1}{\sqrt{n}}R_i; \tau_i\Big)\,, \tag{132}$$

Using the fact that $\|R\|_\infty \le C\sigma\sqrt{s_0(\log p)^2/n}$, with high probability (cf. Theorem 3.8) and proceeding along the same lines as above, we obtain

$$\begin{aligned}
\mathbb{P}(\widehat{\theta}_i^{(2)} \neq 0) &\le \mathbb{P}\left(|(\Omega\widehat{\Sigma}\Omega)_{ii}^{1/2} Z| \ge \sqrt{2\Omega_{ii}\log(p/s_0)} - C\sqrt{\frac{s_0(\log p)^2}{n}}\right) \\
&\quad + \mathbb{P}\left(\|R\|_\infty > C\sigma\sqrt{\frac{s_0(\log p)^2}{n}}\right) \\
&\le \left(\frac{s_0}{p}\right)^{1-\delta} \exp\left\{C\sqrt{\frac{s_0(\log p)^2\log(p/s_0)}{n}}\right\} + e^{-cn\delta^2} + o(1) \\
&\le \frac{Cs_0}{p},
\end{aligned}$$

where in the final step we used the assumption $s_0(\log p)^3/n \to 0$. Hence, by Markov inequality, we have $\|\widehat{\theta}^{(2)}\|_0 \le s_0 L_n$, with high probability as claimed.

Therefore , with high probability,

$$\begin{aligned}
\|\widehat{\theta}^{(2)} - \widehat{\theta}^{(1)}\|_2 &\le \frac{1}{\sqrt{n}}\|R\|_\infty \sqrt{\|\widehat{\theta}^{(2)}\|_0 + \|\widehat{\theta}^{(1)}\|_0} \\
&\le \frac{C\sigma}{\sqrt{n}}\sqrt{\frac{s_0(\log p)^2}{n}}\sqrt{2s_0L_n} = \sqrt{2L_n}C\sigma\,\frac{s_0\log p}{n}\,. \qquad (133)
\end{aligned}$$

Analogously, we have

$$\begin{aligned}
\|\overline{\theta}^{(1)} - \widehat{\theta}^{(1)}\|_2 &\le \sqrt{\|\overline{\theta}^{(1)}\|_0 + \|\widehat{\theta}^{(1)}\|_0} \cdot \max_{i\in[p]} |\hat{\tau}_i - \tau_i| \\
&\le \sqrt{s_0L_nC_{\max}}\sqrt{\frac{2\sigma^2\log(p/s_0)}{n}} \cdot \max_{i\in[p]} \left|(\Omega\widehat{\Sigma}\Omega)_{ii} - \Omega_{ii}\right|,
\end{aligned}$$

where we used the fact that $C_{\max}^{-1} \le \Omega_{ii} \le C_{\min}^{-1}$ is bounded uniformly and $|\sqrt{x} - \sqrt{y}| \le |x-y|/\sqrt{4c}$ for $x, y \ge c$. Since $(\Omega\widehat{\Sigma}\Omega)_{ii}/\Omega_{ii}$ is distributed as $W_n/n$, for $W_n$ a chi-squared random variable with $n$ degrees of freedom, and $\Omega_{ii} \le C_{\min}^{-1}$, we have $\max_{i\in[p]} \left|(\Omega\widehat{\Sigma}\Omega)_{ii} - \Omega_{ii}\right| \le C\sqrt{(\log p)/n}$. Substituting above, we get

$$\|\overline{\theta}^{(1)} - \widehat{\theta}^{(1)}\|_2 \le \sqrt{2L_nC_{\max}}C\sigma\,\frac{\sqrt{s_0}\log p}{n}\,. \qquad (134)$$

Hence, using triangular inequality together with Equations. (133) and (134), we obtain

$$\begin{aligned}
\|\widehat{\theta}^{(2)} - \theta^*\|_2 &\le \|\overline{\theta}^{(1)} - \theta^*\|_2 + C\sigma\sqrt{L_n}\,\frac{s_0\log p}{n} \\
&\le \|\overline{\theta}^{(1)}_S - \theta^*_S\|_2 + \|\overline{\theta}^{(1)}_{S^c}\|_2 + C\sigma\sqrt{L_n}\,\frac{s_0\log p}{n}\,, \qquad (135)
\end{aligned}$$

for some constant $C > 0$.

We are left with the task of bounding $\|\overline{\theta}^{(1)}_S - \theta^*_S\|_2$ and $\|\overline{\theta}^{(1)}_{S^c}\|_2$.

- *Bounding* $\|\overline{\theta}^{(1)}_S - \theta^*_S\|_2$. Fixing $X \in \mathbb{R}^{n\times p}$, we let $F(w;X) \equiv \|\overline{\theta}^{(1)}_S - \theta^*_S\|_2$. Letting $\sigma_i^2 \equiv \sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}/n$, and denoting by $Z \sim \mathsf{N}(0,1)$ a standard Gaussian random variable, we have

$$\begin{aligned}\mathbb{E}_w\{F(w;X)^2\} &\overset{(a)}{=} \sum_{i\in S}\mathbb{E}_Z\Big\{\big[\eta\big(\theta^*_i + \sigma_i Z;\sigma_i\sqrt{2\log(p/s_0)}\big) - \theta^*_i\big]^2\Big\}\\ &\overset{(b)}{\le} 2\log(p/s_0)\sum_{i\in S}\sigma_i^2\\ &\overset{(c)}{\le} \frac{2s_0\sigma^2}{n}\log(p/s_0)\left(\frac{1}{s_0}\sum_{i\in S}\Omega_{ii}\right)\ \Big\{1 + C\sqrt{\frac{\log p}{n}}\Big\} \equiv \overline{F}^2\,. \end{aligned} \tag{136}$$

Here, $(a)$ follows because $(\Omega X^{\mathsf{T}} w)/n \sim \mathsf{N}(0,\sigma_i^2)$, $(b)$ because the soft-thresholding risk is maximized for $\theta^*_i \to \infty$[28, 24, 29], and $(c)$ because, as remarked above, with high probability we have $\max_{i\in[p]}\big|(\Omega\widehat{\Sigma}\Omega)_{ii} - \Omega_{ii}\big| \le C\sqrt{(\log p)/n}$.

Recall that $\overline{F}^2$ denotes the upper bound on the right-hand side of Eq. (136). Let $\mathcal{G}_0$ denote the set of matrices $X$ for which the bound $\mathbb{E}_w\{F(w;X)^2\} \le \overline{F}^2$ holds. By above argument $\mathbb{P}(\mathcal{G}_0) \to 1$ as $n,p\to\infty$. Now note that, since $\eta(\,\cdot\,;\tau)$ is Lipschitz continuous (with Lipschitz constant equal to one), and denoting by $\widehat{\theta}^{(1)}(w)$ the vector defined in Eq. (131) with noise vector $w$, we have

$$\begin{aligned}\big|F(w;X) - F(w';X)\big| &\le \|\widehat{\theta}^{(1)}(w)_S - \widehat{\theta}^{(1)}(w')_S\|_2\\ &\le \frac{1}{n}C_{\min}^{-1}\|X_S\|_2\|w - w'\|_2\,.\end{aligned}$$

By the Bai-Yin law, we have $\|X_S\|_2 \le 2(\sqrt{n} + \sqrt{s_0}) \le 4\sqrt{n}$ with high probability (since $s_0 \le n$). Define, $\mathcal{G} = \mathcal{G}_0 \cap \{X \in \mathbb{R}^{n\times p}:\ \|X_S\|_2 \le 4\sqrt{n}\}$. By Gaussian isoperimetry [44], we have, on $\mathcal{G}$, $\mathbb{P}_w\Big(F(w;X) \ge \mathbb{E}_w\{F(w;X)\} + t\Big) \le e^{-cnt^2/\sigma^2}$. This implies $\mathbb{E}\{F(w;X)\} = \overline{F} + O(\sigma/\sqrt{n})$. Hence, with high probability,

$$\big\|\overline{\theta}^{(1)}_S - \theta^*_S\big\|_2 \le \overline{F} + \frac{L_n\sigma}{\sqrt{n}}\,. \tag{137}$$

- *Bounding* $\|\overline{\theta}^{(1)}_{S^c}\|_2$. As above, we let $\sigma_i^2 \equiv \sigma^2(\Omega\widehat{\Sigma}\Omega)_{ii}/n$. Denoting by

$Z \sim \mathsf{N}(0,1)$ a standard Gaussian random variable, we write

$$\begin{aligned}
\mathbb{E}_w\{\|\overline{\theta}^{(1)}_{S^c}\|_2^2\} &= \sum_{i\in S^c} \mathbb{E}_Z\Big\{\eta\big(\sigma_i\, Z; \sigma_i\sqrt{2\log(p/s_0)}\big)^2\Big\} \\
&\stackrel{(a)}{\le} \sum_{i\in S^c} \sigma_i^2 \mathbb{E}_Z\Big\{\eta\big(Z; \sqrt{2\log(p/s_0)}\big)^2\Big\} \\
&\stackrel{(b)}{\le} C\frac{s_0}{p}\sum_{i\in S^c}\sigma_i^2 \\
&= C\frac{s_0\sigma^2}{np}\mathrm{Trace}(\Omega\widehat{\Sigma}\Omega)\,.
\end{aligned}$$

Here, $(a)$ follows because $\eta(c\,x; c\lambda) = c\,\eta(x;\lambda)$ and $(b)$ by a Gaussian integral calculation. As mentioned above, $(\Omega\widehat{\Sigma}\Omega)_{ii}/\Omega_{ii}$ is distributed as $W_n/n$ for $W_n$ a chi-squared random variable with $n$ degrees of freedom. Tail bounds on chi-squared random variables, together with the fact that $\Omega_{ii} \le C_{\min}^{-1}$ is bounded uniformly, imply that $\mathrm{Trace}(\Omega\widehat{\Sigma}\Omega) \le Cp$, with high probability. Hence, with high probability with respect to the choice of $X$, $\mathbb{E}_w\{\|\overline{\theta}^{(1)}_{S^c}\|_2^2\} \le Cs_0\sigma^2/n$ for some constant $C > 0$. Hence, with high probability

$$\|\overline{\theta}^{(1)}_{S^c}\|_2 \le \sigma\sqrt{\frac{s_0 L_n}{n}}\,. \tag{138}$$

The proof is completed by putting together Equations (135), (137), (138) and setting $L_n = \log p$.

## APPENDIX H: PROOF OF THEOREM 5.4

Throughout this section, we let $\widehat{\theta} = \widehat{\theta}^{\text{Lasso}}$ denote the Lasso estimator. Define $\widehat{\theta}^0$ by

$$\widehat{\theta}^0 \equiv \eta\Big(\theta^* + \frac{1}{n}X^{\mathsf{T}}w; \lambda\Big)\,, \tag{139}$$

where $\eta(\,\cdot\,;\lambda)$ is componentwise soft thresholding, defined for scalars via $\eta(x;\lambda) \equiv (|x|-\lambda)_+\mathrm{sign}(x)$. Further we denote by $\mathbb{P}_w$ and $\mathbb{E}_w$ probability and expectation with respect to $w$ (conditional on $X$). Finally, let $\widehat{S} \equiv \mathrm{supp}(\widehat{\theta})$, $\widehat{S}^0 \equiv \mathrm{supp}(\widehat{\theta}^0)$ and $\overline{S} \equiv \widehat{S} \cup \widehat{S}^0$.

Expanding the square in the definition of $\widehat{\mathsf{R}}(y,X)$, we obtain

$$\begin{aligned}
\mathsf{R}(y,X,\theta^*) - \widehat{\mathsf{R}}(y,X) &= \frac{2}{n}\,\langle w, X(\widehat{\theta}-\theta^*)\rangle \\
&= \frac{2}{n}\,\langle w, X(\widehat{\theta}^0-\theta^*)\rangle + \frac{2}{n}\,\langle w, X(\widehat{\theta}-\widehat{\theta}^0)\rangle \\
&\equiv \Delta_1(w,X,\theta^*) + \Delta_2(w,X,\theta^*)\,. \qquad (140)
\end{aligned}$$

We will separately study the error terms $\Delta_1$ and $\Delta_2$.

We start by considering a preliminary remark.

LEMMA H.1. *Let $X \in \mathbb{R}^{n\times p}$ have iid entries $X_{ij} \sim \mathsf{N}(0,1)$, and define*

$$\mathcal{G}_1(M) \equiv \left\{ X \in \mathbb{R}^{n\times p} :\ \max_{i\in[p]} \Big|\frac{\|\tilde{x}_i\|_2^2}{n} - 1\Big| \le M\sqrt{\frac{\log p}{n}},\ \ \max_{i\neq j\in[p]} \Big|\frac{\langle \tilde{x}_i, \tilde{x}_j\rangle}{\|\tilde{x}_i\|_2\|\tilde{x}_j\|_2}\Big| \le M\sqrt{\frac{\log p}{n}} \right\}.$$

*Then, for $M$ a large enough constant, we have $\mathbb{P}(X \in \mathcal{G}_1(M)) \ge 1 - p^{-10}$.*

*Further, under the assumptions of Theorem 5.4, we have $\mathbb{P}(\overline{S} \subseteq S) \ge 1 - p^{-3}$.*

PROOF. The lower bound on $\mathbb{P}(X \in \mathcal{G}_1(M))$ is standard, and follows from union bound along with tail bounds on chi-squared random variables.

As for the lower bound on $\mathbb{P}(\overline{S} \subseteq S)$, using the definition (139) we get that

$$\begin{aligned}
\mathbb{P}(\widehat{S}^0 \not\subseteq S) &\le \mathbb{P}(\widehat{S}^0 \not\subseteq S;\ X \in \mathcal{G}_1(M)) + \mathbb{P}(X \notin \mathcal{G}_1(M)) \\
&\le \sum_{i\in S^c} \mathbb{P}\Big(\big|\frac{1}{n}(X^{\mathsf{T}}w)_i\big| \ge \lambda;\ X \in \mathcal{G}_1(M)\Big) + p^{-10} \\
&\le p\,\mathbb{P}\Big(\frac{1.1\sigma}{\sqrt{n}}|Z| \ge \lambda\Big) + p^{-10}.
\end{aligned}$$

where, in the last expression $Z \sim \mathsf{N}(0,1)$, and we used $\max_{i\in[p]} \|\tilde{x}_i\|_2 \le 1.1$. The claim then follows by a direct calculation.

In order to bound $\mathbb{P}(\widehat{S} \not\subseteq S)$ note that, by definition,

$$\widehat{\theta} = \eta\Big(\theta^* + \frac{1}{n}X^{\mathsf{T}}w + \frac{1}{\sqrt{n}}R; \lambda\Big). \tag{141}$$

The proof follows the same lines as above noting that, by Theorem 3.8, $\|R\|_\infty/\sqrt{n} \le \lambda/100$ with high probability. □

LEMMA H.2. *Under the assumptions of Theorem 5.4, there exists a constant $C$ such that, with high probability*

$$|\Delta_2| \le \frac{Cs_0\sigma^2}{n}\sqrt{\frac{s_0(\log p)^3}{n}}. \tag{142}$$

PROOF. We have

$$\begin{aligned}|\Delta_2| &\le \frac{2}{n}\|(X^{\mathsf T}w)_{\overline{S}}\|_2\|\widehat{\theta}-\widehat{\theta}^0\|_2\\ &\le \frac{2}{n}\sqrt{|\overline{S}|}\|X^{\mathsf T}w\|_\infty\|\widehat{\theta}-\widehat{\theta}^0\|_2\\ &\le \frac{2}{n}\sqrt{s_0}\cdot 2\sigma\sqrt{n\log p}\cdot \widetilde{C}\sqrt{\frac{s_0}{n}}\,\|R\|_\infty\,,\end{aligned}$$

where the last inequality follows from Lemma H.1 along with the bound (124), for $\Sigma = \mathrm{I}$. Using Theorem 3.8 we obtain the claim. □

Next consider term $\Delta_1$ in the decomposition (140). We first compute its expectation with respect to the noise vector $w$.

LEMMA H.3. *Assume $X$ to have i.i.d. rows $x_i \sim \mathsf{N}(0,\Sigma)$. Then we have, with high probability with respect to the choice of $X$,*

$$\left|\mathbb{E}_w\{\Delta_1\} - \frac{2\sigma^2}{n}\mathbb{E}_w\{\|\widehat{\theta}^0\|_0\}\right| \le \Big(C\sigma^2\sqrt{\frac{\log p}{n^3}}\Big)\,\mathbb{E}_w\{\|\widehat{\theta}^0\|_0\}\,. \tag{143}$$

PROOF. Using Stein's lemma, we get

$$\begin{aligned}\mathbb{E}_w\{\Delta_1\} &= \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^p X_{ij}\mathbb{E}_w\{w_i(\widehat{\theta}^0-\theta^*)_j\}\\ &= \frac{2\sigma^2}{n}\sum_{i=1}^n\sum_{j=1}^p X_{ij}\mathbb{E}_w\Big\{\frac{\partial\widehat{\theta}^0_j}{\partial w_i}\Big\}\,. \qquad (144)\end{aligned}$$

By differentiating the KKT conditions that follow from the definition of $\eta_\Sigma$, cf. Eq. (47), we get that for $y = \eta_\Sigma(z)$, the following holds true

$$\frac{\partial y_j}{\partial z_k} = \mathbb{I}(y_j \ne 0)\big[(\Sigma_{TT})^{-1}\Sigma_{T,\cdot}\big]_{jk}\,, \tag{145}$$

where $T = \mathrm{supp}(y)$. Recall that $\widehat{\theta}^0 = \eta_\Sigma(z)$ with $z = \theta^* + n^{-1}\Omega X^{\mathsf T}w$ and $\widehat{S}^0 = \mathrm{supp}(\widehat{\theta}^0)$. Therefore,

$$\begin{aligned}\frac{\partial\widehat{\theta}^0_j}{\partial w_i} &= \sum_{k'=1}^p \frac{\partial\widehat{\theta}^0_j}{\partial z_{k'}}\frac{\partial z_{k'}}{\partial w_i} = \sum_{k'=1}^p \frac{1}{n}\mathbb{I}(\widehat{\theta}^0_j \ne 0)\big[(\Sigma_{\widehat{S}^0\widehat{S}^0})^{-1}\Sigma_{\widehat{S}^0,\cdot}\big]_{jk'}(\Omega X^{\mathsf T})_{k'i}\\ &= \frac{1}{n}\mathbb{I}(\widehat{\theta}^0_j \ne 0)\sum_{k=1}^p\Big(\sum_{k'=1}^p\big[(\Sigma_{\widehat{S}^0\widehat{S}^0})^{-1}\Sigma_{\widehat{S}^0,\cdot}\big]_{jk'}\Omega_{k'k}\Big)X_{ik}\\ &= \frac{1}{n}\,\mathbb{I}(\widehat{\theta}^0_j \ne 0)\sum_{k\in\widehat{S}^0}(\Sigma_{\widehat{S}^0\widehat{S}^0})^{-1}_{jk}X_{ik}\,.\end{aligned}$$

Substituting in Eq. (144), after some manipulations we get

$$\mathbb{E}_w\{\Delta_1\} = \frac{2\sigma^2}{n}\,\mathbb{E}_w\big\{\mathrm{Trace}\big((\Sigma_{\widehat{S}^0\widehat{S}^0})^{-1}\widehat{\Sigma}_{\widehat{S}^0,\widehat{S}^0}\big)\big\}\,. \tag{146}$$

Using [42, Lemma 23], we have $|\Sigma^{-1}\widehat{\Sigma} - \mathrm{I}|_\infty \le C\sqrt{(\log p)/n}$, with high probability. Hence,

$$\mathbb{E}_w\{\mathrm{Trace}\big((\Sigma_{\widehat{S}^0\widehat{S}^0})^{-1}\widehat{\Sigma}_{\widehat{S}^0,\widehat{S}^0}\big) - |\widehat{S}^0|\} \le C\sqrt{\frac{\log p}{n}}\mathbb{E}_w\{|\widehat{S}^0|\}\,. \tag{147}$$

The claim follows. □

Lemma H.4. *Under the assumptions of Theorem 5.4, the following holds*

$$\mathbb{P}\left(\Big|\Delta_1 - \mathbb{E}_w\Delta_1\Big| \ge \frac{t\sigma^2}{\sqrt{n}}\right) \le 2\,e^{-ct^2} + o_n(1)\,. \tag{148}$$

Proof. Define the event

$$\mathcal{G}_2(M) \equiv \mathcal{G}_1(M) \cap \Big\{X \in \mathbb{R}^{n\times p}:\ \lambda_{\max}(\widehat{\Sigma}_{S,S}) \le 2\Big\}\,. \tag{149}$$

Using Lemma H.1, together with standard tail bounds on the singular values of Wishart matrices [2], we get $\mathbb{P}(X \in \mathcal{G}_2(M)) \ge 1 - p^{-5} - e^{-cn}$.

Define the set

$$\mathcal{C} \equiv \Big\{w \in \mathbb{R}^n:\ \frac{1}{n}\big\|X^{\mathsf{T}}w\big\|_\infty \le \lambda;\ \ \|w\|_2^2 \le 2n\sigma^2\Big\}\,. \tag{150}$$

By a union bound argument, it is immediate to see that, for any $X \in \mathcal{G}_2(M)$, $\mathbb{P}(w \notin \mathcal{C}) \le p^{-6} + e^{-cn}$. Further note the following:

1. $\mathcal{C}$ is convex.
2. For $w \in \mathcal{C}$, we have $\widehat{S}^0 \subseteq S$.
3. As a consequence, for $w \in \mathcal{C}$,

$$\begin{aligned}\|\widehat{\theta} - \theta^*\|_2^2 &\le s_0\|\widehat{\theta}_S - \theta^*_S\|_\infty^2\\ &\le s_0\Big(\lambda + \frac{1}{n}\|X^{\mathsf{T}}w\|_\infty\Big)^2\\ &\le 4s_0\lambda^2\,.\end{aligned}$$

In order to prove the lemma, we will use Gaussian concentration [44], by proving that $w \mapsto \Delta_1(w, X, \theta^*)$ is Lipschitz continuous on $\mathcal{C}$. We have

$$\begin{aligned}\frac{\partial\Delta_1}{\partial w_i} &= \frac{2}{n}\langle x_i, (\widehat{\theta}^0 - \theta^*)\rangle + \frac{2}{n^2}\sum_{j\in\widehat{S}^0}(X^Tw)_jX_{ij}\\ &= \frac{2}{n}\big(X(\widehat{\theta}^0 - \theta^*)\big)_i + \frac{2}{n^2}(X\mathrm{P}_{\widehat{S}^0}X^{\mathsf{T}}w)_i\,,\end{aligned}$$

where $\mathrm{P}_{\widehat{S}^0} \in \mathbb{R}^{p\times p}$ is the projector onto the indices in $\widehat{S}^0$. Namely, $(\mathrm{P}_{\widehat{S}^0})_{ij} = 0$ if $i \neq j$, and $(\mathrm{P}_{\widehat{S}^0})_{ii} = \mathbb{I}(i \in \widehat{S}^0)$. Hence

$$\begin{aligned}\big\|\nabla\Delta_1\big\|_2^2 &\le \frac{8}{n^2}\langle(\widehat{\theta}^0-\theta^*), X^{\mathsf{T}}X(\widehat{\theta}^0-\theta^*)\rangle + \frac{8}{n^4}\big\|X\mathrm{P}_{\widehat{S}^0}X^{\mathsf{T}}w\big\|_2^2\\ &\le \frac{8}{n}\langle(\widehat{\theta}^0-\theta^*)_S, \widehat{\Sigma}_{SS}(\widehat{\theta}^0-\theta^*)_S\rangle + \frac{8}{n^4}\big\|X\mathrm{P}_{\widehat{S}^0}X^{\mathsf{T}}w\big\|_2^2\\ &\le \frac{8}{n}\lambda_{\max}(\widehat{\Sigma}_{SS})\|\widehat{\theta}^0-\theta^*\|_2^2 + \frac{8}{n^4}\big\|X\mathrm{P}_{\widehat{S}^0}X^{\mathsf{T}}\big\|_2^2\|w\|_2^2\,.\end{aligned}$$

Next note that

$$\begin{aligned}\frac{1}{n}\big\|X\mathrm{P}_{\widehat{S}^0}X^{\mathsf{T}}\big\|_2 &= \frac{1}{n}\big\|X\mathrm{P}_{\widehat{S}^0}\big\|_2^2\\ &= \frac{1}{n}\big\|\mathrm{P}_{\widehat{S}_0}X^{\mathsf{T}}X\mathrm{P}_{\widehat{S}^0}\big\|_2\\ &= \lambda_{\max}(\widehat{\Sigma}_{\widehat{S}^0,\widehat{S}^0}) \le \lambda_{\max}(\widehat{\Sigma}_{S,S})\,.\end{aligned}$$

Substituting above, and using $X \in \mathcal{G}_2(M)$, we get

$$\begin{aligned}\big\|\nabla_w\Delta_1\big\|_2^2 &\le \frac{8}{n}\lambda_{\max}(\widehat{\Sigma}_{SS})\Big\{\|\widehat{\theta}^0-\theta^*\|_2^2 + \frac{1}{n}\lambda_{\max}(\widehat{\Sigma}_{SS})\,\|w\|_2^2\Big\}\\ &\le \frac{16}{n}\big(4s_0\lambda^2 + 2\sigma^2\big)\\ &\le \frac{16}{n}\Big(\frac{4Cs_0\sigma^2\log p}{n} + 2\sigma^2\Big)\\ &\le \frac{C\sigma^2}{n}\,.\end{aligned}$$

Hence, using Gaussian concentration [44] (applied to the Lipschitz extension of $\Delta_1$ from $w \in \mathcal{C}$ to $w \notin \mathcal{C}$), we get

$$\begin{aligned}\mathbb{P}_w\Big(\big|\Delta_1 - \mathrm{Med}_w(\Delta_1)\big| \ge t\Big) &\le \mathbb{P}_w\Big(\big|\Delta_1 - \mathrm{Med}_w(\Delta_1)\big| \ge t;\ w\in\mathcal{C}\Big) + \mathbb{P}_w\big(w\notin\mathcal{C}\big)\\ &\le 2e^{-nt^2/C\sigma^4} + \mathbb{P}_w\big(w\notin\mathcal{C}\big)\,,\end{aligned}$$

where $\mathrm{Med}_w(\,\cdot\,)$ denotes the median w.r.t the measure $\mathbb{P}_w$. The claim follows by bounding $|\mathrm{Med}_w(\Delta_1) - \mathbb{E}_w\{\Delta_1\}|$ in the standard way, and using the fact that $\mathbb{P}(X \notin \mathcal{G}_2(M))$, $\mathbb{P}(w \notin \mathcal{C}) \to 0$. □

LEMMA H.5. *Fix $X \in \mathcal{G}_1(M)$, and let $L_n$ be any sequence with $L_n \to \infty$ as $n \to \infty$. Then, we have*

$$\mathbb{E}_w\{\|\widehat{\theta}^0\|_0\} \le s_0 + 1\,, \tag{151}$$

$$\mathbb{P}_w\Big(\big|\|\widehat{\theta}^0\|_0 - \mathbb{E}_w\{\|\widehat{\theta}^0\|_0\}\big| \ge \frac{L_n s_0(\log p)^{1/4}}{n^{1/4}}\Big) \le \frac{M}{L_n^2}\,. \tag{152}$$

PROOF. By Lemma H.1, $\mathbb{P}(\widehat{\theta}^0_{S^c} = 0) \ge 1 - p^{-3}$. We thus get $\mathbb{E}_w\|\widehat{\theta}^0\|_0 \le \mathbb{E}_w\|\widehat{\theta}^0_{S^c}\|_0 + s_0 \le p\cdot p^{-3} + s_0 \le 1 + s_0$. Since $\mathbb{P}\{\widehat{\theta}^0_{S^c} = 0\} \ge 1 - p^{-3}$, in order to prove Eq. (152), it is sufficient to develop a tail bound on $|\|\widehat{\theta}^0_S\|_0 - \mathbb{E}_w\{\|\widehat{\theta}^0\|_0\}$, which we do via Chebyshev inequality. Letting $T_i \equiv \mathbb{I}(|\theta^*_i + n^{-1}(X^\mathsf{T}w)_i| > \lambda)$, we have $\|\widehat{\theta}^0_S\|_0 = \sum_{i\in S} T_i$, whence the variance of $\|\widehat{\theta}^0_S\|_0$ is given by

$$
\begin{aligned}
\mathrm{Var}_w(\|\widehat{\theta}_S\|_0) &= \sum_{i,j\in S} \mathrm{Cov}_w(T_i;T_j) \\
&\overset{(a)}{\le} \sum_{i,j\in S} \frac{\mathrm{Cov}_w((X^\mathsf{T}w)_i;(X^\mathsf{T}w)_j)}{\sqrt{\mathrm{Var}_w((X^\mathsf{T}w)_i)\mathrm{Var}_w((X^\mathsf{T}w)_j)}} \cdot \sqrt{\mathrm{Var}(T_i)\mathrm{Var}(T_j)} \\
&= \sum_{i,j\in S} \frac{\langle \tilde{x}_i, \tilde{x}_j\rangle}{\|\tilde{x}_i\|_2\|\tilde{x}_j\|_2} \sqrt{\mathrm{Var}(T_i)\mathrm{Var}(T_j)} \\
&\le M\sqrt{\frac{\log p}{n}} \left(\sum_{i\in S}\sqrt{\mathrm{Var}(T_i)}\right)^2 \\
&\le M s_0^2 \sqrt{\frac{\log p}{n}}\,. \qquad (153)
\end{aligned}
$$

Here $(a)$ follows because, for jointly Gaussian random variables $Z_1$, $Z_2$, the correlation coefficient between $f(Z_1)$, $g(Z_2)$ is maximized by linear functions $f$, $g$.

The claim (152) follows from Chebyshev inequality, using Eq. (153). □

LEMMA H.6. *Let $L_n$ be any sequence with $L_n \to \infty$. Then, under the assumptions of Theorem 5.4, we have, with high probability,*

$$
\left|\|\widehat{\theta}\|_0 - \|\widehat{\theta}^0\|_0\right| \le L_n s_0 \sqrt{\frac{s_0(\log p)^2}{n}}\,. \qquad (154)
$$

PROOF. Recall that, by definition

$$
\widehat{\theta} = \eta\Big(\theta^* + \frac{1}{n}X^\mathsf{T}w + \frac{1}{\sqrt{n}}R;\lambda\Big)\,, \qquad (155)
$$

$$
\widehat{\theta}^0 = \eta\Big(\theta^* + \frac{1}{n}X^\mathsf{T}w;\lambda\Big)\,. \qquad (156)
$$

Let $\varepsilon_n = C\sqrt{s_0(\log p)^2/n}$ for $C$ a sufficiently large constant, and define the event

$$
\mathcal{G}_0 \equiv \Big\{\|R\|_\infty \le \sigma\varepsilon_n;\ \ \overline{S} \subseteq S\Big\}\,. \qquad (157)
$$

By Theorem 3.8 and Lemma H.1, $\mathbb{P}(\mathcal{G}_0) \to 1$ as $n, p \to \infty$. On this event, we have

$$\begin{aligned}\Big|\|\widehat{\theta}\|_0 - \|\widehat{\theta}^0\|_0\Big| &\le \sum_{i\in S} \mathbb{I}\left(\Big|\theta_i^* + \frac{1}{n}(X^\mathsf{T} w)_i\Big| \in \Big[\lambda - \frac{1}{\sqrt{n}}\|R\|_\infty, \lambda + \frac{1}{\sqrt{n}}\|R\|_\infty\Big]\right)\\ &\le \sum_{i\in S} \mathbb{I}\left(\Big|\theta_i^* + \frac{1}{n}(X^\mathsf{T} w)_i\Big| \in \Big[\lambda - \frac{\sigma\varepsilon_n}{\sqrt{n}}, \lambda + \frac{\sigma\varepsilon_n}{\sqrt{n}}\Big]\right)\\ &\equiv \sum_{i\in S} W_i\,.\end{aligned}$$

We then have, for any sequence $L_n \to \infty$,

$$\mathbb{P}\Big(\Big|\|\widehat{\theta}\|_0 - \|\widehat{\theta}^0\|_0\Big| \ge L_n s_0 \varepsilon_n\Big) \le \mathbb{P}\Big(\sum_{i\in S} W_i \ge L_n s_0\varepsilon_n; \mathcal{G}_1(M)\Big) + \mathbb{P}(\mathcal{G}_0^c) + \mathbb{P}(\mathcal{G}_1(M)^c)\,.$$

Using Lemma H.1, it is sufficient to show that the first term vanishes. This can be done by Markov inequality, bounding the expectation as follows

$$\begin{aligned}\mathbb{E}\Big\{\sum_{i\in S} W_i; \mathcal{G}_1\Big\} &= \sum_{i\in S}\mathbb{P}\left(\Big|\theta_i^* + \frac{1}{n}(X^\mathsf{T} w)_i\Big| \in \Big[\lambda - \frac{\sigma\varepsilon_n}{\sqrt{n}}, \lambda + \frac{\sigma\varepsilon_n}{\sqrt{n}}\Big]; \mathcal{G}_1(M)\right)\\ &\overset{(a)}{\le} 2\sum_{i\in S}\sup_{z\in\mathbb{R}}\mathbb{P}\left(\frac{\sigma\|\tilde{x}_i\|_2}{n} Z \in \Big[z - \frac{\sigma\varepsilon_n}{\sqrt{n}}, z + \frac{\sigma\varepsilon_n}{\sqrt{n}}\Big]; \mathcal{G}_1(M)\right)\\ &\overset{(b)}{\le} 2s_0 \sup_{z\in\mathbb{R}}\mathbb{P}\left(Z \in [z - 2\varepsilon_n, z + 2\varepsilon_n]\right)\\ &\le C\, s_0\varepsilon_n\,,\end{aligned}$$

where $(a)$ holds for $Z \sim \mathsf{N}(0,1)$, and $X \in \mathcal{G}_1(M)$ was used in $(b)$. □

PROOF OF THEOREM 5.4. First notice that

$$\begin{aligned}\Big|\Delta_1 - \frac{2\sigma^2}{n}\|\widehat{\theta}\|_0\Big| &\le \Big|\Delta_1 - \frac{2\sigma^2}{n}\|\widehat{\theta}^0\|_0\Big| + \frac{2\sigma^2}{n}\Big|\|\widehat{\theta}\|_0 - \|\widehat{\theta}^0\|_0\Big|\\ &\le \Big|\mathbb{E}_w\Delta_1 - \frac{2\sigma^2}{n}\mathbb{E}_w\|\widehat{\theta}^0\|_0\Big| + \Big|\Delta_1 - \mathbb{E}_w\Delta_1\Big|\\ &\quad + \frac{2\sigma^2}{n}\Big|\|\widehat{\theta}^0\|_0 - \mathbb{E}_w\|\widehat{\theta}^0\|_0\Big| + \frac{2\sigma^2}{n}\Big|\|\widehat{\theta}\|_0 - \|\widehat{\theta}^0\|_0\Big|\\ &\overset{(a)}{\le} 2Cs_0\sigma^2\sqrt{\frac{\log p}{n^3}} + \frac{2t\sigma^2}{\sqrt{n}} + +2L_n s_0\sigma^2\frac{(\log p)^{1/4}}{n^{5/4}} + 2L_n\sigma^2\Big(\frac{s_0}{n}\Big)^{3/2}\log p\\ &\le \frac{2t\sigma^2}{\sqrt{n}} + \frac{6L_n s_0\sigma^2}{n}\left(\Big(\frac{\log p}{n}\Big)^{1/4} \vee \Big(\frac{s_0(\log p)^2}{n}\Big)^{1/2}\right),\end{aligned}$$

where the inequality $(a)$ holds probability larger than $1 - o_n(1) - 2e^{-ct^2}$ by lemmas H.3, H.4, H.5, H.6 for any sequence $L_n \to \infty$ as $n \to \infty$. We let

$$\varepsilon_n \equiv 6L_n \left( \left(\frac{\log p}{n}\right)^{1/4} \vee \left(\frac{s_0(\log p)^2}{n}\right)^{1/2} \right). \tag{158}$$

Using the decomposition (140), we have

$$\begin{aligned}\left|\mathsf{R}(y, X, \theta^*) - \widehat{\mathsf{R}}(y, X) - \frac{2\sigma^2}{n}\|\widehat{\theta}\|_0\right| &\le \left|\Delta_1 - \frac{2\sigma^2}{n}\|\widehat{\theta}\|_0\right| + |\Delta_2| \\ &\le \frac{2t\sigma^2}{\sqrt{n}} + \frac{\varepsilon_n s_0 \sigma^2}{n} + \frac{C s_0 \sigma^2}{n}\sqrt{\frac{s_0(\log p)^2}{n}} \\ &\le \frac{2t\sigma^2}{\sqrt{n}} + \frac{2\varepsilon_n s_0\sigma^2}{n},\end{aligned}$$

where the last inequality holds for all $n$ large enough.

By choosing $L_n$ to be a sequence with slow enough growth rate, e.g. $L_n = (\frac{n}{s_0(\log p)^2})^{1/4}$, we have $\varepsilon_n \to 0$. This completes the proof for Gaussian designs. □

## APPENDIX I: PROOF OF LEMMA 6.3 (PERTURBATION BOUND)

LEMMA I.1. *For all $\theta \in \mathbb{R}^{p-1}$ the following holds true.*

$$\frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2 \le \mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\rm p}_{\sim i}). \tag{159}$$

Lemma I.1 is proved in Section I.1.

LEMMA I.2. *Let $f_k(x) = \frac{c}{2}(x - a - u_k)^2 + \lambda|x| + b_k$ for $k = 1, 2$. Further assume that $\min_x f_1(x) \le \min_x f_2(x)$. Then,*

$$f_1(a) - f_2(a) \le (c|u_2| + \lambda)|u_1 - u_2| + \frac{c}{2}(u_1 - u_2)^2. \tag{160}$$

Lemma I.2 is proved in Section I.2.

LEMMA I.3. *For $\theta \in \mathbb{R}^{p-1}$ define*

$$u(\theta) \equiv \frac{\tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta^*_{\sim i} - \theta))}{\|\tilde{x}_i\|_2^2} \tag{161}$$

*Also let $c_i \equiv \|\tilde{x}_i\|_2^2/n$. Then, the following relation holds true.*

$$\mathcal{L}(\theta_i, \theta) = \lambda|\theta_i| + \frac{c_i}{2}(\theta_i - \theta_i^* - u(\theta))^2 - \frac{c_i}{2}u(\theta)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*|. \tag{162}$$

Lemma I.3 is proved in Section I.3.

We let $f_1(x) = \mathcal{L}(x, \widehat{\theta}_{\sim i})$ and $f_2(x) = \mathcal{L}(x, \widehat{\theta}^{\rm p}_{\sim i})$. Note that $(\widehat{\theta}_i, \widehat{\theta}_{\sim i})$ is the minimizer of $\mathcal{L}_{y,X}(\theta_i, \theta_{\sim i})$. Therefore, $\min f_1(x) = \mathcal{L}(\widehat{\theta}_i, \widehat{\theta}_{\sim i}) \le \min f_2(x)$. Using decomposition (162) and applying Lemma I.2 with

$$\begin{aligned}
&c = c_i, \quad a = \theta^*_i, \quad u_1 = u(\widehat{\theta}_{\sim i}), \quad u_2 = u(\widehat{\theta}^{\rm p}_{\sim i}),\\
&b_1 = -\frac{c_i}{2} u(\widehat{\theta}_{\sim i})^2 + \mathcal{L}(\theta^*_i, \widehat{\theta}_{\sim i}) - \lambda|\theta^*_i|,\\
&b_2 = -\frac{c_i}{2} u(\widehat{\theta}^{\rm p}_{\sim i})^2 + \mathcal{L}(\theta^*_i, \widehat{\theta}^{\rm p}_{\sim i}) - \lambda|\theta^*_i|\,,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\mathcal{L}(\theta^*_i, \widehat{\theta}_{\sim i}) - \mathcal{L}(\theta^*_i, \widehat{\theta}^{\rm p}_{\sim i}) \le \\
&\Big(c_i|u(\widehat{\theta}^{\rm p}_{\sim i})| + \lambda\Big)\left|u(\widehat{\theta}_{\sim i}) - u(\widehat{\theta}^{\rm p}_{\sim i})\right| + \frac{c_i}{2}\left(u(\widehat{\theta}_{\sim i}) - u(\widehat{\theta}^{\rm p}_{\sim i})\right)^2 . \qquad (163)
\end{aligned}$$

We next write

$$\begin{aligned}
\frac{c_i}{2}(u(\widehat{\theta}_{\sim i}) - u(\widehat{\theta}^{\rm p}_{\sim i}))^2 &= \frac{1}{2n}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})^{\sf T} X^{\sf T}_{\sim i}\tilde{x}_i\tilde{x}_i^{\sf T} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\\
&= \frac{1}{2n}\|{\rm P}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2\,, \qquad (164)
\end{aligned}$$

where ${\rm P}_{\tilde{x}_i} \equiv \tilde{x}_i\tilde{x}_i^{\sf T}/\|\tilde{x}_i\|_2^2$ denotes the projection on the direction of $\tilde{x}_i$.

We lower bond the left-hand side of Equation (163) using Lemma I.1 and employing the above identity to get

$$\frac{1}{2n}\|{\rm P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2 \le (c_i|u(\widehat{\theta}^{\rm p}_{\sim i})| + \lambda)|u(\widehat{\theta}_{\sim i}) - u(\widehat{\theta}^{\rm p}_{\sim i})|\,. \qquad (165)$$

Next preposition bounds $c_i|u(\widehat{\theta}^{\rm p}_{\sim i})|$. We defer the proof of Proposition I.4 to Section I.4.

PROPOSITION I.4. *Let $\mathcal{B} \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n, s_0, 3)$, where the events $\mathcal{B}_\delta(n, s_0, 3)$ and $\tilde{\mathcal{B}}(n,p)$ are given by Lemma 3.2 and Equation (17). The following holds true.*

$$\mathbb{P}\Big(|c_i u(\widehat{\theta}^{\rm p}_{\sim i})| \ge 1.25\rho\lambda; \mathcal{B}\Big) \le 2\exp\Big(-\frac{c_* n}{s_0}\Big)\,.$$

*where $c_* \equiv (1-\delta)^2 C_{\min}/8$.*

We further have

$$|u(\widehat{\theta}_{\sim i}) - u(\widehat{\theta}^{\rm p}_{\sim i})| = \frac{|\tilde{x}_i X_{\sim i}(\widehat{\theta}^{\rm p}_{\sim i} - \widehat{\theta}_{\sim i})|}{\|\tilde{x}_i\|_2^2} \le \frac{\|X_{\sim i}(\widehat{\theta}^{\rm p}_{\sim i} - \widehat{\theta}_{\sim i})\|_2}{\|\tilde{x}_i\|_2}\,. \qquad (166)$$

We next upper bound the term $\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2$.

The corollary below follows from Proposition 3.5 and its proof is given in Section I.5.

COROLLARY I.5. *Set $\lambda = \kappa\sigma\sqrt{(\log p)/n}$, for a constant $\kappa \ge 8$. On the event $\mathcal{B}(C_*) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}_\delta(n,(C_*+1)s_0,3)$, the following holds.*

$$\frac{1}{n}\|X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2 \le (1+\delta)^2 C_{\max}\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}\|_2^2\,. \tag{167}$$

We next lower bound $\|\tilde{x}_i\|_2$. Observe that the entries $\tilde{x}_{i\ell}^2 - 1$, $\ell \in [n]$, are zero-mean sub-exponential random variables. We obtain the following tail-bound inequality by applying Bernstein-type inequality for sub-exponential random variables. (See e.g. [41, Equation (190)].)

$$\mathbb{P}\Big(\|\tilde{x}_i\|_2 > \frac{\sqrt{n}}{5}\Big) \le e^{-n/1000}\,. \tag{168}$$

Combining the results of Proposition (I.4) and equations (167) and (168), we obtain that on event $\mathcal{B}$, with probability at least $1 - e^{-n/1000} - 2e^{-c_* n/s_0}$, the following holds:

$$\frac{1}{2n}\|{\rm P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2 \le 12\rho(1+\delta)\sqrt{C_{\max}}\lambda\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}\|_2\,. \tag{169}$$

The last step is to lower bound the left-hand side of Equation (169). Write

$$\begin{aligned}{\rm P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}) &= X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}) - {\rm P}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\\ &= X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}) - \tilde{x}_i\Big\langle \frac{\tilde{x}_i}{\|\tilde{x}_i\|_2^2}, X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\Big\rangle\,.\end{aligned}$$

Define vector $\mu \in \mathbb{R}^p$ with

$$\mu_i \equiv -\Big\langle \frac{\tilde{x}_i}{\|\tilde{x}_i\|_2^2}, X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\Big\rangle\,, \qquad \mu_{\sim i} = \widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}\,.$$

Then $\mu \in \mathcal{C}(C_\delta s_0, 3)$, by Proposition 3.5, with $C_\delta = C_* + 1$ . Hence, on the event $\mathcal{B}(n, C_\delta s_0, 3)$, we have

$$\begin{aligned}\frac{1}{2n}\|{\rm P}^{\perp}_{\tilde{x}_i} X_{\sim i}(\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\|_2^2 &= \frac{1}{2n}\|X\mu\|_2^2\\ &\ge \frac{1}{2}(1-\delta)^2 C_{\min}\|\mu\|_2^2\\ &\ge \frac{1}{2}(1-\delta)^2 C_{\min}\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}\|_2^2\,. \end{aligned} \tag{170}$$

Finally, note that $\mathcal{C}(s_0,3) \subseteq \mathcal{C}(C_\delta s_0, 3)$, since $C_\delta \geq 1$. Therefore, $\mathcal{B}_\delta(n, C_\delta s_0, 3) \subseteq \mathcal{B}_\delta(n, s_0, 3)$, by definition. Letting $\mathcal{B}(C_\delta) \equiv \tilde{\mathcal{B}}(n,p) \cap \mathcal{B}(n, C_\delta s_0, 3)$, we have $\mathcal{B}(C_\delta) \subseteq \mathcal{B}$. Combining equations (169) and (170), we obtain

$$\mathbb{P}\Big(\|\widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}\|_2 \geq \frac{24\rho(1+\delta)\sqrt{C_{\max}}}{(1-\delta)^2 C_{\min}}\lambda\,; \mathcal{B}(C_\delta)\Big) \leq 2\exp\Big(-\frac{c_* n}{s_0}\Big) + \exp\Big(-\frac{n}{1000}\Big)\,.$$

This completes the proof.

**I.1. Proof of Lemma I.1.** For $\theta$ we have

$$\mathcal{L}_{y,X}(\theta_i^*, \theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta_i^* - X_{\sim i}\theta\|^2 + \lambda\|\theta\|_1 + \lambda|\theta_i^*|$$

Let $\tilde{y} \equiv y - \tilde{x}_i\theta_i^*$. We then have

$$\begin{aligned}\mathcal{L}_{y,X}(\theta_i^*, \theta) &= \frac{1}{2n}\|\tilde{y} - X_{\sim i}\widehat{\theta}^{\rm p}_{\sim i} - X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\|^2 + \lambda\|\theta\|_1 + \lambda|\theta_i^*| \\ &= \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\rm p}_{\sim i}) + \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\|^2 - \frac{1}{n}\langle \tilde{y} - X_{\sim i}\widehat{\theta}^{\rm p}_{\sim i}, X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\rangle \\ &\quad + \lambda\|\theta\|_1 - \lambda\|\widehat{\theta}^{\rm p}_{\sim i}\|_1 \end{aligned} \tag{171}$$

Since $\widehat{\theta}^{\rm p}_{\sim i}$ is the minimizer of $\mathcal{L}_{y,X}(\theta_i^*, \theta)$ by KKT condition we have

$$\frac{1}{n}X_{\sim i}^{\mathsf{T}}(\tilde{y} - X_{\sim i}\widehat{\theta}^{\rm p}_{\sim i}) = \lambda\xi, \qquad \xi \in \partial\|\widehat{\theta}^{\rm p}_{\sim i}\|_1\,. \tag{172}$$

Applying equation (172) in equation (171) we get

$$\begin{aligned}\mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}_{y,X}(\theta_i^*, \widehat{\theta}^{\rm p}_{\sim i}) &= \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\|^2 + \lambda\Big(\|\theta\|_1 - \|\widehat{\theta}^{\rm p}_{\sim i}\|_1 - \langle \xi, \theta - \widehat{\theta}^{\rm p}_{\sim i}\rangle\Big) \\ &\geq \frac{1}{2n}\|X_{\sim i}(\theta - \widehat{\theta}^{\rm p}_{\sim i})\|^2\,,\end{aligned}$$

where the last step follows from the definition of a subgradient.

**I.2. Proof of Lemma I.2.** Define $x_{{\rm opt},1} = \arg\min_x f_1(x)$. It is simple to see that $x_{{\rm opt},1} = \eta(a + u_1; \lambda/c_i)$, where $\eta(x;\alpha)$ is the soft-thresholding function given by

$$\eta(x;\alpha) = \begin{cases} x - \alpha & x \geq \alpha\,, \\ 0 & |x| \leq \alpha\,, \\ x + \alpha & x \leq -\alpha\,. \end{cases}$$

By substituting for $x_{\rm opt,1}$ in equation (162) and after some algebraic manipulations, we obtain

$$f_1(x_{\rm opt,1}) = c_i \mathcal{H}(a+u_1; \lambda/c_i) + b_1\,,$$

where $\mathcal{H}(x;\alpha)$ is the Huber function:

$$\mathcal{H}(x;\alpha) = \begin{cases} \alpha|x| - \frac{\alpha^2}{2} & \text{if } |x| > \alpha\,,\\ \\ \frac{x^2}{2} & \text{if } |x| \le \alpha\,. \end{cases}$$

Similarly, setting $x_{\rm opt,2} = \arg\min_x f_2(x)$ we have

$$f_1(x_{\rm opt,2}) = c_i \mathcal{H}(a+u_2; \lambda/c_i) + b_2\,.$$

Define $\Delta_1 \equiv f_1(a) - f(x_{\rm opt,1})$ and $\Delta_2 \equiv f_2(a) - f(x_{\rm opt,2})$. Substituting for $f_1(a)$ and $f_2(a)$, we get

$$\Delta_1 = c_i \frac{u_1^2}{2} + \lambda|a| - c_i\mathcal{H}(a+u_1;\lambda/c_i)\,, \tag{173}$$

$$\Delta_2 = c_i \frac{u_2^2}{2} + \lambda|a| - c_i\mathcal{H}(a+u_2;\lambda/c_i)\,. \tag{174}$$

We then write

$$f_1(a) - f_2(a) = \Delta_1 - \Delta_2 + f_1(x_{\rm opt,1}) - f_2(x_{\rm opt,2}) \le \Delta_1 - \Delta_2\,, \tag{175}$$

where we use the assumption $\min_x f_1(x) \le \min_x f_2(x)$.

Finally we bound $\Delta_1 - \Delta_2$ as follows:

$$\begin{aligned}\Delta_1 - \Delta_2 &= c_i\frac{u_1^2 - u_2^2}{2} + c_i\Big\{\mathcal{H}(a+u_2;\lambda/c_i) - \mathcal{H}(a+u_1;\lambda/c_i)\Big\}\\ &\le c_i u_2(u_1-u_2) + \frac{(u_1-u_2)^2}{2} + \lambda|u_1-u_2|\end{aligned}$$

where the last inequality holds since $\mathcal{H}'(x;\alpha) = x - \eta(x;\alpha)$ and hence $|\mathcal{H}'(x;\alpha)| \le \alpha$ and due to the mean-value theorem.

**I.3. Proof of Lemma I.3.** To lighten the notation, we drop the subscripts $y, X$ in $\mathcal{L}_{y,X}(\cdot)$. Recall that $\Delta(\theta) \equiv \mathcal{L}_{y,X}(\theta_i^*, \theta) - \mathcal{L}^+(\theta)$. We start by expanding $\mathcal{L}(\theta_i, \theta)$.

$$\mathcal{L}(\theta_i,\theta) = \frac{1}{2n}\|y - \tilde{x}_i\theta_i - X_{\sim i}\theta\|_2^2 + \lambda|\theta_i^*| + \lambda\|\theta\|_1\,.$$

Plugging in $y = \tilde{x}_i\theta_i^* + X_{\sim i}\theta_{\sim i}^* + w$ and rearranging the terms, we obtain

$$\mathcal{L}(\theta_i, \theta) = \frac{1}{2n}\|w + X_{\sim i}(\theta_{\sim i}^* - \theta)\|_2^2 + \frac{1}{n}\langle \theta_i^* - \theta_i, \tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\rangle$$
$$+ \frac{1}{2n}\|\tilde{x}_i\|^2(\theta_i^* - \theta_i)^2 + \lambda|\theta_i| + \lambda\|\theta\|_1 . \tag{176}$$

Therefore,

$$\mathcal{L}(\theta_i^*, \theta) = \frac{1}{2n}\|w + X_{\sim i}(\theta_{\sim i}^* - \theta)\|_2^2 + \lambda|\theta_i^*| + \lambda\|\theta\|_1 . \tag{177}$$

Combining equations (176) and (177), we rewrite $\mathcal{L}(\theta_i, \theta)$ as

$$\mathcal{L}(\theta_i, \theta) = \mathcal{L}(\theta_i^*, \theta) + \frac{1}{n}\langle \theta_i^* - \theta_i, \tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\rangle$$
$$+ \frac{1}{2n}\|\tilde{x}_i\|^2(\theta_i^* - \theta_i)^2 + \lambda|\theta_i| - \lambda|\theta_i^*|$$
$$= \lambda|\theta_i| + \frac{1}{2n}\|\tilde{x}_i\|^2\Big(\theta_i - \theta_i^* - \frac{\tilde{x}_i^{\mathsf{T}}}{\|\tilde{x}_i\|^2}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\Big)^2$$
$$- \frac{1}{2n\|\tilde{x}_i\|^2}\Big(\tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta_{\sim i}^* - \theta))\Big)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*| . \tag{178}$$

Writing expression (178) in terms of $c_i \equiv \|\tilde{x}_i\|^2/n$ and $u(\theta)$, given by (161), we get

$$\mathcal{L}(\theta_i, \theta) = \lambda|\theta_i| + \frac{c_i}{2}(\theta_i - \theta_i^* - u(\theta))^2 - \frac{c_i}{2}u(\theta)^2 + \mathcal{L}(\theta_i^*, \theta) - \lambda|\theta_i^*| . \tag{179}$$

The result follows.

**I.4. Proof of Preposition I.4.** Let $T = \text{supp}(\widehat{\theta}^{\text{p}}_{\sim i}) \cup \text{supp}(\theta_*)$. By Lemma 3.5, $|T| < C_\delta s_0$, where

$$C_\delta \equiv C_* + 1 = \frac{16}{(1-\delta)^2}\frac{C_{\max}}{C_{\min}} + 1 .$$

For $i \in [p]$ define

$$\Sigma_{i|T} \equiv \Sigma_{i,i} - \Sigma_{i,T}(\Sigma_{T,T})^{-1}\Sigma_{T,i} .$$

Since $\tilde{x}_i$ and $X_T$ are jointly Gaussian, we have

$$\tilde{x}_i = X_T(\Sigma_{T,T})^{-1}\Sigma_{T,i} + \Sigma_{i|T}^{1/2} z , \tag{180}$$

where $z \in \mathbb{R}^n$ is independent of $X_T$ with i.i.d standard normal coordinates.

Recalling the definition of $c_i \equiv \|\tilde{x}_i\|^2/n$ and $u(\theta)$, given by equation (161), we write $c_i|u(\widehat{\theta}^{\rm p}_{\sim i})|$ as

$$
\begin{aligned}
c_i|u(\widehat{\theta}^{\rm p}_{\sim i})| =& \frac{1}{n}\Big|\tilde{x}_i^{\mathsf{T}}(w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}))\Big| \\
=& \frac{1}{n}\Big|\tilde{x}_i^{\mathsf{T}}(w + X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T))\Big| \\
\le& \frac{1}{n}|\tilde{x}_i^{\mathsf{T}} w| + \frac{1}{n}\Sigma_{i|T}^{1/2}\Big|z^{\mathsf{T}} X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)\Big| + \frac{1}{n}\Big|\Sigma_{i,T}(\Sigma_{T,T})^{-1}X_T^{\mathsf{T}} X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)\Big| \\
\le& \frac{1}{n}|\tilde{x}_i^{\mathsf{T}} w| + \frac{1}{n}\Sigma_{i|J}^{1/2}\Big|z^{\mathsf{T}} X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)\Big| + \frac{1}{n}\|\Sigma_{i,T}(\Sigma_{T,T})^{-1}\|_1\|X_T^{\mathsf{T}} X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)\|_\infty\,.
\end{aligned} \tag{181}
$$

The first inequality here follows from equation (180).

In the following we bound each term on the RHS of equation (181) individually.

On the event $\tilde{\mathcal{B}}(n,p)$, defined by equation (17), we have

$$
\frac{1}{n}\|\tilde{x}_i^{\mathsf{T}} w\| \le \frac{1}{n}\|X^{\mathsf{T}} w\|_\infty \le 2\sigma\sqrt{\frac{\log p}{n}} \le \frac{\lambda}{4}\,. \tag{182}
$$

We use Corollary 6.2 to bound the second term of expression (181). We recall the event $\mathcal{B}_\delta(n,s_0,3)$, given by Lemma 3.2 and let $\mathcal{B} \equiv \mathcal{B}_\delta(n,s_0,3) \cap \tilde{\mathcal{B}}(n,p)$. Further, recall the notation $\zeta_i \equiv X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})/\sqrt{n} = X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)/\sqrt{n}$ and the event $\mathcal{E}_i$ defined by equation (63). We write

$$
\begin{aligned}
\mathbb{P}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|T}^{1/2}\;|z^{\mathsf{T}}\zeta_i| \ge \lambda; \mathcal{B}\Big) &\le \mathbb{P}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|T}^{1/2}\;|z^{\mathsf{T}}\zeta_i| \ge \lambda; \mathcal{E}_i\Big) \\
&= \mathbb{E}\Big\{\mathbb{I}\Big(\frac{1}{\sqrt{n}}\Sigma_{i|T}^{1/2}\;|z^{\mathsf{T}}\zeta_i| \ge \lambda\Big)\cdot\mathbb{I}(\mathcal{E}_i)\Big\} \\
&\overset{(a)}{\le} 2\mathbb{E}\Big(\exp\Big[-\frac{n\lambda^2}{2\|\zeta_i\|^2}\Big]\cdot\mathbb{I}(\mathcal{E}_i)\Big) \\
&\overset{(b)}{\le} 2\exp\Big(-c_*\frac{n}{s_0}\Big)\,,
\end{aligned} \tag{183}
$$

with $c_* \equiv (1-\delta)^2 C_{\min}/8$. Here, the penultimate inequality follows from Fubini's theorem where we first integrate w.r.t $z$ and then w.r.t $\zeta_i$. Note that $z$ and $\zeta_i$ are independent. Therefore, $z^{\mathsf{T}}\zeta_i|\zeta_i \sim \mathsf{N}(0,\|\zeta_i\|^2)$. In Step $(a)$, we use the fact that $(\Sigma_{T,T})^{-1} \succeq 0$ and hence $\Sigma_{i|T} \le \Sigma_{ii} \le 1$ by Condition *(i)*. Step $(b)$ follows from Corollary 6.2.

We next bound the third term on the RHS of equation (181). Note that the KKT conditions for optimization (56) reads

$$\frac{1}{n} X_{\sim i}^{\mathsf{T}} (w + X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i})) = \lambda \xi \,, \tag{184}$$

for $\xi \in \partial \|\widehat{\theta}^{\rm p}_{\sim i}\|_1$. Since $\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}$ is supported on $T$, we have $X_{\sim i}(\theta^*_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}) = X_T(\theta^*_T - \widehat{\theta}^{\rm p}_T)$. To lighten the notation, let

$$\nu \equiv \frac{1}{n} X_T^{\mathsf{T}} X_T (\theta^*_T - \widehat{\theta}^{\rm p}_T) \,.$$

We know by equation (184),

$$\|\nu\|_\infty \le \frac{1}{n} \|X_T^{\mathsf{T}} w\|_\infty + \lambda \|\xi_T\|_\infty \,.$$

On the event $\tilde{\mathcal{B}}(n,p)$ we have

$$\frac{1}{n} \|X_T^{\mathsf{T}} w\|_\infty \le 2\sigma \sqrt{\frac{\log p}{n}} \le \frac{\lambda}{4} \,.$$

Combining the above two inequalities we obtain

$$\|\nu\|_\infty \le 5\lambda/4 \,. \tag{185}$$

We next employ Condition *(iii)* to bound $\|\Sigma_{i,T}(\Sigma_{T,T})^{-1}\|_1$. Define $\tilde{T} = T \cup \{i\}$ and write the inverse of $\Sigma_{\tilde{T},\tilde{T}}$ using Schur complement:

$$\Sigma^{-1}_{\tilde{T},\tilde{T}} = \begin{pmatrix} \Sigma^{-1}_{i|T} & -\Sigma^{-1}_{i|T} \Sigma_{i,T} \Sigma^{-1}_{T,T} \\ -\Sigma^{-1}_{T,T} \Sigma_{T,i} \Sigma^{-1}_{i|T} & \Sigma^{-1}_{T,T} + \Sigma^{-1}_{T,T} \Sigma_{T,i} \Sigma^{-1}_{i|T} \Sigma_{i,T} \Sigma^{-1}_{T,T} \end{pmatrix} .$$

By Condition *(iii)* and as $|\tilde{T}| \le C_\delta s_0$, $\|\Sigma^{-1}_{\tilde{T},\tilde{T}} e_i\|_1 \le \rho$. Further, by Condition *(i)*, $\Sigma_{i|T} \le \Sigma_{i,i} \le 1$. Hence, we get

$$\rho \ge \|\Sigma^{-1}_{\tilde{T},\tilde{T}} e_i\|_1 \ge 1 + \|\Sigma_{i,T}(\Sigma_{T,T})^{-1}\|_1 \,. \tag{186}$$

Using equations (182) to (186), we bound the RHS of equation (181) as follows. Under the event $\mathcal{B}$,

$$c_i |u(\widehat{\theta}^{\rm p}_{\sim i})| \le \frac{5\lambda}{4} \rho \,.$$

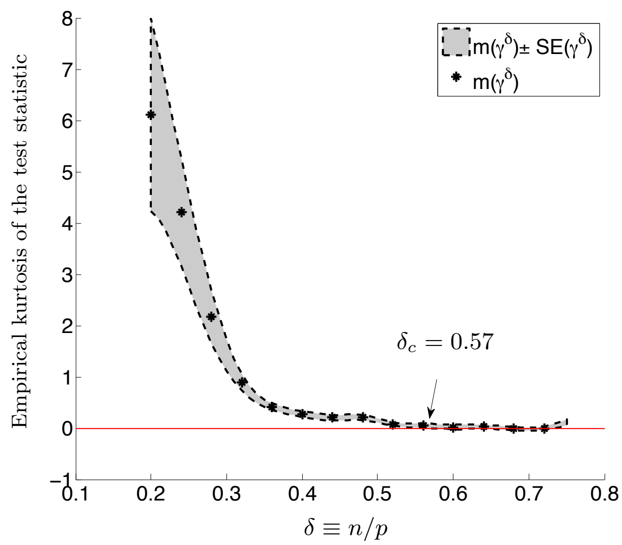



Fig 2: Empirical kurtosis of the (rescaled) debiased Lasso estimator $T_i = \sqrt{n}(\widehat{\theta}_i^{\rm d} - \theta_i^*)/(\sigma[M\widehat{\Sigma}M]_{i,i}^{1/2})$. We plot the kurtosis $m(\gamma^\delta)$ (over coordinates and 100 independent realizations) versus $\delta$ along with the upper and lower one standard error curves, as a function of the number of samples per parameter $\delta$. Here, $\varepsilon = 0.2$ and $\delta_c = 0.57$ is our empirical estimate for the number of samples above which the debiased estimator is approximately Gaussian.

**I.5. Proof of Corollary I.5.** Note that $\widehat{\theta}^{\rm p}_{\sim i}$ is the Lasso estimators corresponding to $(\tilde{y}, X_{\sim i})$, according to equation (56). As a corollary of Proposition 3.5, on event $\mathcal{B}$, $\|\widehat{\theta}^{\rm p}_{\sim i}\|_0 \le C_* s_0$, with $C_* \equiv (16 C_{\max}/C_{\min})(1-\delta)^{-2}$. Also, $\|\widehat{\theta}_{\sim i}\|_0 \le s_0$. Therefore,$(0, \widehat{\theta}_{\sim i} - \widehat{\theta}^{\rm p}_{\sim i}) \in \mathcal{C}((C_*+1)s_0, 3)$ and, by definition, on event $\mathcal{B}_\delta(n, (C_*+1)s_0, 3)$, the claim holds true.

## APPENDIX J: NUMERICAL ILLUSTRATION

Our goal in this section is to numerically corroborate the results of Theorem 3.8 and Theorem 3.13. More specifically, we would like to check whether the debiased estimator exhibits an unbiased Gaussian distribution provided that the sample size scales linearly with the number of nonzero parameters.

We generate data from linear model (1) with the following configuration. We fix $p = 3000$ and consider regression parameter $\theta_0$ with support $S_0$ chosen uniformly at random from the index set $[p]$ and $\theta_i^* = 0.15$ for $i \in S_0$ and zero otherwise. The design matrix $X$ has i.i.d. rows drawn from $\mathsf{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is the circulant matrix with entries $\Sigma_{i,j} = 0.8^{|i-j|}$. The measurement noise $w$ has i.i.d. standard normal entries.

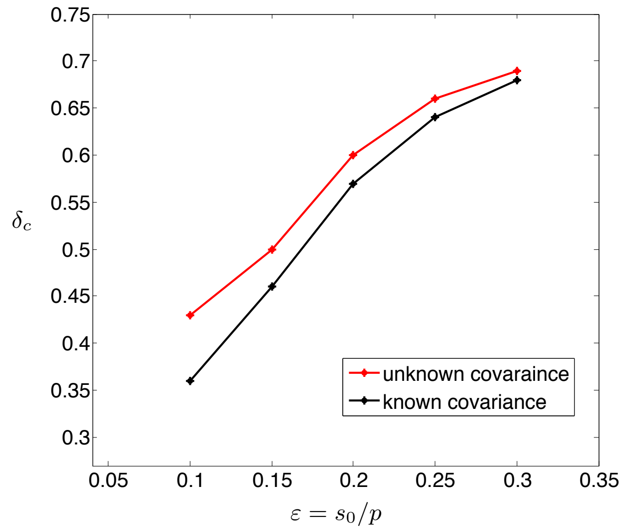



Fig 3: Critical number of samples per coordinate $\delta_c$, versus fraction of non-zero coordinates $\varepsilon$. For $\delta > \delta_c(\varepsilon)$ the debiased Lasso estimator is empirically Gaussian distributed in our experiment. The approximately linear relationship at small $\varepsilon$ is in agreement with our theory.

Let $s_0 = |S_0|$ and $\varepsilon = s_0/p$ be the sparsity level and $\delta = n/p$ denote the under sampling rate. We vary $\varepsilon$ in the set $\{0.1, 0.15, 0.2, 0.25, 0.3\}$ and for each value of $\varepsilon$ we compute critical value of $\delta$ above which the unbiased estimator admits a Gaussian distribution. We will denote this critical value as $\delta_c$ and define it as follows. We vary $\delta$ and for each pair $(\varepsilon, \delta)$, compute the debiased estimator (with $M = \Sigma^{-1}$) for 100 realizations of noise $w$. We then compute the empirical kurtosis of each coordinate $T_i = \sqrt{n}(\widehat{\theta}_i^{\rm d} - \theta_i^*)/(\sigma[M\widehat{\Sigma}M]_{i,i}^{1/2})$. For $i \in [p]$, let $\gamma_i^\delta$ denote the empirical kurtosis of $T_i$, where we make the dependence on $\delta$ explicit in the notation. Denote by $m(\gamma^\delta)$ and $\mathsf{SD}(\gamma^\delta)$ the mean and the standard deviation of $\gamma^\delta = (\gamma_1^\delta, \dots, \gamma_p^\delta)$, respectively. We further define the standard error $\mathsf{SE}(\gamma^\delta) = \mathsf{SD}(\gamma^\delta)/\sqrt{p}$. We use one standard error rule to decide the value of $\delta_c$. Namely,

$$\delta_c = \arg\min\{\delta \in (0,1),\ \text{s.t.},\ m(\gamma^\delta) \le \mathsf{SE}(\gamma^\delta)\}\,. \tag{187}$$

Figure 2 corresponds to $\varepsilon = 0.2$. The dots indicate $m(\gamma^\delta)$ and the dotted lines correspond to $m(\gamma^\delta) \pm \mathsf{SE}(\gamma^\delta)$. By one standard error rule, the estimated value of $\delta_c$ works out at $\delta_c = 0.57$.

Figure 3 shows $\delta_c$ versus $\varepsilon$. The black curve corresponds to the case of known covariance, where we set $M = \Omega$ and the red curve corresponds to the case of unknown covariance, where $M$ is set as in Equation (32). The figure confirms that $\delta_c$ scales roughly linearly in $\varepsilon$ (for small $\varepsilon$). In other words, in order for the debiased estimator to have unbiased Gaussian distribution, the sample size $n$ has only to scale linearly in the support size $s_0$. (Note that for the circulant covariance chosen in this example, $s_\Omega = 2$).

Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089, USA
E-mail: ajavanma@usc.edu

Department of Electrical Engineering
and Department of Statistics
Stanford University
Stanford, CA 94305, USA
E-mail: montanar@stanford.edu