The Annals of Statistics
2018, Vol. 46, No. 6A, 2593–2622
https://doi.org/10.1214/17-AOS1630

DEBIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS

BY ADEL JAVANMARD[1] AND ANDREA MONTANARI[2]

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 θ_i^*, $i \in \{1, \dots, p\}$. A key role in these new inferential methods is played by a certain debiased estimator $\widehat{\theta}^{\mathrm{d}}$. Earlier work establishes that, under suitable assumptions on the design matrix, the coordinates of $\widehat{\theta}^{\mathrm{d}}$ are asymptotically Gaussian provided the true parameters vector θ^* 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 θ^*, 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 $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ with $y_i \in \mathbb{R}$, and $x_i \in \mathbb{R}^p$. The response variable y_i is a linear function of x_i, contaminated by noise w_i independent of x_i

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

Received June 2016; revised August 2017.

[1]Supported in part by a Google Faculty Research Award.

[2]Supported in part by NSF Grants CCF-1319979 and DMS-1106627, AFOSR Grant FA9550-13-1-0036 and the Office of the Provost at the University of Southern California through the Zumberge Fund Individual Grant Program.

MSC2010 subject classifications. Primary 62J05, 62J07; secondary 62F12.

Key words and phrases. Lasso, high-dimensional regression, confidence intervals, hypothesis testing, bias and variance, sample size.

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 $x_1^{\mathsf{T}}, \dots, x_n^{\mathsf{T}}$, 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 [8, 10, 14]. A prominent approach is the Lasso [18, 53] defined through the following convex optimization problem:

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

[We will omit the arguments of $\widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ 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 [9, 38, 39, 54, 59] develop methods to compute confidence intervals for single coordinates of the parameters vector θ^*. 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 [38, 39, 54, 59] is to construct a debiased (or de-sparsified) estimator that takes the form

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

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 θ^* is s_0-sparse, that is, it has only s_0 nonzero entries. The key result that allows the construction of confidence intervals in [39, 54, 59] is the following (holding under suitable conditions on the design matrix). If M is "sufficiently close" to Ω, and the sparsity level is

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

then $\widehat{\theta}_i^{\mathrm{d}}$ is approximately Gaussian with mean θ_i^* and variance of order σ^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 [8, 10, 14, 56, 58]

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

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 Ω. Of course, if M is not close to Ω, then $\widehat{\theta}^{\mathrm{d}}$ will not have a Gaussian limit. However, *earlier proofs* [39, 54, 59] *cannot establish the Gaussian limit for $s_0 \gtrsim \sqrt{n}/\log p$, even if Ω 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., $\Omega = \mathrm{I}$) is only understood in the asymptotic limit $s_0, n, p \to \infty$ with s_0/p, n/p having constant limits in $(0, 1)$ [38].

In order to describe the challenge, let us set $M = \Omega$, and recall the common step of the proofs in [39, 54, 59]. Using the definitions (2), (5), we get

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

where $\widehat{\Sigma} = X^{\mathsf{T}}X/n \in \mathbb{R}^{p\times p}$ is the empirical design covariance. Since $w \sim \mathsf{N}(0, \sigma^2\mathrm{I}_n)$, 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 $\widehat{\theta}^{\mathrm{d}}$ to be approximately Gaussian, we need the second term (which can be interpreted as a bias) to vanish. Earlier papers [39, 54, 59] address this by a simple ℓ_1–ℓ_∞ bound. Namely (denoting by $|Q|_\infty$ the maximum absolute value of any entry of matrix Q),

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

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 [8, 10].

This simple argument implies that the debiased estimator is approximately Gaussian if the upper bound in equation (9) is negligible, that is, if $s_0 = o(\sqrt{n}/\log p)$. We see therefore that this requirement is not imposed as to control the error in estimating Ω. It instead follows from the simple ℓ_1–ℓ_∞ bound *even if* Ω *is known.*

1.2. *Main results.* The above exposition should clarify that the $\ell_1 - \ell_\infty$ bound is quite conservative. Considering the ith entry in the bias vector $\mathsf{bias} = (\Omega\widehat{\Sigma} - \mathrm{I})(\theta^* - \widehat{\theta}^{\mathrm{Lasso}})$, the ℓ_1–ℓ_∞ 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 $(\Omega\widehat{\Sigma} - \mathrm{I}) = (\Omega X X^{\mathsf{T}}/n - \mathrm{I})$ 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 ith 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* Σ *satisfies the technical conditions stated in Theorem* 3.8. *Define the debiased estimator* $\widehat{\theta}^{\mathrm{d}}$ *via equation* (5) *with* $M = \Omega$ *and* $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ *with* $\lambda = 8\sigma\sqrt{(\log p)/n}$.

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

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

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 nonasymptotic 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 Ω?*

If the precision matrix Ω is sufficiently structured, then it can be reliably estimated from the design matrix X. Both [59] and [54] assume that Ω is sparse, and use the node-wise Lasso to construct an estimate $\widehat{\Omega}$ [44]. 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 nonasymptotic probability bounds.

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

Let s_Ω the maximum number of nonzero entries in any row of Ω. 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

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

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 λ, and the estimation of the noise level σ. These can be addressed as in [39, 54, 59] 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 θ^*, we recover the condition $s_0 = o(\sqrt{n}/\log p)$ which is assumed in the results of [54, 59]. (Note that [39] obtain the same condition without sparsity assumption on Ω.) In this regime, our improved analysis does not bring any advantage, since the bottleneck is due to the inaccurate estimation of Ω.

On the other hand, if the precision matrix is sparser, we obtain a much weaker condition on the coefficients θ^*. 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 [22, 23] implied accurate reconstruction from a number of random samples scaling quadratically in the number of nonzero coefficients. Subsequent progress was based on the restricted isometry property [14, 15], 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 ℓ_1–ℓ_∞ bound in equation (9) is to modify the definition of debiased estimator in equation (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 $\widehat{\theta}^{\mathrm{Lasso}}$ and the other batch for y and X appearing in equation (5) (and possibly for computing M).

In the Supplementary Material [40], 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 [39, 54, 59]. 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 θ_i:

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

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

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 $H_{0,i}\colon \theta_i^* = 0$.

Noise level and regularization. The construction of the confidence interval $J_i(\alpha)$ in equations (12), (13) requires a suitable choice of the regularization parameter λ, and an estimate of the noise level $\widehat{\sigma}$. The same difficulty was present in [39, 54, 59]. The approaches used there (for instance, using the scaled Lasso [51]) can be followed in the present case as well. Under the assumptions of Theorem 1.1, the same proofs of [39] show that the additional error due to the choice of λ and $\widehat{\sigma}$ are negligible.

Semi-supervised learning. In some applications, the precision matrix Ω can be estimated more accurately thanks to additional information. For instance, in semi-supervised learning, the statistician is given additional samples $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N \in \mathbb{R}^p$ 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 [16] and, therefore, $N \gg n$ or even $N \gg p$. In this setting, we can estimate Ω 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 Σ.

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 [44]. Remarkably, recent contributions have shown the utility of debiasing methods in this context [17, 32, 33].

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 ℓ_∞ norm of the non-Gaussian component in $\widehat{\theta}^{\mathrm{d}}$. 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 [50]).

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 [40].

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 [19, 31, 41, 52]. 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 θ^* 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 [2, 3, 34–36]. In particular, [36] uses the idea of debiased estimator to construct an ℓ_∞ 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 [45, 55]. As mentioned above, sample splitting can also be used to define a modified debiased estimator; see the Supplementary Material [40]. However, sample splitting techniques typically use only part of the data for inference, and are therefore suboptimal. Also, the result depends on the random split of the data.

A method for inference without assumptions on the design matrix was developed in [43]. The resulting confidence intervals are typically quite conservative.

The debiasing method was developed independently from several points of view [9, 38, 39, 54, 59]. The present authors were motivated by the AMP analysis of the Lasso [4–6, 25], and by the Gaussian limits that this analysis implies. In particular, [38] 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 $s_0 \le Cn/\log(p/s_0)$ (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 [38] only holds asymptotically. Also [38] 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 nonasymptotic result under weaker sparsity conditions, and for designs with dependent columns, was proved in [37]. However, this only establishes Gaussianity of $\widehat{\theta}_i^{\mathrm{d}}$ 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 [54]. Robustness to model misspecification was studied in [11]. An R-package for inference in high dimension that uses the node-wise Lasso is available [20]. An R implementation of the method [39] (which does not make sparsity assumptions on Ω) is also available.[3]

3. Main results: Gaussian limit theorems.

3.1. *General notation.* We use e_i to refer to the ith standard basis element, for example, $e_1 = (1, 0, \dots, 0)$. 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 $I, J \subseteq \{1, \dots, p\}$, 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 θ and a subset S, θ_S is the restriction of θ to indices in S. For an integer $p \ge 1$, we use the notation $[p] = \{1, \dots, p\}$ and the shorthand $\sim i$ for the set $[p] \backslash i$. We write $\|v\|_p$ for the

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

standard ℓ_p norm of a vector v, that is, $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$ 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 ℓ_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, that is, $f(n) \le Cg(n)$ for some positive constant C. The notation $f(n) \ll g(n)$ 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 $\widehat{\theta}^{\mathrm{Lasso}}$ 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 Σ, 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 ϕ 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 [8]. 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 nonsingularity of $\widehat{\Sigma}$ for vectors in $\mathcal{C}(s_0, L)$. Rudelson and Zhou [48] 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 ([48], 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*:

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

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

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

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–Schwarz 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, [10]).

LEMMA 3.4 ([10], Lemma 6.2). *Suppose that* $\widehat{\Sigma}_{ii} \le 1$ *for* $i \in [p]$. *Then we have*

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

The following lemma states that the Lasso estimator is sparse. Its proof is given in the Supplementary Material [40].

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 ℓth 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 ℓ_∞ operator norm (maximum ℓ_1 norm of the rows). It is clear that $\rho(A, k)$ is nondecreasing 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 [40].

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* Σ *and* θ^* *is* s_0*-sparse. Suppose that* Σ *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* $\widehat{\theta}^{\mathrm{d}}$ *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*:

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

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

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

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

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 [37], 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 Σ). Assumption (i) sets the normalization of the design matrix. Assumptions (ii) on the eigenvalues of Σ 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 ρ in equation (23), we recover the bound $\|R\|_\infty \lesssim s_0\log p/\sqrt{n}$ which is established in previous work [39, 54, 59]. Note that this bound on the bias does not

require Assumption (iii) (namely, that ρ is a bounded constant). However, Theorem 3.8 asserts that, if ρ 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* $\widehat{\theta}^{\mathrm{d}}$ *is normally distributed. More precisely, let* $\widehat{\sigma} = \widehat{\sigma}(y, X)$ *be an estimator of the noise level satisfying, for any* $\varepsilon > 0$,

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

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

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

Proof of Corollary 3.10 is given in the Supplementary Material [40].

There are several proposals for a consistent estimator of σ. A nonexhaustive list includes [3, 7, 21, 28–30, 47, 49, 51, 57]. For concreteness, we use the the scaled Lasso [51] given by

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

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 [39], Appendix C, for the proof of Lemma 3.11.

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

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

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

REMARK 3.12. It is worth noting that the sample splitting approach, discussed in the Supplementary Material [40], 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 [54, 59] we construct the debiasing matrix M using node-wise Lasso on matrix X. For the 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 ith column of X against all the other columns. Formally,

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

where $X_{\sim i}$ is the submatrix obtained by removing the ith 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

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

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 Σ. Suppose that Assumptions* (i), (ii), (iii) *in Theorem* 3.8 *hold true for Σ. We further let s_Ω be the maximum sparsity of the rows of $\Omega \equiv \Sigma^{-1}$, that is,*

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

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*:

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

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

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 the Supplementary Material [40].

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 equation* (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*

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

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

Using the above distributional characterization, we can construct confidence intervals for the individual model parameters θ_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. equation (27). For the 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:

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

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_Ω is necessary.

To answer this question, we develop a lower bound for $\|R\|_\infty$ based on a minimax theorem for the estimation of coefficient θ_1. This also clarifies the connection between our results and the ones of [13], whose general approach we build on here.

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

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

For $\alpha \in (0, 1)$ and a given parameter space Γ, denote by $\mathcal{I}_\alpha(\Gamma)$ the set of all $(1-\alpha)$-confidence intervals for θ_1 over the entire space Γ,

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

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 θ. Here, and below we focus on the first coordinate θ_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 Γ,

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

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 Γ as follows:

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

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

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

Quantities c, $C_{\min}$ and $C_{\max} \geq 1$ 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 λ. 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 λ such that

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

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

for some known Δ_n and for some known constant C. Then we have

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

Note that since Q is a function of only X, the arguments θ^* and σ^2 in equation (43) are superfluous. To establish the above upper bound, we construct a confidence interval J_α^{d} using a debiased estimator, such that $J_\alpha^{\mathrm{d}} \in \mathcal{I}(\Gamma(s_0, s_\Omega, \rho))$. We refer to the Supplementary Material [40] 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 θ_1 over $\Gamma(s_0, s_\Omega, \rho)$ satisfies*

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

Proposition 4.2 generalizes the result of [13], Theorem 2, which shows that without the sparsity constraint on Ω 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 [40] 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 Δ_n in Proposition 4.1, we obtain the following upper bound on $\ell_\alpha^*(\Gamma(s_0, s_\Omega, \rho))$:

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

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 ℓ_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 [27] 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

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

Note that the minimizer is always unique because Σ is strictly positive definite. In the case $\Sigma = \mathrm{I}$, η_Σ coincides with component-wise soft thresholding at level λ. More generally, $\eta_\Sigma(\cdot)$ can be viewed as a denoising operator associated to the problem of estimating θ^* from the noisy observation $z = \theta^* + \tilde{w}$, where $\tilde{w}$ has covariance Σ. Our next theorem connects the Lasso to this denoising problem and its proof is given in the Supplementary Material [40].

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

(iv) *Letting $C_* \equiv 32 C_{\max}/C_{\min}$, 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 $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ 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 $C_{\min}$, $C_{\max}$, ρ, $\tilde{\rho}$, $\kappa_{\max}$), such that for $n \ge \max(25 \log p, c s_0 \log(p/s_0))$, the following holds true with high probability:

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

Under the hypothesis of this theorem, the Lasso ℓ_2 error is known to be bounded as $\|\widehat{\theta}^{\mathrm{Lasso}} - \theta^*\|_2^2 \le C(s_0 \log p)/n$ [8]. 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 ℓ_2 error for standard designs. We defer its proof to the Supplementary Material [40].

COROLLARY 5.2. *Consider the linear model* (2) *where X has independent Gaussian rows, with zero mean and covariance $\Sigma = \mathrm{I}$. Let $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ 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*

$$\begin{aligned}\|\widehat{\theta}^{\mathrm{Lasso}} - \theta^*\|_2^2 &= \sum_{i \in \mathrm{supp}(\theta^*)} \mathbb{E}_Z\{[\eta(\theta_i^* + n^{-1/2}Z_i; \lambda) - \theta_i^*]^2\} \\ &\quad + O_P\left(\sigma^2 \frac{\sqrt{s_0 \log p}}{n} \vee \sigma^2 \left(\frac{s_0 \log p}{n}\right)^{3/2}\right),\end{aligned}$$

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 [24, 26]. An asymptotic characterizations of the Lasso mean square error for standard Gaussian designs was first obtained in [6]. However, in the present case we recover this as a corollary of a result for general Gaussian designs, and in a nonasymptotic 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 Σ 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 $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ with $\lambda = 8\sigma \times \sqrt{(\log p)/n}$.
(ii) Compute the debiased estimator $\widehat{\theta}^{\mathrm{d}} = \widehat{\theta}^{\mathrm{Lasso}} + n^{-1}\Omega X^{\mathsf{T}}(y - X\widehat{\theta}^{\mathrm{Lasso}})$.
(iii) Compute a new estimator $\widehat{\theta}^{(2)}$ by soft thresholding $\widehat{\theta}^{\mathrm{d}}$ component-wise, namely

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

Here, $\eta(x;\tau) \equiv (|x| - \tau)_+ \operatorname{sign}(x)$ 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 λ 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 Σ, 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*

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

We refer to the Supplementary Material [40] 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 $n, s_0, p \to \infty$ [50]. Candés and Su [50] recently proved that SLOPE achieves the same guarantee for Gaussian designs with $\Sigma = \mathrm{I}$. On one hand, the approach of [50] 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

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

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

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

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

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

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 [27] 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 [42]. Zou, Hastie and Tibshirani [60] showed that the number of degrees of freedom (which enters Efron's formula) coincides with the number of nonzero parameters $\|\widehat{\theta}^{\mathrm{Lasso}}\|_0$. They also proved that $\widehat{\mathsf{R}}_{\mathrm{SURE}}(y, X)$ 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 $\widehat{\mathsf{R}}_{\mathrm{SURE}}(y, X)$ 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 $\Sigma = \mathrm{I}$. Let $\widehat{\theta}^{\mathrm{Lasso}} = \widehat{\theta}^{\mathrm{Lasso}}(y, X; \lambda)$ be the Lasso estimator with $\lambda \geq 9\sigma\sqrt{(\log p)/n}$. 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 $1 - e^{-ct^2} - o_n(1)$:*

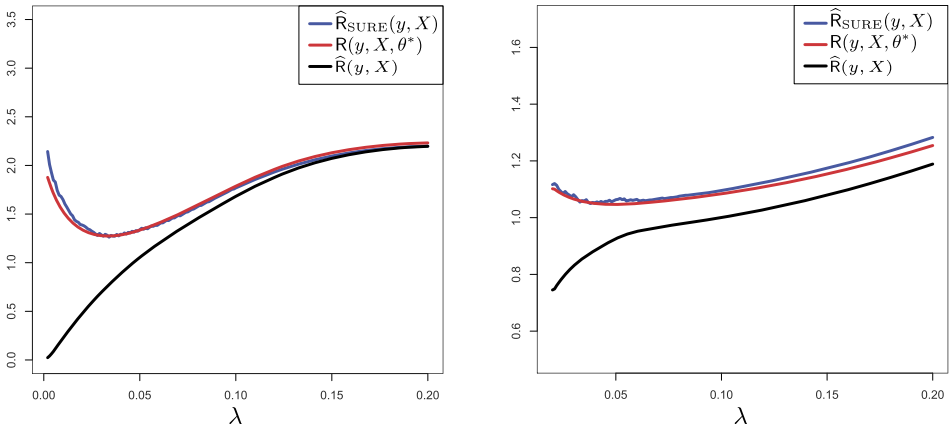
$$\left|\widehat{\mathsf{R}}_{\mathrm{SURE}}(y, X) - \mathsf{R}(y, X, \theta^*)\right| \leq \frac{t\sigma^2}{\sqrt{n}} + \frac{s_0\sigma^2\varepsilon_n}{n}. \tag{54}$$

Proof of Theorem 5.4 is provided in the Supplementary Material [40]. Let us emphasize a few important points:

- The error bound in equation (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 $\widehat{\mathsf{R}}_{\mathrm{SURE}}(y, X)$ is perfectly well defined for arbitrary design covariance Σ.
- While our proof applies to standard designs, $\Sigma = \mathrm{I}$, 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 σ appearing in equation (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 $\widehat{S}$ to obtain an estimate $\widehat{\theta}^{\mathrm{LS}}$. The noise variance is computed as $\widehat{\sigma} = \|y - X\widehat{\theta}^{\mathrm{LS}}\|_2/\sqrt{n}$.

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



(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* $\widehat{\mathsf{R}}_{\mathrm{SURE}}$ *curves versus* λ *for the simulation setting described in Section 5.3.*

Let us mention that [3] also studied estimators similar to $\widehat{\mathsf{R}}_{\mathrm{SURE}}(y, X)$, and related ideas were developed in [46] on the basis of nonrigorous but insightful statistical mechanics techniques. Other approaches to the risk estimation (e.g., [12]) 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 ith coordinate θ_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 $\widehat{\theta}^{\mathrm{p}}$ be the Lasso estimator when one forces $\widehat{\theta}_i^{\mathrm{p}} = \theta_i^*$. With a slight abuse of notation, we use the representation $\theta = (\theta_i, \theta_{\sim i})$.[4] Adopting this convention, we have $\widehat{\theta}^{\mathrm{p}} = (\theta_i^*, \widehat{\theta}_{\sim i}^{\mathrm{p}})$ where

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

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

We observe that $\widehat{\theta}_{\sim i}^{\mathrm{p}}$ 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

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

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

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

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

We decompose the above expression into the following terms:

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

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 $\widehat{\theta}_{\sim i}^{\mathrm{p}}$ 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 $X_{\sim i}(\theta_{\sim i}^* - \widehat{\theta}_{\sim i}^{\mathrm{p}})$.

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

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

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

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

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

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

Using the result of [39], Lemma 23, for $n \ge (A^2 C_{\min})/(4e^2 C_{\max}) \log p$, we have

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

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

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

In addition, on the event $\mathcal{B} \equiv \mathcal{B}_\delta(n, s_0, 3) \cap \tilde{\mathcal{B}}(n, p)$ we have [10]

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

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

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

As discussed, $\widehat{\theta}^{\mathrm{p}}_{\sim i}$ 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 ([10], 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 $\phi^2(S, \widehat{\Sigma}_{\sim i,\sim i}) \geq \phi^2(S, \widehat{\Sigma})$. 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* $\lambda \geq 8\sigma\sqrt{(\log p)/n}$,

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

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

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

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

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

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

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. ζ_i using the fact that v_i and ζ_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

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

• *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 [40] for the proof of Lemma 6.3.

LEMMA 6.3 (Perturbation bound). *Suppose that* $\Sigma_{ii} \leq 1$, *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*:

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

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

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

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

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

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

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

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: 10.1214/17-AOS1630SUPP; .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 [40].

REFERENCES

[1] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. System identification and time-series analysis. MR0423716

[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876

[3] BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[4] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755

[5] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. MR2810285

[6] BAYATI, M. and MONTANARI, A. (2012). The Lasso risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312

[7] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

[8] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

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

[10] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761

[11] BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Stat.* **9** 1449–1473. MR3367666

[12] CAI, T. T. and GUO, Z. (2016). Accuracy assessment for high-dimensional linear regression. Available at arXiv:1603.03474.

[13] CAI, T. T., GUO, Z. et al. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

[14] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[15] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846

[16] CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

[17] CHEN, M., REN, Z., ZHAO, H. and ZHOU, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111** 394–406.

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

[19] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Ann. Rev. Econ.* **7** 649–688.

[20] DEZEURE, R., BÜHLMANN, P., MEIER, L., MEINSHAUSEN, N. et al. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statist. Sci.* **30** 533–558. MR3432840

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

[22] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[23] DONOHO, D. L. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** 2845–2862. MR1872845

[24] DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory* **59** 3396–3433. MR3061255

[25] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[26] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise sensitivity phase transition in compressed sensing. *IEEE Trans. Inform. Theory* **57** 6920–6941.

[27] EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR2090899

[28] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

[29] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

[30] FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038.

[31] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Available at arXiv:1410.2597.

[32] JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. MR3354336

[33] JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST* **26** 143–162.

[34] JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065. MR3689308

[35] JANSON, L., SU, W. et al. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975.

[36] JAVANMARD, A. and LEE, J. D. (2017). A Flexible Framework for Hypothesis Testing in High-dimensions. Available at arXiv:1704.07971.

[37] JAVANMARD, A. and MONTANARI, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In 51*st Annual Allerton Conference* 1427–1434.

[38] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

[39] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909.

[40] JAVANMARD, A. and MONTANARI, A. (2018). Supplement to "Debiasing the Lasso: Optimal sample size for Gaussian designs." DOI:10.1214/17-AOS1630SUPP.

[41] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the Lasso. *Ann. Statist.* **42** 413–468. MR3210970

[42] MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

[43] MEINSHAUSEN, N. (2014). Group bound: Confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 923–945. MR3414134

[44] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[45] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

[46] OBUCHI, T. and KABASHIMA, Y. (2016). Cross validation in LASSO and its acceleration. *J. Stat. Mech. Theory Exp.* **2016** 53304–53339.

[47] REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2013). A study of error variance estimation in Lasso regression. Available at arXiv:1311.5274.

[48] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256

[49] STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

[50] SU, W., CANDES, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953

[51] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[52] TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. Available at arXiv:1401.3889.

[53] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[54] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[55] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

[56] YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

[57] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[58] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[59] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242.

[60] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967

DATA SCIENCES AND OPERATIONS DEPARTMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: ajavanma@usc.edu

DEPARTMENT OF ELECTRICAL ENGINEERING
AND DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: montanar@stanford.edu