The Annals of Statistics
2023, Vol. 51, No. 5, 2194–2220
https://doi.org/10.1214/23-AOS2327

THE LASSO WITH GENERAL GAUSSIAN DESIGNS WITH APPLICATIONS TO HYPOTHESIS TESTING

BY MICHAEL CELENTANO[1,a], ANDREA MONTANARI[2,b], AND YUTING WEI[3,c]

[1]*Department of Statistics, University of California at Berkeley,* [a]*mcelentano@berkeley.edu*
[2]*Department of Statistics, Stanford University,* [b]*montanar@stanford.edu*
[3]*Department of Statistics and Data Science, University of Pennsylvania,* [c]*ytwei@wharton.upenn.edu*

The Lasso is a method for high-dimensional regression, which is now commonly used when the number of covariates p is of the same order or larger than the number of observations n. Classical asymptotic normality theory does not apply to this model due to two fundamental reasons: (1) The regularized risk is nonsmooth; (2) The distance between the estimator $\widehat{\boldsymbol{\theta}}$ and the true parameters vector $\boldsymbol{\theta}^*$ cannot be neglected. As a consequence, standard perturbative arguments that are the traditional basis for asymptotic normality fail.

On the other hand, the Lasso estimator can be precisely characterized in the regime in which both n and p are large and n/p is of order one. This characterization was first obtained in the case of Gaussian designs with i.i.d. covariates: here we generalize it to Gaussian correlated designs with nonsingular covariance structure. This is expressed in terms of a simpler "fixed-design" model. We establish nonasymptotic bounds on the distance between the distribution of various quantities in the two models, which hold uniformly over signals $\boldsymbol{\theta}^*$ in a suitable sparsity class and over values of the regularization parameter.

As an application, we study the distribution of the debiased Lasso and show that a degrees-of-freedom correction is necessary for computing valid confidence intervals.

1. Introduction. Questions of statistical inference and decision theory are often addressed by characterizing the distribution of the estimator of interest under a variety of assumptions on the data distribution. A central role is played by normal theory, which guarantees that broad classes of estimators are asymptotically normal with prescribed covariance structure [31, 40]. Normality theory can serve as the basis for inference, facilitate the comparison of estimators and justify claims of efficiency.

In high dimensions, the distributional theory available for many estimators of interest is more limited. Frequently, we have access to upper and lower bounds on important quantities like the estimation or prediction error or the size of a selected model. These may have the correct dependence on sample size, dimensionality and certain structural parameters, but are usually loose in their leading constants. Asymptotic normality often breaks down in high dimensions, even when considering low-dimensional projections of the coefficients vector [4, 37, 53, 59]. There has been substantial progress in recovering normality in special cases by resorting to careful constructions designed to remove bias and target normality [4, 9, 20, 37, 59]. It is of substantial interest to identify precisely the conditions under which such constructions succeed and fail. This challenge is compounded by the fact that resampling methods also fail in this context [30].

Received July 2022; revised June 2023.
MSC2020 subject classifications. Primary 62J07, 62E17; secondary 62F05, 62F12.
Key words and phrases. Lasso, debiased Lasso, exact asymptotics, convex Gaussian min–max theorem, Gaussian designs, Gaussian width.

The Lasso is arguably the prototypical method in high-dimensional statistics. Given data $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$, with $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^p$, it performs linear regression of the y_i's on the $\boldsymbol{x}_i$'s by solving the optimization problem

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{R}(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\theta}\|_1 \right\}. \tag{1}$$

Here, $\boldsymbol{y} \in \mathbb{R}^n$ is the vector with ith entry equal to y_i, and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the matrix with ith row given by $\boldsymbol{x}_i^\top$. Throughout the paper, we will assume the model to be well specified. Namely, there exist $\boldsymbol{\theta}^* \in \mathbb{R}^p$ such that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \sigma \boldsymbol{z}, \tag{2}$$

where $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_n)$ is a Gaussian noise vector.[1] In the informal discussion below, we will assume $\boldsymbol{\theta}^*$ to be s-sparse (i.e., to have at most s nonzero entries), although our theorems apply more generally to coefficient vectors that are only approximately sparse.

Distribution theory for the Lasso. A substantial body of theoretical work studies the Lasso with fixed (nonrandom) designs $\boldsymbol{X}$ in the regime $s\log(p/s)/n = O(1)$ [6, 12, 13, 48] by providing estimation error bounds that are rate optimal. These results have two types of limitations. First, they usually require that λ be chosen larger than the approximate minimax choice $\lambda_{\mathrm{MM}} = c\sigma\sqrt{\log(p/s)}$ (with c a constant which cannot be taken arbitrarily small). In practice, however, λ is chosen by cross-validation and is often significantly smaller than λ_{MM} because the coefficient $\boldsymbol{\theta}^*$ is not the least favorable one [21, 45]. Second, these require restricted eigenvalue or similar compatibility conditions on the design matrix $\boldsymbol{X}$. These conditions only hold for sample sizes that are strictly larger than what is necessary for accurate estimation when $\boldsymbol{X}$ is random.

A more recent line of research attempts to address these limitations by characterizing the distribution of $\widehat{\boldsymbol{\theta}}$ with Gaussian design matrices [4, 38, 45, 55]. For example, [4] proved in the case of i.i.d. Gaussian designs an exact characterization of the distribution of $\widehat{\boldsymbol{\theta}}$, which is simple enough to be described in words. Imagine, instead of observing $\boldsymbol{y}$ according to the linear model (2), we are given $\boldsymbol{y}^f = \boldsymbol{\theta}^* + \tau\boldsymbol{g}$ where $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_p)$, and $\tau > \sigma$ is the original noise level inflated by the effect of undersampling. Then $\widehat{\boldsymbol{\theta}}$ is approximately distributed as $\eta(\boldsymbol{y}^f; \zeta)$ where $\eta(x;\zeta) := (|x| - \lambda/\zeta)_+ \operatorname{sign}(x)$ is the soft thresholding function (applied to vectors entrywise) and ζ controls the threshold value. The values of τ, ζ are determined by a system of two nonlinear equations (see below). This analysis, as well as that in [45, 55], assumes n, p and the number of nonzero coefficients s to be large and of the same order. It further applies to any λ scaling as $c\sigma\sqrt{\log(p/s)}$. In particular, unlike the Lasso results in [6, 12, 13, 48], the constant c here can be taken arbitrarily small, though nonvanishing asymptotically, which covers the typical values of the regularization selected by standard procedures such as cross-validation [21, 45].

Of course the case of i.i.d. Gaussian covariates is highly idealized and one can think of two directions in which the results of [4, 45, 55] could be brought closer to reality:

1. Non-Gaussian but still independent and—say— sub-Gaussian covariates. Both numerical simulations and universality arguments suggest that the same characterization that was proven for Gaussian covariates also applies to this case. Rigorous universality results were proven in [3, 46, 49] in closely related settings. Hence, while mathematically interesting, this generalization yields limited new statistical insight.

[1] The assumption of Gaussian noise is not necessary for our results, but is made throughout to simplify our exposition and proofs. See Remark 4.2.

2. Gaussian but correlated designs. As we will see, in this case the asymptotic characterization is different and depends on the covariance $\mathbf{\Sigma} = \mathbb{E}\{\boldsymbol{x}_i \boldsymbol{x}_i^\top\}$. The covariance $\mathbf{\Sigma}$ (or an estimate of $\mathbf{\Sigma}$) plays a key role in statistically important tasks such as debiasing and hypothesis testing. This will be the focus of the present paper.

By analogy with the uncorrelated designs, we expect our results for correlated Gaussian designs to apply also to correlated non-Gaussian designs. A set of results proved after a first appearance of this manuscript work supports this expectation [33, 35, 47].

Throughout the paper, we assume that the covariates (each row of $\boldsymbol{X}$) have distribution

$$\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{\Sigma})$$

for some well-conditioned and known covariance matrix $\mathbf{\Sigma}$. As in the i.i.d. case, our results present two advantages with respect to fixed-design theory. First, they allow for any λ of the order $c\sigma\sqrt{\log(p/s)}$, with c an arbitrarily small (nonzero) constant. Second, they provide guarantees for sample sizes n at which the restricted eigenvalue condition does not hold.

In fact, we provide guarantees for all sample sizes above the Gaussian dimension of the relevant descent cone. This critical sample size marks a sharp transition in the ability of ℓ_1-based methods to achieve noiseless and stable sparse recovery in compressed sensing [19, 56]. We will refer to this as the *Donoho–Tanner* phase transition (although the original work of [25, 26] was limited to i.i.d. designs). More details can be found in our Section 3.

In the case of correlated designs, [38] proved a similar characterization in the regime $s\log(p)/n = o(1)$ assuming a bound on $\|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_1$. The regime studied [38] is substantially simpler than the one studied here. In particular, the characterization proved here simplifies in that regime, in that one can take $\tau = \sigma$ and $\zeta = 1$.

An important consequence of our theory is the asymptotic optimality of a hyperparameter tuning method based on the following degrees-of-freedom adjusted residuals:

$$\widehat{\tau}(\lambda)^2 := \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)^2}. \tag{3}$$

It was already observed in [45] that minimizing $\widehat{\tau}(\lambda)$ over λ provides a good selection procedure for the regularization parameter. Our results provide theoretical support for this approach under general Gaussian designs. Recently (and after this paper was originally posted), this criterion has been generalized to a wider class of losses and penalties [7].

Distribution theory for the debiased Lasso. The debiased Lasso is a recently popularized approach for performing hypothesis testing and computing confidence regions for low-dimensional projections of $\boldsymbol{\theta}^*$. Most constructions take the form

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}),$$

for an appropriate and possibly data-dependent choice of the matrix $\boldsymbol{M}$. Under appropriate choices of $\boldsymbol{M}$, low-dimensional projections of $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ are approximately normal with mean $\boldsymbol{\theta}^*$.

The first constructions for the debiased Lasso took $\boldsymbol{M}$ to be suitable estimators of the precision matrix $\mathbf{\Sigma}^{-1}$ and proved approximate normality when $\|\boldsymbol{\theta}^*\|_0 =: s = o(\sqrt{n}/\log p)$ [36–38, 57, 59]. Later work considered the case of Gaussian covariates with known covariance, and set $\boldsymbol{M} = \mathbf{\Sigma}^{-1}$. In this idealized setting, the sparsity condition was relaxed to $s = o(n/(\log p)^2)$ under an ℓ_1-constraint on $\mathbf{\Sigma}^{-1}\boldsymbol{e}_j$ [38], and to $s = o(n^{2/3}/\log(p/s)^{1/3})$ for general $\mathbf{\Sigma}$ [9]. The latter conditions turn out to be tight for $\boldsymbol{M} = \mathbf{\Sigma}^{-1}$.

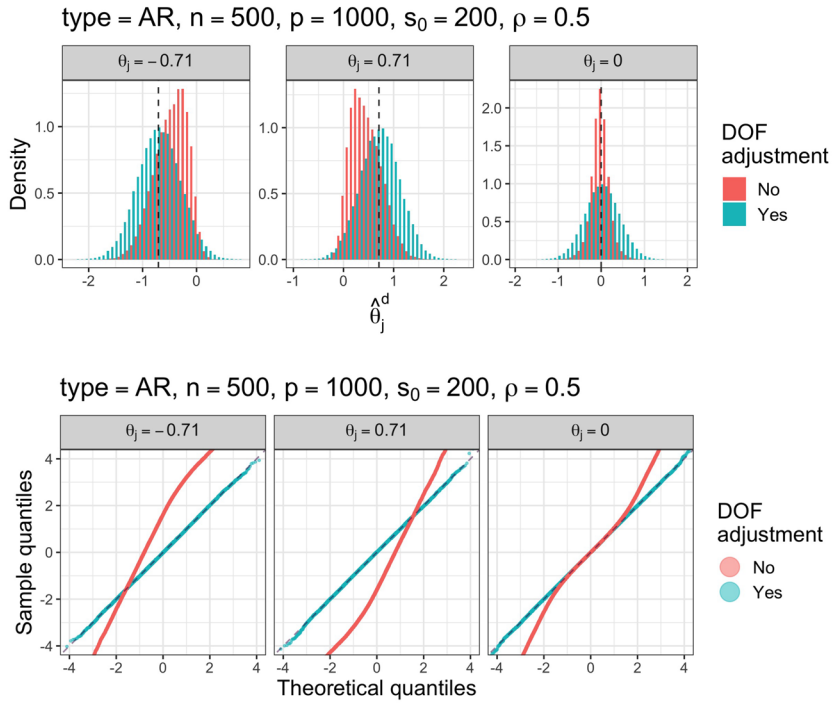

FIG. 1. *The debiased Lasso with and without degrees-of-freedom (DOF) adjustment. Here,* $p = 1000$, $n = 500$, $s = 200$, $\Sigma_{ij} = \rho^{|i-j|} = 0.5^{|i-j|}$, $\lambda = 4/\sqrt{n} = .18$, $\sigma = 1$. *The coefficients vector* $\boldsymbol{\theta}^*$ *contains* 100 *entries* $\theta_i^* = +.707$, *and* 100 *entries* $\theta_i^* = -.707$. *The histogram plots the raw values of* $\widehat{\theta}_j^{\mathrm{d}}$ *without standardization, with the true value of* θ_j^* *drawn as the vertical dashed line. The qqplot is made with theoretical quantiles from the standard normal distribution.*

For larger values of s, it is necessary to adjust the previous construction for the degrees-of-freedom by setting[2] $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}/(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)$:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \frac{1}{n - \|\widehat{\boldsymbol{\theta}}\|_0} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^{\top} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}). \tag{4}$$

Figure 1 illustrates the difference between the debiased estimator with and without degrees-of-freedom correction. It is clear that debiasing without degrees-of-freedom correction can lead to invalid inference.

Recently, Bellec and Zhang [9, 10] established asymptotic normality and unbiasedness of the coordinates $\widehat{\theta}_j^{\mathrm{d}}$ of the debiased estimator of equation (4). As in the present work, they assumed correlated Gaussian designs in the proportional regime $s \asymp n \asymp p$. Our results on debiasing are not directly comparable with the ones of [10]: on the one hand, we assume weaker condition on the regularizations and the sample size; on the other hand, we establish normality in a weaker sense. See Section 4.5 for further discussion.

Our results on the debiased Lasso do not imply that a fixed coordinate of $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ is approximately unbiased and normally distributed. Indeed, without additional assumptions, there can be a small subset of coordinates for which normality does not hold [10]. Instead, we present an alternative *leave-one-out* method to construct confidence intervals for which we prove

[2]More precisely, [37, 45] showed that the degrees-of-freedom correction is needed for uncorrelated designs with $s = \Theta(n)$, [9] showed that it is needed for correlated designs with $n \gg s \gg n^{2/3}/\log(p/s)^{1/3}$, and [10] studied it for correlated designs with $s = \Theta(n)$, but under stronger conditions on the sample size and regularization parameter than considered here.

asymptotic validity via a direct argument. An advantage of the leave-one out method is that it produces p-values for single coordinates that are exact (not just asymptotically valid for large n, p). Empirically, the leave-one-out intervals almost exactly agree with the debiased intervals in several settings. On the other hand, we demonstrate that—for certain carefully designed $(\boldsymbol{\theta}^*, \boldsymbol{\Sigma})$—the leave-one-out intervals can be smaller than the debiased intervals.

Notation. We generally use lowercase for scalars (e.g., $x, y, z, \dots$), boldface lowercase for vectors (e.g., $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \dots$) and boldface uppercase for matrices (e.g., $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \dots$). We denote the support of vector $\boldsymbol{x}$ as $\operatorname{supp}(\boldsymbol{x}) := \{i \,|\, \boldsymbol{x}_i \neq 0\}$. In addition, the ℓ_q norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{x}\|_q^q \equiv \sum_{i=1}^n |x_i|^q$. For $r \geq 0$ and $q \in (0, \infty)$, we use $\mathsf{B}_q(\boldsymbol{v}; r)$ to represent the corresponding ℓ_q-ball of radius r and center $\boldsymbol{v}$, namely

$$\mathsf{B}_q(\boldsymbol{v}; r) := \big\{\boldsymbol{x} \in \mathbb{R}^p \,|\, \|\boldsymbol{x} - \boldsymbol{v}\|_q \leq r\big\} \quad \text{for } q > 0 \quad \text{and} \quad \mathsf{B}_0(s) := \big\{\boldsymbol{\theta} \in \mathbb{R}^p \,|\, \|\boldsymbol{\theta}\|_0 \leq s\big\}.$$

If the center is omitted, it should be understood that the ball is centered at $\mathbf{0}$. A function $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz if for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p \times \mathbb{R}^p$, it satisfies $|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$. The notation $\mathbb{S}_{\geq 0}^n$ is used to denote the set of $n \times n$ positive semidefinite matrices. We reserve n for the sample size, p for the dimension of the unknown parameter $\boldsymbol{\theta}^*$ and always define $\delta := n/p$.

2. A glimpse of our results. Our main result establishes an approximate equivalence between the undersampled linear model of equation (2) and a related statistical model:

$$\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \frac{\tau}{\sqrt{n}}\boldsymbol{g}. \tag{5}$$

Here, $\boldsymbol{g} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$ and $\tau \geq 0$. We may take any square root of the matrix $\boldsymbol{\Sigma}$. For simplicity, we always assume that we take a symmetric square root. The reader should have in mind a setting in which the singular values of $\boldsymbol{\Sigma}$ and the noise parameter τ are of order 1.

We call equation (5) the *fixed-design model* (hence the superscript f) and call model (2) the *random-design model*. The Lasso estimator in the fixed-design model can be written as

$$\widehat{\boldsymbol{\theta}}^f := \eta(\boldsymbol{y}^f, \zeta) := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \frac{\zeta}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\theta}\|_1 \right\}, \tag{6}$$

with predictions given by $\widehat{\boldsymbol{y}}(\boldsymbol{y}^f, \zeta) := \boldsymbol{\Sigma}^{1/2}\eta(\boldsymbol{y}^f, \zeta)$. We define the debiased Lasso in the fixed-design model as

$$\widehat{\boldsymbol{\theta}}^{f,\mathrm{d}} := \widehat{\boldsymbol{\theta}}^f + \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}^f = \boldsymbol{\theta}^* + \frac{\tau}{\sqrt{n}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}. \tag{7}$$

The approximate equivalence between the random design and fixed design models holds for particular choices of τ and ζ, which we denote τ^* and ζ^*. Such an equivalence is relatively straightforward in the low-dimensional regime: in that case, it is sufficient to take $\tilde{\boldsymbol{y}}^f = n^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\boldsymbol{y}$, and check that for $n \gg p$, this is approximately distributed as $\boldsymbol{y}^f$ of equation (5) with $\tau = \sigma$. This equivalence was extended by [38], Theorem 5.1, to $n \gg s\log(p)/n$, assuming $\max_j \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1 = O(1)$. As long as these conditions are met, we can keep $\tau = \sigma$ and $\zeta = 1$.

Here, we consider the more interesting case $s\log(p/s)/n = \Theta(1)$ without an ℓ_1-restriction on the rows of $\boldsymbol{\Sigma}^{-1}$. In this regime, the equivalence only holds if we properly select $\tau^* > \sigma$ and $\zeta^* < 1$.

To specify these choices of τ and ζ, let the in-sample prediction risk and degrees-of-freedom of the Lasso estimator in the fixed-design model be

$$\mathsf{R}(\tau^2, \zeta) := \mathbb{E}\left[\left\|\widehat{\boldsymbol{y}}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \frac{\tau}{\sqrt{n}}\boldsymbol{g}, \zeta\right) - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*\right\|_2^2\right], \tag{8a}$$

$$\mathsf{df}(\tau^2,\zeta) := \frac{\sqrt{n}}{\tau}\mathbb{E}\Big[\Big\langle \widehat{\boldsymbol{y}}\Big(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \frac{\tau}{\sqrt{n}}\boldsymbol{g},\zeta\Big),\boldsymbol{g}\Big\rangle\Big] = \mathbb{E}\Big[\Big\|\eta\Big(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \frac{\tau}{\sqrt{n}}\boldsymbol{g},\zeta\Big)\Big\|_0\Big], \tag{8b}$$

where the expectation is taken over $\boldsymbol{g} \sim \mathsf{N}(0,\mathbf{I}_p)$. Here, for notational simplicity, we leave the dependence of $\mathsf{R}(\tau^2,\zeta)$ and $\mathsf{df}(\tau^2,\zeta)$ on $\boldsymbol{\theta}^*$, $\boldsymbol{\Sigma}$, n, p and λ implicit. The notion of "degrees-of-freedom" is standard to quantify the model complexity of statistical procedures (see, e.g., [27, 28, 34] and references therein), and its equivalence to the expected sparsity of the Lasso estimate holds, for example, by [60], Theorem 1. The parameters τ^*, ζ^* are chosen as solutions to the system of equations

$$\tau^2 = \sigma^2 + \mathsf{R}(\tau^2,\zeta), \tag{9a}$$

$$\zeta = 1 - \frac{\mathsf{df}(\tau^2,\zeta)}{n}. \tag{9b}$$

We refer to these equations as the *fixed-point equations*. As asserted in Section 4.1, there exists a unique pair of solution to the above fixed-point equations under weak conditions.

Role of fixed-point equations. Before presenting our assumptions and results formally, it is useful to discuss the interpretation of τ^* and ζ^*. In what follows, we take $\widehat{\boldsymbol{\theta}}^f$ and $\widehat{\boldsymbol{\theta}}^{f,\mathrm{d}}$ to be computed according to equation (7) in the fixed-design model with parameters $\tau = \tau^*$, $\zeta = \zeta^*$, which solve the fixed-point equations (9a) and (9b).

- *Prediction and estimation error of the Lasso.* We can interpret τ^{*2} as a theoretical prediction for the test error $\mathbb{E}[(y_{\mathrm{test}} - \boldsymbol{x}_{\mathrm{test}}^\top\widehat{\boldsymbol{\theta}})^2]$ on an independent test sample $(\boldsymbol{x}_{\mathrm{test}}, y_{\mathrm{test}})$. Indeed, we obviously have $\mathbb{E}[(y_{\mathrm{test}} - \boldsymbol{x}_{\mathrm{test}}^\top\widehat{\boldsymbol{\theta}})^2] = \sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$. We will prove that the prediction risk $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ concentrates on the prediction risk of the fixed-design model $\mathsf{R}(\tau^{*2},\zeta^*) = \mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2]$; cf. equation (8a). Similarly, we will prove that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$ concentrates on $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_2^2]$. We conclude that $\mathbb{E}[(y_{\mathrm{test}} - \boldsymbol{x}_{\mathrm{test}}^\top\widehat{\boldsymbol{\theta}})^2]$ concentrates on τ^{*2} by equation (9a).
- *Model size of the Lasso.* ζ^* is interpreted as (a theoretical prediction for) the fraction of coordinates *not* selected by the Lasso. Indeed, we will prove that the model size in the random design model $\|\widehat{\boldsymbol{\theta}}\|_0$ concentrates around $\mathsf{df}(\tau^{*2},\zeta^*)$, that is, the expected model size in the fixed-design model; cf. equation (8b). The interpretation follows by the second fixed-point equation (9b). By equation (6), we can also interpret ζ^* as an inverse effective regularization parameter. Thus, the larger the size of the selected model, the smaller the effective regularization.
- *False discovery proportion (FDP) of the debiased Lasso.* Consider the task of constructing confidence intervals for coordinates of $\boldsymbol{\theta}^*$. For each $j \in [p]$, define the interval

  $$\mathsf{CI}_j^{\mathrm{d}} := [\widehat{\theta}_j^{\mathrm{d}} - \Sigma_{j|-j}^{-1/2}\widehat{\tau}z_{1-q/2}/\sqrt{n}, \widehat{\theta}_j^{\mathrm{d}} + \Sigma_{j|-j}^{-1/2}\widehat{\tau}z_{1-q/2}/\sqrt{n}],$$

  where $z_{1-q/2}$ is the $(1-q/2)$-quantile of the standard normal distribution, $\widehat{\tau}$ is an empirical estimate of τ^* (defined formally in (3)) and

  $$\Sigma_{j|-j} := \Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}(\boldsymbol{\Sigma}_{-j,-j})^{-1}\boldsymbol{\Sigma}_{-j,j}.$$

  We prove that the false-coverage proportion (FCP) concentrates around q, where

  $$\mathsf{FCP} := \frac{1}{p}\sum_{j=1}^p \mathbf{1}_{\theta_j^* \notin \mathsf{CI}_j^{\mathrm{d}}} = \frac{1}{p}\sum_{j=1}^p \mathbf{1}\{|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*| > \Sigma_{j|-j}^{-1/2}\widehat{\tau}z_{1-\alpha/2}/\sqrt{n}\}.$$

In other words, confidence intervals based on the debiased Lasso achieve nominal false coverage. Combining this with the fact that $q = \mathbb{E}[\frac{1}{p}\sum_{j=1}^{p} \mathbf{1}\{|\widehat{\theta}_j^{f,\mathrm{d}} - \theta_j^*| \geq \Sigma_{j|-j}^{-1/2}\tau z_{1-\alpha/2}/\sqrt{n}\}]$, we conclude the FCP in the random-design model concentrates on the expectation of the analogous quantity in the fixed-design model.

The above result provides an additional interpretation of the fixed-point parameter τ^{*2} as the effective noise-level for the debiased Lasso estimates. Note that in the low-dimensional limit, which takes p fixed, $n \to \infty$, the asymptotic standard error of the OLS estimate for θ_j^* is given by $\Sigma_{j|-j}^{-1/2}\sigma/\sqrt{n}$. The first fixed-point equation states that we should inflate this standard error by replacing σ^2 with $\sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$, which concentrates around τ^{*2}. Of course, under a low-dimensional asymptotics, we expect $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 \xrightarrow{\mathrm{p}} 0$, recovering the low-dimensional theory.

Versions of these results and the corresponding interpretations of τ^*, ζ^* have appeared elsewhere [2, 4, 5, 22, 45, 54]. The present paper is the first one establishing these results under correlated Gaussian designs and optimal sample size requirements.

3. Preliminaries. This section summarizes several important concepts that shall be used throughout the paper and discusses the assumptions under which our main results are derived.

Gaussian width and the Donoho–Tanner phase transition. The success probability of ℓ_1-norm based methods changes abruptly at a critical sampling rate δ_{DT}, which depends on the sparsity of the signal and the geometry of the covariates. We will refer to this phenomenon as the Donoho–Tanner phase transition [25, 26]. Below the transition (roughly speaking, for $n/p < \delta_{\mathrm{DT}}$), ℓ_1-penalized methods fail to achieve exact noiseless recovery, stable noisy recovery, bounded minimax noisy recovery over sparse balls and full power for variable selection [19, 23, 24, 51, 56, 58]. Above the transition (for $n/p > \delta_{\mathrm{DT}}$), ℓ_1-penalized methods are able to succeed according to these metrics.

This paper uses Gaussian comparison techniques [19, 45], and our results hold for all sampling rates n/p exceeding δ_{DT}, where δ_{DT} is defined below in terms of a certain Gaussian width. We anticipate that our definition of this threshold is (for general $\boldsymbol{\Sigma}$) slightly different from the standard one in the literature. Importantly, the restricted eigenvalue conditions, which are often used to derive estimation error bounds on the Lasso need not occur near the Donoho–Tanner phase transition. Hence, our results could not be established using those conditions.

Given a vector $\boldsymbol{x} \in \{+1, -1, 0\}^p$, define the closed convex cone $\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma})$ and the homogeneous convex function $F(\cdot; \boldsymbol{x}, \boldsymbol{\Sigma}) : \mathbb{R}^p \to \mathbb{R}$ as follows:

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma}) := \{\boldsymbol{v} \in \mathbb{R}^p : F(\boldsymbol{v}; \boldsymbol{x}, \boldsymbol{\Sigma}) \leq 0\},$$

$$F(\boldsymbol{v}; \boldsymbol{x}, \boldsymbol{\Sigma}) := \langle \boldsymbol{x}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\rangle + \|(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v})_{S^c}\|_1 \quad \text{for } S := \mathrm{supp}(\boldsymbol{x}).$$

(The reader should think of $\boldsymbol{v}$ as $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}$ is the argument appearing in the Lasso optimization.)

Consider $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with $\boldsymbol{x} = \mathrm{sign}(\boldsymbol{\theta}^*)$, that is, $x_j = 1$ for $\theta_j^* > 0$, $x_j = -1$ for $\theta_j^* < 0$, and $x_j = 0$ for $\theta_j^* = 0$. Then $\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma})$ is the descent cone of the function $\boldsymbol{v} \mapsto \|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\|_1$ at $\boldsymbol{v} = \boldsymbol{0}$, namely (denoting by $\mathrm{cl}(A)$ the closure of set A)

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma}) := \mathrm{cl}(\{\boldsymbol{v} \in \mathbb{R}^p : \exists \varepsilon > 0 \text{ s.t. } \|\boldsymbol{\theta}^* + \varepsilon\Sigma^{-1/2}\boldsymbol{v}\|_1 \leq \|\boldsymbol{\theta}^*\|_1\}).$$

The connection between this cone and the Lasso is most easily seen in the case of minimum ℓ_1-norm interpolation (basis pursuit), corresponding to the $\lambda \to 0$ limit of the Lasso (1):

$$\widehat{\boldsymbol{\theta}}_{\mathrm{BP}} := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \{\|\boldsymbol{\theta}\|_1 \text{ s.t. } \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}\}.$$

In the noiseless case $\sigma = 0$ (i.e., $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^*$), $\widehat{\boldsymbol{\theta}}_{\mathrm{BP}} = \boldsymbol{\theta}^*$ if and only if $\mathrm{null}(\boldsymbol{G}) \cap \mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma}) = \{\boldsymbol{0}\}$ where $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ is a Gaussian matrix with i.i.d. entries [1]. As proven in [1], the probability of the event $\mathrm{null}(\boldsymbol{G}) \cap \mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma}) = \{\boldsymbol{0}\}$ transitions rapidly from 0 to 1 when the sampling ratio n/p crosses $\mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma})^2$. Specifically, [1], Theorem II, ensures that

$$\text{if } \frac{n}{p} \le \mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma})^2 - \Delta, \quad \mathbb{P}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{BP}} = \boldsymbol{\theta}^*\big) \le 4\exp\big(-p\Delta^2/8\big);$$

$$\text{if } \frac{n-1}{p} \ge \mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma})^2 + \Delta, \quad \mathbb{P}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{BP}} = \boldsymbol{\theta}^*\big) \ge 1 - 4\exp\big(-p\Delta^2/8\big).$$

Here, $\mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma})$ is the Gaussian width of $\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma})$ defined as follows [19, 32, 56]:

$$\mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{p}}\mathbb{E}\Big[\max_{\substack{\boldsymbol{v}\in\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma}) \\ \|\boldsymbol{v}\|_2^2\le 1}} \langle \boldsymbol{v}, \boldsymbol{g}\rangle\Big]. \tag{10}$$

We next introduce the modified width that is relevant for our results. Consider the probability space $(\mathbb{R}^p, \mathcal{B}, \gamma_p)$ with $\mathcal{B}$ being the Borel σ-algebra and γ_p the standard Gaussian measure in p dimensions. We denote by $L^2 := L^2(\mathbb{R}^p; \mathbb{R}^p)$ the space of functions $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ that are square integrable in $(\mathbb{R}^p, \mathcal{B}, \gamma_p)$. This space is equipped with the scalar product

$$\langle \boldsymbol{f}_1, \boldsymbol{f}_2\rangle_{L^2} = \mathbb{E}[\langle \boldsymbol{f}_1(\boldsymbol{g}), \boldsymbol{f}_2(\boldsymbol{g})\rangle] = \int \langle \boldsymbol{f}_1(\boldsymbol{g}), \boldsymbol{f}_2(\boldsymbol{g})\rangle \gamma_p(\mathrm{d}\boldsymbol{g}).$$

The standard notion of Gaussian width defined in equation (10) can be rewritten as

$$\mathcal{G}_d(\boldsymbol{x}, \boldsymbol{\Sigma}) := \sup_{\boldsymbol{v}\in L^2}\Big\{\frac{1}{\sqrt{p}}\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2} : \mathbb{P}\big(\|\boldsymbol{v}\|_2 \le 1\big) = 1, \mathbb{P}\big(F(\boldsymbol{v}; \boldsymbol{x}, \boldsymbol{\Sigma}) \le 0\big) = 1\Big\}, \tag{11}$$

where $\boldsymbol{g}$ denotes the identity function on L^2. Let us emphasize that the supremum is taken over functions $\boldsymbol{v} : \mathbb{R}^p \to \mathbb{R}^p$, $\boldsymbol{g} \mapsto \boldsymbol{v}(\boldsymbol{g})$.

Instead of (11), we will make use of the following relaxed version of Gaussian width:

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) := \sup_{\boldsymbol{v}\in L^2}\Big\{\frac{1}{\sqrt{p}}\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2} : \|\boldsymbol{v}\|_{L^2} \le 1, \mathbb{E}\big[F(\boldsymbol{v}; \boldsymbol{x}, \boldsymbol{\Sigma})\big] \le 0\Big\}. \tag{12}$$

In words, $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma})$ is the maximal correlation of a random direction with a standard Gaussian vector $\boldsymbol{g}$ subject to $F(\boldsymbol{w}; \boldsymbol{x}, \boldsymbol{\Sigma})$ being nonpositive *on average*.

Properties of the Gaussian width. In the case $\boldsymbol{\Sigma} = \mathbf{I}_p$, $\mathcal{G}(\boldsymbol{x}, \mathbf{I}_p)$ depends on $\boldsymbol{x}$ only through $\varepsilon := \|\boldsymbol{x}\|_0/p$. Denote

$$\omega^*(\varepsilon) := \mathcal{G}_d(\boldsymbol{x}, \mathbf{I}_p) \quad \text{for any } \boldsymbol{x} \text{ with } \|\boldsymbol{x}\|_0/p = \varepsilon.$$

Indeed $\omega^*(\varepsilon)$ can be computed explicitly, and is given in parametric form by

$$\omega^*(\varepsilon)^2 = \varepsilon + 2(1-\varepsilon)\Phi(-\alpha),$$

$$\text{where } \alpha \text{ satisfies } \varepsilon = \frac{2[\varphi(\alpha) - \alpha\Phi(-\alpha)]}{\alpha + 2[\varphi(\alpha) - \alpha\Phi(-\alpha)]}.$$

Here, $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the standard Gaussian density, and $\Phi(x) = \int_{-\infty}^{x}\varphi(t)\,\mathrm{d}t$ is the Gaussian cumulative distribution function. One can show that $\omega^*(\varepsilon)$ is increasing and continuous in ε, goes to 1 as $\varepsilon \to 1$, and satisfies

$$\omega^*(\varepsilon) = \big(1 + o(\varepsilon)\big)\sqrt{2\varepsilon\log(1/\varepsilon)}.$$

Thus, $n/p \ge \mathcal{G}(\mathrm{sign}(\boldsymbol{\theta}^*), \mathbf{I}_p)^2$ is equivalent to $2(1 + o(s/p))s\log(p/s)/n \le 1$.

For general Gaussian designs $\boldsymbol{\Sigma}$, the critical sampling rate depends not only on the sparsity of $\boldsymbol{\theta}^*$ but also on the location and sign of its active coordinates. However, the value of $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma})$ changes at most by a factor equal to the condition number of $\boldsymbol{\Sigma}$, as stated in the next lemma.

LEMMA 1. *Assume that* $\mathbf{\Sigma}$ *has condition number upper bounded by* κ_{cond}. *Then for any* $\boldsymbol{x} \in \{-1, 0, 1\}^p$,

$$\kappa_{\mathrm{cond}}^{-1/2} \cdot \omega^*\big(\|\boldsymbol{x}\|_0/p\big) \le \mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma}) \le \kappa_{\mathrm{cond}}^{1/2} \cdot \omega^*\big(\|\boldsymbol{x}\|_0/p\big).$$

In particular, if $2(1 + o(s/p))s \log(p/s)/n \le \kappa_{\mathrm{cond}}^{-1}$, *then* $n/p \ge \mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma})^2$.

We prove Lemma 1 in Appendix C.2.

The definitions (12) and (11) immediately imply $\mathcal{G}_d(\boldsymbol{x}, \mathbf{\Sigma}) \le \mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma})$. The next lemma establishes that the two definitions of Gaussian width differ by a factor that is often negligible.

PROPOSITION 2. *For* c' *depending only on* κ_{cond}, *we have*

$$\mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma}) - c' \min\left(\frac{\sqrt{p}}{s}; \sqrt{\frac{s}{p}\log(p/s)}\right) \le \mathcal{G}_d(\boldsymbol{x}, \mathbf{\Sigma}) \le \mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma}),$$

where $s = \|\boldsymbol{x}\|_0$.

We prove Proposition 2 in Section C.

For designs with a bounded condition number, $\mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma})^2 \asymp (s/p)\log(p/s)$; cf. Lemma 1. Comparing with the lower bound in Proposition 2, we obtain that the difference between $\mathcal{G}_d(\boldsymbol{x}, \mathbf{\Sigma})$ and $\mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma})$ is negligible provided $s \gg p^{2/3}/(\log p)^{1/3}$.

For sublinear sparsity $s = o(p)$, we do not expect the bound of Proposition 2 to be tight. Because the results in this paper provide nontrivial control of the Lasso and debiased Lasso estimates for sampling rates n/p of order 1 (see parameter $\Delta_{\min}$ in Assumption (A1)(d) below), we do not pursue a more careful comparison of the standard and functional Gaussian widths for sublinear sparsities here. Indeed, under sublinear sparsity, any sampling rate of order 1 is well above the Donoho–Tanner phase transition.

Assumptions. We are ready to formally state the assumptions, which will hold throughout the paper. The distribution of the random design $\boldsymbol{X}$, response vector $\boldsymbol{y}$ and Lasso estimate $\widehat{\boldsymbol{\theta}}$ is determined by the tuple $(\boldsymbol{\theta}^*, \mathbf{\Sigma}, \sigma, \lambda)$, the number of samples n and the dimensionality p. Our results hold uniformly over choices of $(\boldsymbol{\theta}^*, \mathbf{\Sigma}, \sigma, \lambda)$ and sampling rates n/p that satisfy the following conditions:

(A1) There exist $0 < \lambda_{\min} \le \lambda_{\max} < \infty$, $0 < \kappa_{\min} \le \kappa_{\max} < \infty$, and $0 < \sigma_{\min} \le \sigma_{\max} < \infty$, $M < \infty$, $\Delta_{\min} \in (0, 1)$ such that
- (a) The Lasso regularization parameter λ is bounded $\lambda_{\min} \le \lambda \le \lambda_{\max}$.
- (b) The singular values $\kappa_j(\mathbf{\Sigma})$ of the population covariance $\mathbf{\Sigma}$ are bounded $\kappa_{\min} \le \kappa_j(\mathbf{\Sigma}) \le \kappa_{\max}$ for all j. We define $\kappa_{\mathrm{cond}} := \kappa_{\max}/\kappa_{\min} \ge 1$.
- (c) The noise variance σ^2 is bounded $\sigma_{\min}^2 \le \sigma^2 \le \sigma_{\max}^2$.
- (d) There exists $\bar{\boldsymbol{\theta}}^* \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}^*\|_1/p \le M/\sqrt{n}$ and

$$\frac{n}{p} \ge \mathcal{G}\big(\mathrm{sign}(\bar{\boldsymbol{\theta}}^*), \mathbf{\Sigma}\big)^2 + \Delta_{\min}.$$

We denote the collections of constants appearing in assumptions (A1) by

$$\mathcal{P}_{\mathrm{model}} := (\lambda_{\min}, \lambda_{\max}, \kappa_{\min}, \kappa_{\max}, \sigma_{\min}, \sigma_{\max}, \Delta_{\min}, M). \tag{13}$$

The choice of the constants $\mathcal{P}_{\mathrm{model}}$ determines via Assumption (A1) the space of parameters $(\boldsymbol{\theta}^*, \mathbf{\Sigma}, \sigma, \lambda)$ and sampling rates n/p (the uniformity class) within which the results stated below apply. With a slight abuse of language, we will occasionally use $\mathcal{P}_{\mathrm{model}}$ to refer to the uniformity class as well.

Assumption (A1)(d) can be viewed as an approximate sparsity condition: $\boldsymbol{\theta}^*$ is approximated in ℓ_1-norm by a vector $\bar{\boldsymbol{\theta}}^*$ whose sparsity places it above the Donoho–Tanner phase transition. As established in the next proposition, Assumption (A1)(d) is implied by existing popular notions of approximate sparsity which appear elsewhere in the Lasso literature.

PROPOSITION 3. *Assumption* (A1)(*d*) (*with the specified choice of* M) *is implied by any of the following*:

(a) *If* $\|\boldsymbol{\theta}^*\|_0 \le s$, *then Assumption* (A1)(*d*) *is satisfied with* $M = 0$ *if*

$$\kappa_{\text{cond}}^{1/2}\omega^*(s/p) \le 1 - \Delta_{\min}. \tag{14}$$

In particular, it suffices that

$$2\kappa_{\text{cond}}\big(1 + o(s/p)\big)s\log(p/s)/n \le (1 + \Delta_{\min})^{-1}. \tag{15}$$

(b) *If* $\boldsymbol{\theta}^* \in \mathsf{B}_q(\nu)$ *for* $q, \nu > 0$, *then Assumption* (A1)(*d*) *is satisfied by taking* $M = \sqrt{n}\nu(1 - s/p)/p^{1/q}$ *for any* s *satisfying equation* (14) *or equation* (15).

(c) *If* $\sum_{j=1}^{p} \min(1, \sqrt{n}|\theta_j^*|/a_0) \le s$ *for a certain* a_0, *then Assumption* (A1)(*d*) *is satisfied with* $M = a_0 s/p$ *provided equation* (14) *or equation* (15) *is satisfied.*

Proposition 3 follows from Lemma 1. Its proof is given in Appendix D.2.

In words, Assumption (A1)(d) allows $\boldsymbol{\theta}^*$ to be unbounded on a certain signed support, and requires that it be small in ℓ_1-norm on its remaining coordinates. Here "small" means $O(1/\sqrt{n})$ per coordinate on average, with leading constant given by M. The location and sign of the coordinates on which $\boldsymbol{\theta}^*$ can be unbounded is determined by the Gaussian width $\mathcal{G}(\boldsymbol{\Sigma}, \boldsymbol{x})$ of the corresponding vector $\boldsymbol{x}$. Assumption (A1)(d) permits that the number of coordinates in which $\boldsymbol{\theta}^*$ is unbounded is proportional to p, but does not allow for arbitrarily large proportionality constant. For example, as is clear from Proposition 3, we require at least that $s \le n$, and in fact will require something stronger than this.

Proposition 3 uses Lemma 1 to bound $\mathcal{G}(\boldsymbol{\Sigma}, \boldsymbol{x})$ with a suitable $\boldsymbol{x} = \operatorname{sign}(\bar{\boldsymbol{\theta}}^*)$. Since Lemma 1 is loose in general, the sufficient notions of approximate sparsity in Proposition 3 are not sharp and do not identify the whole domain of validity of our results. In contrast, Assumption (A1)(d) will imply that our results hold down to the Donoho–Tanner phase transition for a good ℓ_1-approximation of $\boldsymbol{\theta}^*$.

4. Main results. We now turn to the statement of our main results and a discussion of some of their consequences. The proof details are deferred to the Appendix.

4.1. *Control of the fixed-point parameters.* Each of our results involves a comparison of the Lasso or debiased Lasso estimators in the random- and fixed-design models. The comparison will be valid provided we choose τ, ζ to be the solution to the fixed-point equations (9a) and (9b). This solution we call τ^*, ζ^*. The next lemma establishes that the solution is unique, and satisfies uniform bounds under Assumption (A1).

LEMMA 4. *If* $\boldsymbol{\Sigma}$ *is invertible and* $\sigma^2 > 0$, *then equations* (9a) *and* (9b) *have a unique solution* τ^*, ζ^*. *Under Assumption* (A1), *there exists* $\tau_{\max} < \infty$ *and* $\zeta_{\min} > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* δ *such that* $\sigma^2 \le \tau^{*2} \le \tau_{\max}^2$ *and* $\zeta_{\min} \le \zeta^* \le 1$.

We prove Lemma 4 in Appendix C. An important consequence of Lemma 4 is that, due to the fixed-point equations (9a) and (9b), the quantity $\mathsf{R}(\tau^{*2}, \zeta^*)$ is bounded above by $\tau_{\max}^2 - \sigma^2$ and the quantity $\mathsf{df}(\tau^{*2}, \zeta^*)/n$ is bounded away from 1 by $1 - \zeta_{\min}$. As we will see (and as

described in Section 3), $\mathsf{R}(\tau^{*2}, \zeta^*)$ and $\mathsf{df}(\tau^{*2}, \zeta^*)$ are good approximations of the prediction risk $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ and the degrees-of-freedom $\|\widehat{\boldsymbol{\theta}}\|_0$ of the Lasso estimator in the random-design model (1). Thus, Lemma 4, in addition to being a technical tool, which shall be used repeatedly in our proofs, has substantive consequences on the behavior of the Lasso: under an arbitrarily small separation from the Donoho–Tanner phase transition, it gives nontrivial upper bounds on the Lasso prediction error and model size.

REMARK 4.1. The challenge in proving Lemma 4 lies in the fact that τ^*, ζ^* are implicitly defined as the solutions to the fixed-point equations (9a) and (9b). While in the case of i.i.d. Gaussian designs, one can exploit the explicit analytic formulas for $\mathsf{R}(\tau^2, \zeta)$ and $\mathsf{df}(\tau^2, \zeta)$ as in [45], no such formulas are available under correlated designs. Thus, we resort to a novel argument based on viewing equations (9a) and (9b) as KKT conditions for an infinite-dimesional optimization problem defined in Section B (see also Section C). The Gaussian width plays a central and natural role in the analysis of this optimization problem. Restricted eigenvalues or similar ideas do not yield a tight analysis of this optimization problem.

For the remainder of the document, we always assume $\widehat{\boldsymbol{\theta}}^f$ and $\widehat{\boldsymbol{\theta}}^{f,\mathrm{d}}$ are computed with parameters τ^*, ζ^*.

4.2. *Control of the Lasso estimate.* Our first result states that the random-design Lasso behaves like the fixed-design Lasso from the point of view of Lipschitz test functions. The proof of this result is deferred to Section C.7.

THEOREM 5. *Assume* (A1) *holds. Then there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\mathrm{model}}$ *and* δ *such that the following holds*: *if* $n \geq \sqrt{2}/\Delta_{\min}$, *then for any* 1*-Lipschitz function* $\phi : \mathbb{R}^p \to \mathbb{R}$ *we have for all* $\epsilon < c'$,

$$\mathbb{P}\big(\exists \lambda \in [\lambda_{\min}, \lambda_{\max}],\ \big|\phi(\widehat{\boldsymbol{\theta}}) - \mathbb{E}[\phi(\widehat{\boldsymbol{\theta}}^f)]\big| > \epsilon\big) \leq \frac{C}{\epsilon^4} e^{-cp\epsilon^4}.$$

Here, $\widehat{\boldsymbol{\theta}}^f$ *is the fixed-design Lasso with* τ^*, ζ^* *solving equations* (9a) *and* (9b).

The proof of this theorem is presented in Section C.8.

Theorem 5 has an obvious corollary which we spell out for future reference. For any fixed $\lambda \in [\lambda_{\min}, \lambda_{\max}]$,

$$\mathbb{P}\big(\big|\phi(\widehat{\boldsymbol{\theta}}) - \mathbb{E}[\phi(\widehat{\boldsymbol{\theta}}^f)]\big| > \epsilon\big) \leq \frac{C}{\epsilon^4} e^{-cp\epsilon^4}. \tag{16}$$

Namely, any Lipschitz function of the Lasso estimate concentrates around its expectation in the fixed-design model with high probability—provided that the sampling rate exceeds the Donoho–Tanner phase transition for a good ℓ_1 approximation of $\boldsymbol{\theta}^*$ and p is large. In particular, this concentration holds true even in the case where the sparsity s and dimension p are proportional to n, although the proportionality constants cannot be arbitrary.

We make note that since $\boldsymbol{\theta}^*$ is deterministic, ϕ may depend implicitly on $\boldsymbol{\theta}^*$. In particular, Theorem 5 applies, for example, to the estimation error and prediction error by taking $\phi(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$ and $\phi(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$, respectively. (In the latter case, the constants must be adjusted to account for the fact that $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$ does not have Lipschitz constant equal to 1. The adjustment is by at most constant factors because the Lipschitz constant is bounded under (A1).) Thus, the ℓ_2-estimation error and the prediction error concentrate on their expectations in the fixed-design models. By equation (9a), this implies that the prediction error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ concentrates on $\mathsf{R}(\tau^*, \zeta^*) = \tau^{*2} - \sigma^2$.

Comparison with earlier results. It is worth comparing this result to the existing fixed-design results for the Lasso (e.g., [6, 12, 13, 48]). To be definite, we consider ℓ_q-estimation error for $1 \le q \le 2$. The optimal fixed-design results establish the existence of constants $c, C > 0$ such that

$$\lambda \ge c\sqrt{\log(2ep/s)} \quad \Rightarrow \quad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_q \le C\frac{s^{1/q}\lambda}{\mathrm{RE}^2\sqrt{n}}, \tag{17}$$

where RE is an appropriate restricted eigenvalue of $\boldsymbol{X}$ (see [6] for precise statements), and C may depend on q.

Consider the proportional sparsity regime $s = \Omega(p)$, which is our focus in the present paper. We make the following comparisons.

Regularization parameter. When s is proportional to p, $c\sqrt{\log(2ep/s)}$ is of order one, so that $\lambda \ge c\sqrt{\log(2ep/s)}$ implies Assumption (A1)(d). On the other hand, Assumption (A1)(d) permits smaller regularization parameters than are permitted by [6], since $\lambda_{\min}$ in Assumption (A1)(d) can be arbitrarily small (but nonvanishing as $n, p, s \to \infty$), while c in equation (17) and [6] is a fixed numerical constant bounded away from 0. The case when λ is taken to be exactly zero is considered in recent works (see, e.g., [43]).

Estimation error. Because $\boldsymbol{\theta} \mapsto \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_q / p^{1/q-1/2}$ is 1-Lipschitz, we can apply Theorem 5. Further using the bound on τ^* from Lemma 4, one can show that $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q] = O(p^{1/q}/n^{1/2})$ under Assumption (A1), where O hides constants depending on $\mathcal{P}_{\mathrm{model}}$ (see (13)). Summarizing, we obtain with probability at least $1 - p^{-A}$ for any constant A,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_q = \mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q] + O(p^{1/q-3/4}\log(p)), \tag{18}$$

$$\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q] = O(p^{1/q}/n^{1/2}).$$

In the present setting, $p^{1/q}/n^{1/2}$ is of the same order as $Cs^{1/q}\lambda/(\mathrm{RE}^2\sqrt{n})$, so that the estimate is consistent with the results of [6]. If in addition $n = \tilde{o}(p^{3/2})$, then the error term in equation (18) is much smaller than $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q]$. In other words, we obtain a more precise concentration around a deterministic theoretical prediction, which we characterize.

Restricted eigenvalues and sampling rates. The previous bullet point describes a scenario in which the restricted eigenvalue RE is of order 1 (and, in particular, is bounded away from 0). In the random-design setting, this implicitly corresponds to an assumption on the number of samples. In Section 4.7, we show that restricted eigenvalues can be 0 for $n/p \ge (1 + \varepsilon)\mathcal{G}(\boldsymbol{\Sigma}, \boldsymbol{x})$ with ε a positive constant. Our results provide precise control in an interval of sampling rates that is excluded by [6] and related work [12, 13, 48].

Exact characterization. By establishing that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_q$ concentrates on $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q]$, Theorem 5 establishes upper and lower bounds on the risk that hold pointwise with respect to $\boldsymbol{\theta}^*$ and match up to negligible errors. It is a promising research direction to analyze $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_q]$ for specific correlation structures $\boldsymbol{\Sigma}$ (e.g., block diagonal or low-rank plus identity).

Theorem 5 and the later results in this paper can be used to design estimators for τ^*, ζ^*, derive the distribution of the debiased Lasso, and construct confidence intervals for single coordinates. A recent example of this strategy was given in [17] in a different setting. These exact concentration results are inaccessible from existing results like those in [6, 12, 13, 48], which are loose in their leading constants.

REMARK 4.2. Although we assume that the error $\boldsymbol{z}$ in the linear model is Gaussian with independent components, this assumption is not necessary, and Theorem 5 holds provided that $\|\boldsymbol{z}\|_2/\sqrt{n}$ concentrates on σ (the rate of this concentration may affect the right-hand side of equation (16)). This results from the rotational invariance of the ℓ_2-norm. In settings

similar to ours, the extension to non-Gaussian noise is common in the literature (see, e.g., [16]). We choose to develop theory with Gaussian noise to simplify the exposition and proofs.

REMARK 4.3. Up to logarithmic factors, Theorem 5 demonstrates a concentration at the rate $p^{-1/4}$. Such a rate is typical of results proved using Gordon's comparison inequality, which we use to derive all the results in this paper (see Section B). We suspect this rate is an artifact of our proof technique, and the correct rate should be $p^{-1/2}$. Recently, [41, 42] developed a nonasymptotic theory to analyze the approximate message passing algorithm, which offers another possible path to improve upon the current rate.

At a high level, the source of the rate appearing in Theorem 5 is as follows. Gordon's proof technique allows us to localize $\widehat{\boldsymbol{\theta}}$ within a region across which the growth of the objective value exceeds the size of its typical fluctuations. The size of the typical fluctuations are $O_{\mathrm{p}}(n^{-1/2})$, and as a function of distance r from the minimizer, we expect to growth to be $O_{\mathrm{p}}(r^2)$. Thus, we get the rate $n^{-1/4}$. This rate appears again in Theorems 5 and 7. Theorem C.11, Theorem 10 and Corollary 11 require approximations, which degrade the rate further. We expect that here, too, the rate appearing in the theorem is not optimal.

Simultaneous control over λ. So far, we only discussed the consequences of Theorem 5 for a fixed value of λ, namely equation (16). However, Theorem 5 establishes a characterization which holds simultaneously over all λ in a bounded interval $[\lambda_{\min}, \lambda_{\max}]$. This is particularly useful to analyze adaptive procedures to select λ.

In particular, it implies that with high probability the minimum estimation error over choices of $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, is nearly achieved at a deterministic value λ_*. Namely, writing $\widehat{\boldsymbol{\theta}}_\lambda$ and $\widehat{\boldsymbol{\theta}}^f_\lambda$ for the Lasso estimator and fixed-design estimator at regularization λ, we have

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{\theta}}_{\lambda_*} - \boldsymbol{\theta}^*\|_2 - \min_{\lambda\in[\lambda_{\min},\lambda_{\max}]}\frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_2\right| > \epsilon\right) \le \frac{C}{\epsilon^4}e^{-cp\epsilon^4},$$

$$\text{for } \lambda_* := \operatorname*{arg\,min}_{\lambda\in[\lambda_{\min},\lambda_{\max}]}\frac{1}{\sqrt{p}}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f_\lambda - \boldsymbol{\theta}^*\|_2].$$

Recall that it is standard to choose λ on the order of $\sqrt{\log(p/s)}$ (see, e.g., [6]). As we have already described, applying existing fixed-design analysis to the current setting where s is proportional to p requires taking $\lambda_{\min} \ge c > 0$ for an explicit constant c that is bounded away from 0. As shown in [45], choosing λ based on such conservative lower bounds can be suboptimal by a large factor. By allowing $\lambda_{\min}$ to be arbitrarily close to 0, our results can capture the full range of regularization parameters on which the Lasso behaves well.

Control of the empirical distribution. Previous work on i.i.d. covariates has mainly focused on establishing the convergence of the joint empirical distribution of the coordinates of the Lasso estimator and the true parameter vector:

$$\hat{\mu}_{n,p} := \frac{1}{p}\sum_{i=1}^{p}\delta_{\sqrt{n}\theta^*_i,\sqrt{n}\widehat{\theta}_i},$$

to a limiting distribution either weakly or in Wasserstein distance [4, 45]. When covariates are i.i.d., the behavior of $\hat{\mu}_{n,p}$ captures all nontrivial behavior of the distribution of $\widehat{\boldsymbol{\theta}}$: indeed, the exchangeability of the model implies that conditional on $\hat{\mu}_{n,p}$, the distribution of $\widehat{\boldsymbol{\theta}}$ is uniform over permutations of the coordinates, which map each coordinate of $\boldsymbol{\theta}^*$ to a coordinate with the same value. This is no longer the case for correlated covariates, and Theorems 5 capture this this additional structure.

Nevertheless, the empirical distribution $\hat{\mu}_{n,p}$ may be of interest, in part because it is easily interpretable. By applying Theorem 5 to several test functions at once, we can establish concentration of the empirical distribution simultaneously in λ. We use a particular metrization of the weak topology[3] on the space of probability measures on $\mathbb{R}^2$, namely

$$d_{w^*}(\mu,\nu)=\sum_{k=1}^{\infty}2^{-k}\left|\mathbb{E}_{\boldsymbol{A}\sim\mu}\left[\phi_k(\boldsymbol{A})\right]-\mathbb{E}_{\boldsymbol{B}\sim\nu}\left[\phi_k(\boldsymbol{B})\right]\right|.$$

Here, $\{\phi_k\}$ denotes a countable subset of the 1-Lipschitz functions $\mathbb{R}^2\to\mathbb{R}$ such that for any compact set $K\subset\mathbb{R}^2$, $\{\phi_k|_K\}$ is dense with respect to the ℓ_∞-norm.

COROLLARY 6. *Assume Assumption* (A1) *and additionally that* $n/p\le\Delta_{\max}$. *There exists* μ_**—a probability distribution on* $\mathbb{R}^2$*—and constants* $C,C',c>0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* $\Delta_{\max}$ *such that*

$$\mathbb{P}\left(\exists\lambda\in[\lambda_{\min},\lambda_{\max}],\, d_{w^*}\left(\frac{1}{p}\sum_{i=1}^{p}\delta_{\sqrt{n}\theta_i^*,\sqrt{n}\widehat{\theta}_i},\mu_*\right)\ge\frac{C'}{\sqrt{p}}+\epsilon\right)\le\frac{C}{\epsilon^4}e^{-cn\epsilon^4},$$

and

$$\mathbb{P}\left(\exists\lambda\in[\lambda_{\min},\lambda_{\max}],\, d_{w^*}\left(\frac{1}{p}\sum_{i=1}^{p}\delta_{\sqrt{n}\theta_i^*,\sqrt{n}\widehat{\theta}_i^{\,f}},\mu_*\right)\ge\frac{C'}{\sqrt{p}}+\epsilon\right)\le 2e^{-cn\epsilon^2}.$$

Corollary 6 states that in both the random-design model and the fixed-design model, the joint empirical distribution of the estimate and the true parameter concentrates with respect to weak-$*$ distance, and that moreover, they concentrate on the same value. Using Theorem 5, one can also control properties of μ_* such as its second moments in terms of $\mathcal{P}_{\text{model}}$. We prove Corollary 6 in Appendix C.15.

REMARK 4.4. The proof of Theorem 5 is similar to the proof of Theorem 3.1 of [45] in the i.i.d. design case. The proof of simultaneous control over λ (Theorem 5) and the control of the Lasso residual (Theorem 7), stated below, are similar to the proofs of analogous results in [45]. We emphasize, however, that these proofs rely heavily on the boundedness and uniqueness of the fixed-point parameters τ^* and ζ^* (see Lemma 4). Regarding the Lasso estimate, establishing these properties of the fixed-design characterization is the main technical innovation of the present paper (see Remark 4.1). Below we will see that further technical innovations are required for analyzing the Lasso sparsity and the debiased Lasso.

Note that the ϵ^4 appearing in the exponent in Theorem 5 is faster than the rate appearing in Theorem 3.1 of [45]. This is because [45] provide a good approximation of the empirical distribution of the coordinates of $\widehat{\boldsymbol{\theta}}$ in Wasserstein metric, which is a more complex object to control than a single Lipschitz function (see [45], Proposition F.2). Corollary 6 controls the empirical distribution of the coordinates of $\widehat{\boldsymbol{\theta}}$, but in a metric, which is weaker than the Wasserstein metric.

4.3. *Control of the Lasso residual.* In this section, we establish control for the residual of the Lasso estimator. The behavior of this residual is of interest because it can be used in estimators of important quantities. For example, we shall use it to construct an empirical estimate $\widehat{\tau}$ of τ^*. Informally, the Lasso residual behaves like a normally distributed random vector with zero mean and covariance $(\tau^*\zeta^*)^2\mathbf{I}_n$.

[3]The metric d_{w^*} metrizes weak convergence in the sense that $\mu_i\xrightarrow{\text{d}}\mu$ if and only if $d_{w^*}(\mu_i,\mu)\to 0$.

THEOREM 7. *Under Assumption* (A1), *there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* δ *such that for any* 1*-Lipschitz function* $\phi : \mathbb{R}^p \to \mathbb{R}$, *we have for all* $\epsilon < c'$,

$$\mathbb{P}\left(\left|\phi\left(\frac{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}}{\sqrt{n}}\right) - \mathbb{E}\left[\phi\left(\frac{\tau^*\zeta^*\boldsymbol{h}}{\sqrt{n}}\right)\right]\right| > \epsilon\right) \le \frac{C}{\epsilon^2} e^{-cn\epsilon^4},$$

where $\boldsymbol{h} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_n)$. *Consequently*,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}} - \tau^*\zeta^*\right| > \epsilon\right) \le \frac{C}{\epsilon^2} e^{-cn\epsilon^4}.$$

The proof of Theorem 7 is provided in Section C.9.

4.4. *Control of the Lasso sparsity.* This section characterizes the sparsity of the Lasso estimator. In particular, we show that the number of selected parameters per observation $\|\widehat{\boldsymbol{\theta}}\|_0/n$ concentrates on $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f\|_0]/n = 1 - \zeta^*$.

THEOREM 8. *Under Assumption* (A1), *there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* δ *such that for all* $\epsilon < c'$,

$$\mathbb{P}\left(\left|\frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n} - (1 - \zeta^*)\right| > \epsilon\right) \le \frac{C}{\epsilon^3} e^{-cn\epsilon^6}.$$

The proof of this result is presented in Section C.11.

Note that the ϵ^6 in the exponent in Theorem 8 is worse than the ϵ^4 appearing in the exponent in Theorem 5, Theorem 5, Corollary 6 and Theorem 7. This is because the function $\|\widehat{\boldsymbol{\theta}}\|_0/n$ is not a Lipschitz function. The proof involves instead analyzing the subgradient of the ℓ_1 penalty at the Lasso solution and applying certain Lipschitz approximations for indicator functions. Because the Lipschitz constants diverge as $\epsilon \to 0$, this results in a weaker probability bound (see Section C.11 for details). We suspect this rate is not tight, and a dependence of ϵ^2 may be possible, but proving such a tighter dependence may require new tools. The estimators in the coming sections, which involve $\|\widehat{\boldsymbol{\theta}}\|_0/n$, will also suffer this degraded rate.

We make a note that recently Bellec and Zhang [8], Section 3.4, establish that $\frac{1}{n}\|\widehat{\boldsymbol{\theta}}\|_0 | \boldsymbol{X}$ concentrates around its expectation with deviations of order $O(n^{-1/2})$ using the second-order Stein's formula. Our result is different and complementary in that it shows that $\frac{1}{n}\|\widehat{\boldsymbol{\theta}}\|_0$ has large-deviation probabilities (w.r.t. randomness of both the noise and the design), which decay exponentially, and characterizes the value around which it concentrates. Moreover, our result also implies that under Assumption (A1) (and, in particular, above the Donoho–Tanner phase transition), the value on which $\frac{1}{n}\|\widehat{\boldsymbol{\theta}}\|_0$ concentrates is uniformly bounded away from 1.

REMARK 4.5. The proof of Theorem 8 is fundamentally different from the proof of the analogous result for i.i.d. designs [45], Theorem F.1. Indeed, the proof of [45], Theorem F.1, draws heavily on simple expression for the empirical distribution of the coordinates of $\widehat{\boldsymbol{\theta}}$ and of the subgraident of the ℓ_1-norm at the Lasso solution. For general covariances, such simple expressions are unavailable due to the nonexchangeability of the model. See Section C.11 for details.

Prediction error and hyperparameter tuning. Using Theorem 7 and 8, we can construct an estimator $\widehat{\tau}$ of τ^*. This gives rise to a provably optimal method for parameter tuning and a consistent estimate of the standard error of the debiased Lasso, which can be used to construct confidence intervals. In particular, Theorem 7 shows that the residuals $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}$ are

approximately $\mathsf{N}(0, (\tau^*\zeta^*)^2\mathbf{I}_n)$ and, moreover, that $\|\boldsymbol{\theta}^*\|_0/n$ concentrates on $1-\zeta^*$. Thus, the parameters τ^* is consistently estimated by

$$\widehat{\tau}(\lambda) := \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)}. \tag{19}$$

Since τ^* controls the noise in the fixed design model, its estimation is of particular interest. Indeed, because $\tau^{*2} = \sigma^2 + \mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2]$ and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2$ concentrates on $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2]$, $\widehat{\tau}(\lambda)^2$ concentrates, up to an additive constant, which does not depend on λ, on the prediction error. Because of their importance, we collect these facts in the next theorem.

THEOREM 9. *Under Assumption* (A1), *let* $\tau^* = \tau^*(\lambda)$ *be the unique solution of the system of equations* (9a), (9b). *Then there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* δ *such that for all* $\epsilon < c'$,

$$\mathbb{P}\big(\exists\lambda \in [\lambda_{\min}, \lambda_{\max}], |\widehat{\tau}(\lambda) - \tau^*(\lambda)| \geq \epsilon\big) \leq \frac{C}{\epsilon^6} e^{-cn\epsilon^6}.$$

Further defining $\hat{\lambda} := \arg\min\{\widehat{\tau}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$, *we have*

$$\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\theta}}_{\hat{\lambda}} - \boldsymbol{\theta}^*\big\|_{\boldsymbol{\Sigma}}^2 \geq \min_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \big\|\widehat{\boldsymbol{\theta}}_{\lambda} - \boldsymbol{\theta}^*\big\|_{\boldsymbol{\Sigma}}^2 + \epsilon\Big) \leq \frac{C}{\epsilon^6} e^{-cn\epsilon^6}.$$

Thus, minimizing $\widehat{\tau}(\lambda)^2$ over λ gives a provably optimal parameter tuning method. Importantly, $\widehat{\tau}(\lambda)$ does not depend on any unknown model parameters, namely σ, $\boldsymbol{\Sigma}$ or $\boldsymbol{\theta}^*$. It was already observed in [45] that minimizing $\widehat{\tau}(\lambda)$ over λ provides a good selection procedure for the regularization parameter. Our results provide theoretical support for this approach under general Gaussian designs. After the current paper was posted, similar results were recently obtained for a wide class of losses and penalties in [7].

4.5. *Control of the debiased Lasso.* Recall that the debiased Lasso with degrees-of-freedom adjustment is defined according to expression (4),

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \widehat{\boldsymbol{\theta}} + \frac{1}{n - \|\widehat{\boldsymbol{\theta}}\|_0} \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}).$$

The next theorem establishes that the debiased Lasso behaves like the debiased Lasso in the fixed-design model $\widehat{\boldsymbol{\theta}}^{f,\mathrm{d}}$ (defined in equation (7)), which follows a Gaussian distribution with mean $\boldsymbol{\theta}^*$ and covariance $\tau^{*2}\boldsymbol{\Sigma}^{-1}/n$. The proof of this result is provided in Section C.13.

THEOREM 10. *Under Assumption* (A1), *there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* δ *such that for any* 1*-Lipschitz* $\phi : \mathbb{R}^p \to \mathbb{R}$, *we have for all* $\epsilon < c'$,

$$\mathbb{P}\big(|\phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}}) - \mathbb{E}[\phi(\widehat{\boldsymbol{\theta}}^{f,\mathrm{d}})]| > \epsilon\big) \leq \frac{C}{\epsilon^3} e^{-cp\epsilon^6},$$

where $\boldsymbol{g} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$.

Note that the rate of convergence obtained here is faster than the one appearing in Theorem 3.3 of [45] in the case of i.i.d. Gaussian designs. The results, however, are not directly comparable, since [45], Theorem 3.3, controls the empirical distribution of the coordinates of $\widehat{\boldsymbol{\theta}}$ in Wasserstein distance, whereas we control a single Lipschitz function (see Remark 4.3 for a similar discussion). Further, our proof techniques differ substantially from that of [45]. While their results rely on a gluing argument (see Section F.2 of the Supplementary Material to [45]), we connect the debiased Lasso to a "smoothed Lasso" estimator (see Section C.13). In neither this paper nor in [45] do we expect the rates of concentration to be tight.

Confidence intervals using the debiased Lasso. Equipped with Theorem 10, one may construct confidence intervals for any individual coordinate of $\boldsymbol{\theta}^*$ with guaranteed coverage on average. Because τ^* is unknown, we use the estimator $\widehat{\tau}(\lambda)$ given by equation (3). We refer to the resulting intervals as the *debiased confidence intervals*.

COROLLARY 11. *Fix $q \in (0,1)$. For each $j \in [p]$, define the interval*

$$\mathsf{CI}_j^{\mathrm{d}} := \big[\widehat{\theta}_j^{\mathrm{d}} - \Sigma_{j|-j}^{-1/2}\widehat{\tau}(\lambda) z_{1-q/2}/\sqrt{n}, \widehat{\theta}_j^{\mathrm{d}} + \Sigma_{j|-j}^{-1/2}\widehat{\tau}(\lambda) z_{1-q/2}/\sqrt{n}\big],$$

where $z_{1-q/2}$ is the $(1-q/2)$-quantile of the standard normal distribution, $\widehat{\tau}(\lambda)$ is given by equation (3), and

$$\Sigma_{j|-j} = \Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}(\boldsymbol{\Sigma}_{-j,-j})^{-1}\boldsymbol{\Sigma}_{-j,j}.$$

Define the false-coverage proportion

$$\mathsf{FCP} := \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{\theta_j^* \notin \mathsf{CI}_j^{\mathrm{d}}}.$$

Under assumptions (A1) *and if $n/p \le \Delta_{\max} < \infty$, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$ and $\Delta_{\max}$ such that for all $\epsilon < c'$,*

$$\mathbb{P}\big(|\mathsf{FCP} - q| > \epsilon\big) \le \frac{C}{\epsilon^6} e^{-cn\epsilon^{12}}.$$

We prove Corollary 11 in Section C.13. Importantly, we are able to show that the debiased Lasso is successful, at least in the sense of Corollary 11, down to the Donoho–Tanner phase transition and allow λ to be arbitrarily close to zero (though not vanishing asymptotically).

As we have already described in Section 3, in the low-dimensional limit, which takes p fixed, $n \to \infty$, the asymptotic standard error of the OLS estimate for θ_j^* is given by $\Sigma_{j|-j}^{-1/2}\sigma/\sqrt{n}$. The first fixed-point equation (9a) states that we should inflate this standard error by replacing σ^2 with $\sigma^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$. By Lemma 4, we have that τ^* is $O(1)$. Thus, Theorem 10 shows above the Donoho–Tanner phase transition the debiased Lasso achieves the parametric $n^{-1/2}$ rate in most coordinates, with standard error inflated at most by a constant.

It is worth emphasizing that the debiasing construction of equation (4) assumes that the population covariance $\boldsymbol{\Sigma}$ is known. In practice, $\boldsymbol{\Sigma}$ often needs to be estimated from data. Replacing $\boldsymbol{\Sigma}$ with $\widehat{\boldsymbol{\Sigma}}$ in equation (4) introduces an error $(\boldsymbol{\Sigma}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1})\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})/(n - \|\widehat{\boldsymbol{\theta}}\|_0)$, which we can crudely bound as $O_p(\|\boldsymbol{\Sigma}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op}})$ (because under Assumption (A1) $\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})\|_2/(n - \|\widehat{\boldsymbol{\theta}}\|_0) = O_p(1)$). Operator norm consistency of $\widehat{\boldsymbol{\Sigma}}$ can be achieved under two scenarios: (i) when sufficiently strong information is known about the structure of $\boldsymbol{\Sigma}$ (for instance $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ are band diagonal or very sparse); see, for example, [11, 14, 29] and (ii) when additional "unlabeled" data $(\boldsymbol{x}_i')_{i\ge 1}$ is available. Alternatively, if one is interested in a particular coordinate j of $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$, one needs only to control the corresponding row of $\widehat{\boldsymbol{\Sigma}}^{-1}$, which can be achieved using, for example, the nodewise Lasso and sufficient sparsity conditions [38], Section 3.3.2. Finally, we remark that the recent paper [17] studies the problem of debiasing in a regime where the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ cannot be estimated well, although much about this difficult regime remains open.

REMARK 4.6. It is instructive to compare the degrees-of-freedom adjusted debiased Lasso of equation (4) with the more standard construction without adjustment [36–38, 57, 59]:

$$\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \frac{1}{n}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}).$$

The degrees-of-freedom adjustment adjusts the second term by a factor $1/(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)$. Intuitively, when the sparsity s is much smaller than n, this factor should be close to 1, and the two constructions $\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}}, \widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ should behave comparably. The paper [9] made this precise by showing that the impact of the adjustment on a single coordinate $\widehat{\theta}_{0j}^{\mathrm{d}}$ is $o_{\mathrm{p}}(n^{-1/2})$ provided $s = o(n^{2/3}/\log(p/s)^{1/3})$. For larger values of s, the impact of the adjustment on a single coordinate can be nonnegligble on the $n^{-1/2}$ scale, so becomes relevant for inference on a single coordinate (see the next section). In the proportional regime $s = \Theta(n)$, we can have $\|\widehat{\boldsymbol{\theta}}^{\mathrm{d}} - \widehat{\boldsymbol{\theta}}_0^{\mathrm{d}}\|_2 = \Theta(1)$, whence we expect the degrees-of-freedom adjustment to have a nonnegligible impact on all or almost all coordinates simultaneously. The degrees-of-freedom adjustment in equation (4) is crucial for Theorem 10 and Corollary 11.

4.6. *Inference on a single coordinate.* While Theorem 10 and Corollary 11 establish coverage of the debiased confidence intervals $\mathsf{CI}_j^{\mathrm{d}}$ on average across coordinates, they do not guarantee the coverage of $\mathsf{CI}_j^{\mathrm{d}}$ for a fixed j. To illustrate the problem, recall that Theorem 10 implies that for any 1-Lipschitz $\phi : \mathbb{R}^p \to \mathbb{R}$, we have with high probability $\phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}}) - \mathbb{E}[\phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}})] = \tilde{O}(p^{-1/6})$, where $\tilde{O}$ hides factors which only depend on $\mathcal{P}_{\mathrm{model}}$ and δ or are polylogarithmic in p. Applied to $\phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}}) = \widehat{\theta}_j^{\mathrm{d}} - \theta_j^*$, this implies that the difference $\sqrt{n}(\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*)$ lies with high probability in an interval of length $\tilde{O}(\sqrt{n}/p^{1/6})$. In contrast, Theorem 10 and Corollary 11 suggest that the typical fluctuations of $\sqrt{n}(\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*)$ are of order $O(1)$. Thus, the control of a single coordinate provided by Theorem 10 is at a larger scale than the scale of its typical fluctuations.

In fact, the naïve guess based on Theorem 10 that $\sqrt{n}\Sigma_{j|-j}^{1/2}(\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*) \sim \mathsf{N}(0, \tau^{*2})$ can be incorrect. For example, the recent paper [10] studies the distribution of a single coordinate of the debiased Lasso (and other penalized estimators), and establishes that $\sqrt{n}\Sigma_{j|-j}^{1/2}(\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*)/\tau^* \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$ for most, but not all, coordinates of the debiased Lasso. They show that the variance of $\sqrt{n}\Sigma_{j|-j}^{1/2}(\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*)$ is approximately given by (see equation (3.19) of [10])

$$\mathbb{E}\Big[\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)^2} + \frac{(\widehat{\theta}_j - \theta_j^*)^2}{1-\|\widehat{\boldsymbol{\theta}}\|_0/n}\Big].$$

In particular, the standard error estimate $\widehat{\tau}$ will be too small by a nonnegligible amount when $(\widehat{\theta}_j - \theta_j^*)^2/(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)$ does not vanish relative to $\hat{\tau}(\lambda)^2 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2/n(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)^2$. Under a proportional asymptotics, we have shown that both $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2/(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)$ and $\hat{\tau}(\lambda)^2$ are of order 1, which implies that for most coordinates, $(\widehat{\theta}_j - \theta_j^*)^2/(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)$ vanishes relative to $\hat{\tau}(\lambda)^2$. Nevertheless, there may exist a sublinear number of coordinates j for which $(\hat{\theta}_j - \theta_j^*)^2 = \Omega_{\mathrm{p}}(1)$ [9]. Note that this can occur even above the Donoho–Tanner phase transition or when restricted eigenvalue conditions are satisfied. For such coordinates, the standard error $\widehat{\tau}$ will be too small. The bounds $\max_j \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1$ used by [38] prohibit the existence of such coordinates, but need not hold under the Assumption (A1).

In Figure 1 of [10], the authors demonstrate a case in which $\widehat{\tau}$ systematically underestimates the variance of $\widehat{\theta}_j^{\mathrm{d}}$. For convenience, we also include a similar simulation here. Let $\boldsymbol{v} = (0, \mathbf{1}_s, \mathbf{0}_{p-s-1})/\sqrt{s}$. That is, $\boldsymbol{v}$ has unit ℓ_2-norm, sparsity s, and is constant on its active set. We take $s = 100$, $n = 500$, $p = 1000$, $\rho^2 = 0.75$, $\sigma = 1$, $\lambda = \sqrt{2\sigma\log(p/s)}$, $\boldsymbol{\theta}^* = 3\sqrt{s}\lambda\boldsymbol{v}$ and $\boldsymbol{\Sigma} = \mathbf{I}_p + \rho\boldsymbol{e}_1\boldsymbol{v}^\top + \rho\boldsymbol{v}\boldsymbol{e}_1^\top$. One can check that $\boldsymbol{\Sigma}$ is positive definite. For 5000 replications, we generate data from the model (2), fit the debiased Lasso estimate $\hat{\theta}_1^{\mathrm{d}}$, compute the

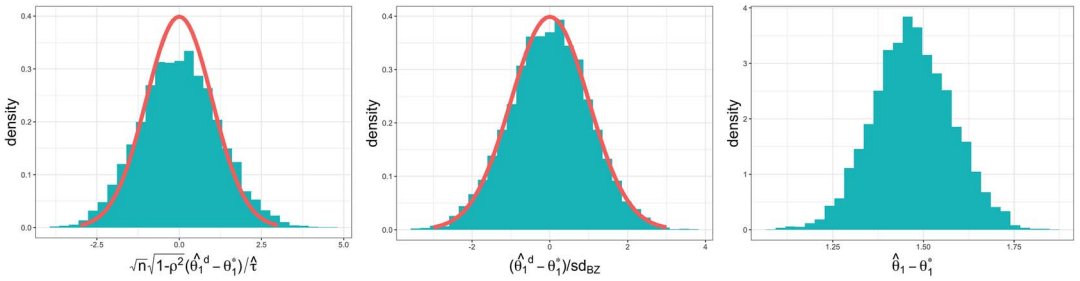

FIG. 2. *The debiased Lasso test statistic* $\widehat{\theta}_1^{\mathrm{d}}$ *for* $p = 1000$, $n = 500$, $s = 100$, $\rho^2 = .5$, $\sigma = 1$, $\lambda = \sqrt{2\sigma \log(p/s)/n} = .096$, $\boldsymbol{\theta}^* = 3\sqrt{s}\lambda \boldsymbol{v}$, *where* $\boldsymbol{v} = (0, \mathbf{1}_s, \mathbf{0}_{p-s-1})/\sqrt{s}$ *and* $\boldsymbol{\Sigma} = \mathbf{I}_p + \rho \boldsymbol{e}_1 \boldsymbol{v}^\top + \rho \boldsymbol{v} \boldsymbol{e}_1^\top$. *On the left, we plot a histogram of the debiased Lasso centered and normalized based on the effective noise* $\widehat{\tau}$ *and the theory in this paper, and we superimpose the standard normal density. In the center plot, we normalize instead by the standard deviation derived in* [10] (*see equation* (20)). *On the right, we plot a histograms of Lasso error* $\widehat{\theta}_1 - \theta_1^*$ *without centering or standarization, demonstrating that the error of the Lasso in this coordinate is* $O(1)$.

estimated standard error $\widehat{\tau}$ and compute

$$\mathrm{sd}_{\mathrm{BZ}}^2 := \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)^2} + \frac{(\widehat{\theta}_j - \theta_j^*)^2}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}. \tag{20}$$

In Figure 2, from left to right, we plot histograms of $\sqrt{1-\rho^2}\sqrt{n}(\widehat{\theta}_1^{\mathrm{d}} - \theta_1^*)/\widehat{\tau}$, $\sqrt{n}(\widehat{\theta}_1^{\mathrm{d}} - \theta_1^*)/\mathrm{sd}_{\mathrm{BZ}}$ and $\hat{\theta}_1 - \theta_1^*$. In the first two plots, we superimpose the standard normal density. In the left plot, we see an overdispersion of $\sqrt{1-\rho^2}\sqrt{n}(\widehat{\theta}_1^{\mathrm{d}} - \theta_1^*)/\widehat{\tau}$ relative to the normal density, which is no longer present when the errors are instead normalized by $\mathrm{sd}_{\mathrm{BZ}}$ in the second plot. This validates that for the first coordinate, $\sqrt{1-\rho^2}\widehat{\tau}/\sqrt{n}$ underestimates the standard error. The right-most histogram shows that the error $\hat{\theta}_1 - \theta_1^*$ is of order 1, whence the second term in $\mathrm{sd}_{\mathrm{BZ}}$ is nonnegligible. (Precisely, the standard division in this plot is about 2.2). We emphasize that $\mathrm{sd}_{\mathrm{BZ}}$ is not an empirical quantity. Our purpose is simply to display evidence that the standard error $\Sigma_{1|-1}^{-1/2}\widehat{\tau}$ is incorrect for the first coordinate. The paper [9] also provides an empirical standard error, which agrees with $\mathrm{sd}_{\mathrm{BZ}}$ to first order.

Figure 2 suggests the conjecture that while $\widehat{\theta}_1^{\mathrm{d}}$ may have standard error larger that τ^* in some coordinates, it is still approximately normally distributed and unbiased. We do not establish this fact, and as far as we know, establishing it remains open. We expect that completing this theory will require different techniques than those in the current work.

An alternative approach. In the current paper, we instead provide an alternative construction of confidence intervals for a single coordinate using a leave-one-out technique. We are able to establish the coordinatewise validity of these confidence intervals even in cases where the Lasso error $\widehat{\theta}_j - \theta_j^*$ is of order 1. We call these confidence intervals, defined below, the *leave-one-out confidence intervals*, denoted by $\mathsf{CI}_j^{\mathrm{loo}}$. According to simulation, the leave-one-out confidence intervals often approximately agree with the debiased confidence intervals, though for some coordinates they may have a larger or smaller width.

To facilitate the construction, let us write the observation vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \begin{pmatrix} \cdots & \check{\boldsymbol{x}}_j & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ \theta_j^* \\ \vdots \end{pmatrix} + \sigma \boldsymbol{z} = \theta_j^* \check{\boldsymbol{x}}_j + \boldsymbol{X}_{-j} \boldsymbol{\theta}^*{}_{-j} + \sigma \boldsymbol{z}, \tag{21}$$

where $\boldsymbol{X}_{-j} \in \mathbb{R}^{n\times(p-1)}$ denotes the original design matrix excluding the jth column and $\check{\boldsymbol{x}}_j$ denotes the jth column. Define $\check{\boldsymbol{x}}_j^{\perp} := \check{\boldsymbol{x}}_j - \boldsymbol{X}_{-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j} \in \mathbb{R}^n$ so that $\check{\boldsymbol{x}}_j^{\perp}$ is independent of $\boldsymbol{X}_{-j}$ (see Section D.3). Let $\widehat{\theta}_{j,\text{init}}$ be any deterministic real number that is chosen *a priori*; for instance, $\widehat{\theta}_{j,\text{init}}$ can be set as 0. According to decomposition (21),

$$\boldsymbol{y} - \check{\boldsymbol{x}}_j^{\perp}\widehat{\theta}_{j,\text{init}} = \boldsymbol{X}_{-j}\underbrace{\big(\boldsymbol{\theta}^*_{-j} + (\theta_j^* - \widehat{\theta}_{j,\text{init}})\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}\big)}_{=:\boldsymbol{\theta}^*_{\text{loo}}} + \check{\boldsymbol{x}}_j^{\perp}(\theta_j^* - \widehat{\theta}_{j,\text{init}}) + \sigma\boldsymbol{z}, \tag{22}$$

and

$$\check{\boldsymbol{x}}_j^{\perp}(\theta_j^* - \widehat{\theta}_{j,\text{init}}) + \sigma\boldsymbol{z} \sim \mathsf{N}(0, \sigma_{\text{loo}}^2\mathbf{I}_n) \quad \text{with } \sigma_{\text{loo}}^2 := \sigma^2 + \Sigma_{j|-j}(\theta_j^* - \widehat{\theta}_{j,\text{init}})^2.$$

Expression (22) can be viewed as defining a linear-model with $p-1$ covariates, with true parameter $\boldsymbol{\theta}^*_{\text{loo}}$, noise variance σ^2_{loo} and outcome $\boldsymbol{y} - \check{\boldsymbol{x}}_j^{\perp}\widehat{\theta}_{j,\text{init}}$. We call this the *leave-one-out model*, and call

$$\boldsymbol{y}_{\text{init}} := \boldsymbol{y} - \check{\boldsymbol{x}}_j^{\perp}\widehat{\theta}_{j,\text{init}}$$

the pseudo outcome. Let τ^*_{loo}, ζ^*_{loo} be the solution to the fixed-point equations (9a) and (9b) in the leave-one-out model, and $\widehat{\boldsymbol{\theta}}_{\text{loo}}$ be the Lasso fit on $\boldsymbol{y}_{\text{init}}$ to $\boldsymbol{X}_{-j}$.

The leave-one-out confidence interval is then constructed based on the variable importance statistic

$$\xi_j := \widehat{\theta}_{j,\text{init}} + \frac{(\check{\boldsymbol{x}}_j^{\perp})^{\top}(\boldsymbol{y}_{\text{init}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}})}{\Sigma_{j|-j}(n - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0)}. \tag{23}$$

Note the statistic ξ_j is a renormalized empirical correlation between residuals from two regressions: the population regression of feature j on the other features (i.e., $\check{\boldsymbol{x}}_j^{\perp}$), and a sample regression of the pseudo-outcome $\boldsymbol{y}_{\text{init}}$ on the other features (i.e., $\boldsymbol{y}_{\text{init}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}$). If $\widehat{\theta}_{j,\text{init}} = \theta_j^*$, these residuals will be independent. Indeed, in this case $\check{\boldsymbol{x}}_j^{\perp}$ is independent of $(\boldsymbol{y}_{\text{init}}, \boldsymbol{X}_{-j})$, and because $\widehat{\boldsymbol{\theta}}_{\text{loo}}$ is a function of $(\boldsymbol{y}_{\text{init}}, \boldsymbol{X}_{-j})$, $\check{\boldsymbol{x}}_j^{\perp}$ is also independent of $\boldsymbol{y}_{\text{init}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}$. In this case, the distribution of ξ_j is easy to understand. We will also quantify the distribution of the variable importance statistic ξ_j when $\widehat{\theta}_{j,\text{init}}$ is sufficiently close to θ_j^*, which will allow us to we estimate the effective confidence intervals.

Similar to $\widehat{\tau}(\lambda)$ defined in equation (3), we estimate the effective noise level in the leave-one-out model by

$$\widehat{\tau}^j_{\text{loo}} := \frac{\|\boldsymbol{y}_{\text{init}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_2}{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n)}.$$

The leave-one-out confidence interval is then defined as

$$\mathsf{CI}^{\text{loo}}_j := \big[\xi_j - \Sigma_{j|-j}^{-1/2}\widehat{\tau}^j_{\text{loo}}z_{1-\alpha/2}/\sqrt{n}, \xi_j + \Sigma_{j|-j}^{-1/2}\widehat{\tau}^j_{\text{loo}}z_{1-\alpha/2}/\sqrt{n}\big].$$

As asserted by the following result, this confidence interval $\mathsf{CI}^{\text{loo}}_j$ achieves approximate coverage for fixed j provided $\widehat{\theta}_{j,\text{init}} - \theta_j^* = o(1)$. We prove this result in Section C.14.2.

THEOREM 12. *Assume $p \geq 2$ and that the leave-one-out model and Lasso estimators satisfy* (A1). *Recall τ^*_{loo}, ζ^*_{loo} are the solution to the fixed-point equations* (9a) *and* (9b) *in the leave-one-out model.*

(a) (Coverage and power of the leave-one-out confidence interval) *For any* $\gamma > 0$, *there exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$ *and* γ *such that for all* $\epsilon < c'$, $|\theta_j^* - \widehat{\theta}_{j,\text{init}}| < c'$, *and* $\theta \in \mathbb{R}$, *we have*

$$\begin{aligned}&\big|\mathbb{P}\big(\theta \notin \mathsf{CI}_j^{\text{loo}}\big) - \mathbb{P}\big(|\theta_j^* + \Sigma_{j|-j}^{-1/2}\tau_{\text{loo}}^* G/\sqrt{n} - \theta| > \Sigma_{j|-j}^{-1/2}\tau_{\text{loo}}^* z_{1-\alpha/2}/\sqrt{n}\big)\big| \\ &\quad \le C\bigg(\big|\theta_j^* - \widehat{\theta}_{j,\text{init}}\big|^{2/3} + n^{2/6+\gamma}\,|\theta_j^* - \theta| + \frac{1}{n}\bigg),\end{aligned}$$

where $G \sim \mathsf{N}(0,1)$. (*See the discussion following the theorem for an interpretation of this bound*).

(b) (Length of the leave-one-out confidence interval). *There exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\text{model}}$, M' *and* δ_{loo} *such that for all* $\epsilon < c'$,

$$\mathbb{P}_{\theta_j^*}\bigg(\bigg|\frac{\widehat{\tau}_{\text{loo}}^{\,j}}{\tau_{\text{loo}}^*} - 1\bigg| > \epsilon\bigg) \le \frac{C}{\epsilon^3} e^{-cn\epsilon^6}.$$

Note that $\mathbb{P}(|\theta_j^* + \Sigma_{j|-j}^{-1/2}\tau_{\text{loo}}^* G/\sqrt{n} - \theta| > \tau_{\text{loo}}^* z_{1-\alpha/2}\Sigma_{j|-j}^{-1/2}/\sqrt{n})$ is the power of the standard two-sided confidence interval under Gaussian observations $\Sigma_{j|-j}^{1/2}\theta_j^* + \tau_{\text{loo}}^* G/\sqrt{n}$ against alternative θ. This normal approximation holds provided $\theta_j^* - \widehat{\theta}_{j,\text{init}} = o(1)$ and $\theta - \theta_j^* = o(n^{-2/6-\gamma})$ for some $\gamma > 0$. In particular, it holds for $\theta - \theta_j^*$ on the $n^{-1/2}$ scale.

It is convenient to consider a few special cases of Theorem 12:

1. $\widehat{\theta}_{j,\text{init}} = 0$ and $\theta_j^* = 0$. In this case, setting $\theta = 0$ yields $|\mathbb{P}(0 \notin \mathsf{CI}_j^{\text{loo}}) - \alpha| \le C/n$. In fact, a moment of reflection shows that this bound can be improved to yield

$$\mathbb{P}\big(0 \notin \mathsf{CI}_j^{\text{loo}}\big) = \alpha.$$

That is, we have exact control of type I errors.

2. $\widehat{\theta}_{j,\text{init}} = 0$ and $\theta_j^* = o(1)$. Setting again $\theta = \theta_j^*$, we obtain

$$\big|\mathbb{P}\big(\theta_j^* \notin \mathsf{CI}_j^{\text{loo}}\big) - \alpha\big| = o(1).$$

That is, we obtain asymptotic coverage for all nonzero coefficients that are small (note that if $\|\boldsymbol{\theta}^*\|_2 = O(1)$, this is the case for most nonzero coefficients).

3. Generally leave-one-out confidence intervals are successful provided $\widehat{\theta}_{j,\text{init}}$ is consistent for θ_j^*. Note that we assume $\widehat{\theta}_{j,\text{init}}$ is deterministic, which accommodates settings in which it is based on prior knowledge or is an estimate based on an independent data set. Note that consistency is a rather weak requirement (indeed $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O(1)$). We also point at the next section for a construction of exact confidence intervals that do not require the initialization $\widehat{\theta}_{j,\text{init}}$.

REMARK 4.7. Even when θ_j^* is 0, it is possible that $\widehat{\theta}_j$ as estimated by the Lasso is of order 1; indeed, Figure 2 presents a simulation of such a scenario. In this case, the naïve standard error for the debiased Lasso is too small, but our leave-one-out construction with $\widehat{\theta}_{j,\text{init}} = 0$ achieves coverage. Moreover, in Section A.2, we provide simulation evidence that in this scenario, the leave-one-out estimates ξ_j have smaller variance than the debiased estimates $\widehat{\theta}_j^{\text{d}}$, indicating that they permit more precise inference. Characterizing in which scenarios the leave-one-out intervals are more or less precise than the debaised confidence intervals is a promising avenue for future work.

In concurrent work, Bellec and Zhang [10] consider debiasing with a arbitrary convex penalties, and establish success of the debiased confidence intervals when (among other assumptions) the initial estimate $\widehat{\theta}_j$ is consistent in coordinate j. Our result is comparable with theirs (for a special choice of the penalty) but has the advantage of holding down to the Donoho–Tanner phase transition and permitting that taking λ be arbitrarily close to 0. We also do not require that $\|\widehat{\boldsymbol{\theta}}\|_0/n \le 1/2$ with high probability.

The leave-one-out construction is a renormalized empirical correlation between the residuals of the regression of $\boldsymbol{y}_{\text{init}}$ on $\boldsymbol{X}_{-j}$ and of $\boldsymbol{x}_j$ on $\boldsymbol{X}_{-j}$. It is thus similar to a method proposed by [50, 52], in which the partial correlation between two features in a Gaussian graphical model is estimated by regressing each of these features on the remaining features. For each regression, [50, 52] use the scaled Lasso and must assume $s = o(\sqrt{n})$ (up to logarithmic terms) to achieve normal inference. In contrast, we assume that one of the regressions—that of $\boldsymbol{x}_j$ on $\boldsymbol{X}_{-j}$—is known perfectly, whereas the second regression—that of $\boldsymbol{y}_{\text{init}}$ on $\boldsymbol{X}_{-j}$—must be estimated and can have much less structure (possibly linear sparsity). For this reason, we require a degrees-of-freedom correction, which is not present in [50, 52].

Relation to the conditional randomization test. It is worth remarking that exact tests and confidence intervals for θ_j^* may be constructed in our setting. In fact, when the feature distribution is known, one can perform an exact test of

$$\boldsymbol{y} \perp\!\!\!\perp \check{\boldsymbol{x}}_j | \boldsymbol{X}_{-j}, \tag{24}$$

even without Gaussianity or any assumption on the conditional distribution of the outcome $\boldsymbol{y}$ given the features $\boldsymbol{X}$ (see, e.g., [15, 39, 44]). The test which achieves this is called the *conditional randomization test* and is feasible to use for any arbitrary variable importance statistic $T(\boldsymbol{y}, \boldsymbol{X})$. The key observation leading to the construction of the conditional randomization test is that under the null, the distribution of $T(\boldsymbol{y}, \boldsymbol{X})|\boldsymbol{X}_{-j}$ is equal to the distribution of $T(\boldsymbol{y}, \boldsymbol{x}_j', \boldsymbol{X}_{-j})|\boldsymbol{X}_{-j}$ where $\boldsymbol{x}_j'$ is drawn by the statistician from the distribution $\boldsymbol{x}_j|\boldsymbol{X}_{-j}$ without using $\boldsymbol{y}$. Under the null, this distribution can be computed to arbitrary precision by Monte Carlo sampling. We refer the reader to [15, 39, 44] for more details about how these observations lead to the construction of an exact test.

When the linear model is well specified, the null (24) corresponds to $\theta_j^* = 0$, and our leave-one-out procedure with $\widehat{\theta}_{j,\text{init}} = 0$ implements the conditional randomization test under this null, as we now explain. The statistic ξ_j, defined in equation (23) and used in the construction of the leave-one-out interval, can also be used as the variable importance statistic in the conditional randomization test. The Gaussian design assumption and the choice of statistic ξ_j permit an explicit description of the null conditional distribution $\xi_j|\boldsymbol{y}, \boldsymbol{X}_{-j}$. Indeed, because $\check{\boldsymbol{x}}_j^{\perp}$ is independent of $(\boldsymbol{y}, \boldsymbol{X}_{-j}, \widehat{\boldsymbol{\theta}}_{\text{loo}})$ under the null $\theta_j^* = 0$, one has

$$\sqrt{n}\xi_j|\boldsymbol{y}, \boldsymbol{X}_{-j} \sim \mathsf{N}\big(0, \Sigma_{j|-j}^{-1}(\widehat{\tau}_{\text{loo}}^j)^2\big).$$

In our setting, we can access the null conditional distribution through its analytic form rather than through Monte Carlo sampling. The test which rejects when $0 \notin \mathsf{CI}_j^{\text{loo}}$ is exactly the conditional randomization test for the null (24) based on the variable importance statistic $|\xi_j|$.[4] As a consequence, the leave-one-out confidence intervals have exact finite sample coverage under the null $\theta_j^* = 0$ when $\widehat{\theta}_{j,\text{init}} = 0$. Moreover, Theorem 12 provides more than what existing theory on the conditional randomization test can provide: it gives confidence intervals, which are valid under proportional asymptotics and a power analysis for the corresponding tests.

[4]This holds provided that the statistician computes $\xi_j|\boldsymbol{y}, \boldsymbol{X}_{-j}$ exactly by taking an arbitrarily large Monte Carlo sample.

The linearity assumption in our setting allows us to push this rationale further. When $\theta_j^* = \widehat{\theta}_{j,\text{init}}$, the jth residualized covariate $\check{\boldsymbol{x}}_j^{\perp}$ is independent of the pseudo-outcome $\boldsymbol{y}_{\text{init}}$ and $\boldsymbol{X}_{-j}$. Thus, by the same logic as above, the leave-one-out confidence interval achieve exact coverage when $\widehat{\theta}_{j,\text{init}} = \theta_j^*$. In particular, we have an exactly valid test of $\theta_j^* = \widehat{\theta}_{j,\text{init}}$ for all values of $\widehat{\theta}_{j,\text{init}}$. The inversion of this collection of tests, indexed by $\widehat{\theta}_{j,\text{init}}$, produces an exact confidence interval. Details of this construction are provided in Appendix C.14.

We prefer the approximate interval $\mathsf{CI}_j^{\text{loo}}$ to the exact interval outlined in the preceding paragraph for computational reasons. The construction of these exact confidence intervals requires recomputing the leave-one-out Lasso estimate using pseudo-outcome $\boldsymbol{y} - \check{\boldsymbol{x}}_j^{\perp}\widehat{\theta}_{j,\text{init}}$ for each value of $\widehat{\theta}_{j,\text{init}}$. In contrast, the leave-one-out confidence interval we provide requires only computing a single leave-one-out Lasso estimate. It achieves only approximate coverage, but our simulations in Section A.2 show that coverage is good already for n, p, s on the order of 10 s or 100 s. An additional benefit of Theorem 12 is its quantification of the length of the leave-one-out confidence intervals and the power of the corresponding tests, which are not in general accessible for the conditional randomization test or confidence intervals based on it. In fact, because the test $0 \notin \mathsf{CI}_j^{\text{loo}}$ is exactly the conditional randomization test, Theorem 12(a) applied under θ_j^* provides an estimate of the power of the conditional randomization test under alternative $\theta_j^* = \omega$.

4.7. *Restricted eigenvalues and the Donoho–Tanner phase transition.* An important feature of our results is that they hold down to the Donoho–Tanner phase transition, which can be weaker than the requirement based on restricted eigenvalue conditions.

Specifically, the standard restricted eigenvalue on support $S \subset [p]$ of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is defined as (see, e.g., [6, 12])

$$\text{RE}(S, c) := \text{RE}(S, c; \boldsymbol{X}) := \min_{\boldsymbol{\theta}\in\mathcal{C}_{\text{RE}}(s,c)} \frac{1}{\sqrt{n}}\|\boldsymbol{X}\boldsymbol{\theta}\|_2 > 0,$$

where $\mathcal{C}_{\text{RE}}(S, c) := \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}_{S^c}\|_1 \le c\|\boldsymbol{\theta}_S\|_1, \|\boldsymbol{\theta}\|_2 = 1\}$. In order for bounds based on restricted eigenvalues to yield the correct estimation error rate, one typically needs $\text{RE}(S, c)$ to be bounded away from zero for some c strictly larger than 1.

In the random design setting of the present paper, we illustrate by the following example that, $\text{RE}(S, c) = 0$ with high probability for some nonvanishing interval of sampling rates above the Donoho–Tanner phase transition.

PROPOSITION 13. *Consider a block diagonal matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ whose first $s/2$ diagonal blocks are $\boldsymbol{K} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ for some constant $\rho > 0$, and whose lower right $(p - s) \times (p - s)$ diagonal block is $\mathbf{I}_{p-s}$. Let $S = \{1, 2, \ldots, s\}$ and $\boldsymbol{x}^* = \mathbf{1}_S \in \mathbb{R}^p$ be the indicator vector on S.*

Consider the limit $s, p, n \to \infty$ with $s/p = \varepsilon$ and $n/p = \delta$ fixed. In this setting, the Gaussian width $\mathcal{G}(\boldsymbol{x}^*, \boldsymbol{\Sigma}) = \overline{\mathcal{G}}(\varepsilon, \delta, \rho) \in (0, \infty)$ only depends on n, p, s through the ratios ε, δ. Further, there exists $\Delta(\varepsilon, \delta, \rho) > 0$ such that if $\mathcal{G}(\boldsymbol{x}^*, \boldsymbol{\Sigma})^2 < \delta < \mathcal{G}(\boldsymbol{x}^*, \boldsymbol{\Sigma})^2 + \Delta(\varepsilon, \delta, \rho)$, then with probability going to 1 as $p \to \infty$, $\text{RE}(S, c; \boldsymbol{X}) = 0$ for all $c \ge 1$.

We prove Proposition 13 in Appendix D.4. We remark that the set $\mathcal{C}_{\text{RE}}(S, 1)$ is closely related to the cone $\mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma})$ used in defining the Gaussian width $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma})$: the former is based on the cone constraint $\|\boldsymbol{\theta}_{S^c}\|_1 \le \|\boldsymbol{\theta}_S\|_1$, whereas the latter is based on the cone constraint $\|\boldsymbol{\theta}_{S^c}\|_1 \le \langle \text{sign}(\boldsymbol{x}), \boldsymbol{\theta}\rangle$, where $S = \text{supp}(\boldsymbol{x})$. The right-hand side $\|\boldsymbol{\theta}_S\|_1$ is the supremum of $\langle \text{sign}(\boldsymbol{x}), \boldsymbol{\theta}\rangle$ over all sign vectors $\boldsymbol{x}$ with support S. Existing proofs based on the restricted eigenvalue condition [6, 12] go through if $\|\boldsymbol{\theta}_S\|_1$ were replaced by $\langle \text{sign}(\boldsymbol{x}), \boldsymbol{\theta}\rangle$ in the definition of the restricted eigenvalue condition (indeed, in these proofs, this quantity serves only as a bound on $\|\boldsymbol{\theta}_S^*\|_1 - \|\boldsymbol{\theta}_S\|_1$). Thus, Proposition 13 as demonstrates the importance of

using $\langle \operatorname{sign}(\boldsymbol{x}), \boldsymbol{\theta} \rangle$ instead of $\|\boldsymbol{\theta}_S\|_1$ in definitions of the Gaussian width or restricted eigenvalue rather than demonstrating a fundamental limitation of prior analyses. A fundamental improvement of our analysis relative to prior analyses is that we can take $c = 1$ rather than $c > 1$. For fixed $c > 1$, even a modified restricted eigenvalue condition using $\langle \operatorname{sign}(\boldsymbol{x}), \boldsymbol{\theta} \rangle$ results in a gap with respect to our condition $\mathcal{G}(\boldsymbol{x}^*, \boldsymbol{\Sigma})^2 < \delta$.

A natural question is whether our results hold for sampling rates below the Donoho–Tanner phase transition. The following proposition gives a partial answer, in the negative direction.

PROPOSITION 14. *Consider $\boldsymbol{x} \in \{-1, 0, 1\}^p$ with $\|\boldsymbol{x}\|_0 \geq 1$ and $\epsilon > 0$. If*

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \geq \sqrt{\frac{n}{p}} + \epsilon,$$

then, for any $r > 0$, there exists $\boldsymbol{\theta}^*$ (depending on $r, \lambda, \sigma, \kappa_{\min}, \kappa_{\max}, n, p$ and $\|\boldsymbol{x}\|_0$) with $\operatorname{sign}(\boldsymbol{\theta}^*) = \boldsymbol{x}$ such that if the data is generated according to (2), then

$$\mathbb{P}\big(\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big\|_2 \geq r\big) \geq 1 - Ce^{-cp\epsilon^2},$$

where $C, c > 0$ depend only on $\kappa_{\max}$.

In particular, the Lasso has unbounded risk on sparse balls below the Donoho–Tanner phase transition, whence Theorem 5 cannot hold with bounded fixed-point parameters. We prove Proposition 14 in Appendix D.1.

Funding. The first author was partially supported by NSF Grants CCF-1714305, IIS-1741162 and ONR Grant N00014-18-1-2729. We thank the anonymous reviewers for their valuable reviews.

The second author was partially supported by the National Science Foundation Graduate Research Fellowship Grant DGE-1656518.

The third author was partially supported by NSF Grants DMS-2015447/2147546, CAREER award DMS-2143215 and the Google Research Scholar Award.

SUPPLEMENTARY MATERIAL

Supplement to "The Lasso with general Gaussian designs with applications to hypothesis testing" [18] (DOI: 10.1214/23-AOS2327SUPP; .pdf). The supplement contains proofs and technical details that were omitted from the main text.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

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

[3] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[4] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[5] BELLEC, P. C. (2023). Out-of-sample error estimation for M-estimators with convex penalty. *Inf. Inference* **12** 2782–2817. MR4660702 https://doi.org/10.1093/imaiai/iaad031

[6] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663 https://doi.org/10.1214/17-AOS1670

[7] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. Loh and M. Raginsky, eds.) *Proceedings of Machine Learning Research* **178** 1912–1947. PMLR.
[8] BELLEC, P. C. and ZHANG, C.-H. (2018). Second order stein: Sure for sure and other applications in high-dimensional inference.
[9] BELLEC, P. C. and ZHANG, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli* **28** 713–743. MR4389062 https://doi.org/10.3150/21-BEJ1348
[10] BELLEC, P. C. and ZHANG, C.-H. (2023). Debiasing convex regularized estimators and interval estimation in linear models. *Ann. Statist.* **51** 391–436. MR4600987 https://doi.org/10.1214/22-aos2243
[11] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600
[12] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620
[13] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9
[14] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885 https://doi.org/10.1214/09-AOS752
[15] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265
[16] CELENTANO, M. (2021). Approximate separability of symmetrically penalized least squares in high dimensions: Characterization and consequences. *Inf. Inference* **10** 1105–1165. MR4312091 https://doi.org/10.1093/imaiai/iaaa037
[17] CELENTANO, M. and MONTANARI, A. (2021). Cad: Debiasing the lasso with inaccurate covariate model.
[18] CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). Supplement to "The Lasso with general Gaussian designs with applications to hypothesis testing." https://doi.org/10.1214/23-AOS2327SUPP
[19] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7
[20] CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116
[21] CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2016). On cross-validated lasso. Available at arXiv:1605.02214.
[22] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[23] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388 https://doi.org/10.1098/rsta.2009.0152
[24] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823
[25] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716 https://doi.org/10.1073/pnas.0502258102
[26] DONOHO, D. L. and TANNER, J. (2009). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.* **22** 1–53. MR2449053 https://doi.org/10.1090/S0894-0347-08-00600-0
[27] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067
[28] EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560. MR1467848 https://doi.org/10.2307/2965703
[29] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559
[30] EL KAROUI, N. and PURDOM, E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *J. Mach. Learn. Res.* **19** Paper No. 5. MR3862412
[31] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond., Ser. A, Contain. Pap. Math. Phys. Character* **222** 309–368.

[32] GEER, S. A. and VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation* **6**. Cambridge University Press, Cambridge.
[33] HAN, Q. and SHEN, Y. (2023). Universality of regularized regression estimators in high dimensions. *Ann. Statist.* **51** 1799–1823. MR4658577 https://doi.org/10.1214/23-aos2309
[34] HASTIE, T. J. (2017). *Generalized Additive Models*. Routledge, London.
[35] HU, H. and LU, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory* **69** 1932–1964. MR4564688
[36] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[37] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629
[38] JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630
[39] KATSEVICH, E. and RAMDAS, A. (2022). On the power of conditional independence testing under model-X. *Electron. J. Stat.* **16** 6348–6394. MR4517344 https://doi.org/10.1214/22-ejs2085
[40] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7
[41] LI, G., FAN, W. and WEI, Y. (2023). Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302930120. MR4637851
[42] LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. ArXiv preprint. Available at arXiv:2208.03313.
[43] LI, Y. and WEI, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. ArXiv preprint. Available at arXiv:2110.09502.
[44] LIU, M., KATSEVICH, E., JANSON, L. and RAMDAS, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109** 277–293. MR4430958 https://doi.org/10.1093/biomet/asab039
[45] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038
[46] MONTANARI, A. and NGUYEN, P.-M. (2017). Universality of the elastic net error. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 2338–2342. IEEE Press, New York.
[47] MONTANARI, A. and SAEED, B. N. (2022). Universality of empirical risk minimization. In *Conference on Learning Theory* 4310–4312. PMLR.
[48] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400
[49] OYMAK, S. and TROPP, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Inf. Inference* **7** 337–446. MR3858331 https://doi.org/10.1093/imaiai/iax011
[50] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286
[51] SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. MR3718164 https://doi.org/10.1214/16-AOS1521
[52] SUN, T. and ZHANG, C.-H. (2012). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm" [MR3027084]. *Statist. Sinica* **22** 1354–1358. MR3027086
[53] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116
[54] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720
[55] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709.
[56] TROPP, J. A. (2015). Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*. *Appl. Numer. Harmon. Anal.* 67–101. Birkhäuser/Springer, Cham. MR3467419
[57] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[58] WANG, H., YANG, Y., BU, Z. and SU, W. (2020). The complete lasso tradeoff diagram. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 20051–20060. Curran Associates, Red Hook.

[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. MR3153940 https://doi.org/10.1111/rssb.12026

[60] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127