Submitted to the Annals of Statistics

THE LASSO WITH GENERAL GAUSSIAN DESIGNS WITH APPLICATIONS TO HYPOTHESIS TESTING

BY MICHAEL CELENTANO[1,†,*], ANDREA MONTANARI[1,2,*] AND YUTING WEI[3,‡]

[1]*Department of Statistics, Stanford University*

[2]*Department of Electrical Engineering, Stanford University*

[3]*Department of Statistics, Carnegie Mellon University*

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 non-smooth; (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 standard Gaussian designs, and subsequently generalized to other high-dimensional estimation procedures. Here we extend the same characterization to Gaussian correlated designs with non-singular covariance structure. This characterization is expressed in terms of a simpler "fixed-design" model. We establish non-asymptotic 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 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. Statistical inference questions 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 [22, 31].

It is by now well understood that asymptotic normality breaks down in high dimension, even when considering low-dimensional projections of the underlying covariates [6, 27, 54]. As a consequence, the statistician has a limited toolbox to address inferential questions. This challenge is compounded by the fact that resampling methods also fail in this context [21].

The Lasso is arguably the prototypical method in high-dimensional statistics [50]. 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

*Partially supported by the NSF grants CCF-1714305, IIS-1741162, and by the ONR grant N00014-18-1-2729.

†Partially supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1656518.

‡Partially supported by the NSF grants CCF-2007911 and DMS-2015447.

MSC2020 subject classifications: Primary 62J07, 62E17; secondary 62F05, 62F12.

Keywords and phrases: Lasso, debiased Lasso, exact asymptotics, Convex Gaussian Min-Max Theorem, Gaussian designs, Gaussian width.

by solving the optimization problem

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

Here $\boldsymbol{y}\in\mathbb{R}^n$ is the vector with i-th entry equal to y_i, and $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ is the matrix with i-th 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 $\boldsymbol{z}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_n)$ is a Gaussian noise vector. We denote j^{th} column of $\boldsymbol{X}$ by $\breve{\boldsymbol{x}}_j\in\mathbb{R}^n$.

Classical normal theory does not apply to the estimator $\widehat{\boldsymbol{\theta}}$ for two reasons that are instructive to revisit. First, the Lasso objective (1) is non-smooth: its gradient is discontinuous on the hyperplanes $\theta_i = 0$. As a consequence, $\widehat{\theta}_i = 0$ with positive probability (indeed, as we will see, with probability bounded away from 0 for large n, p). Second, the estimation error $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2$ is not negligible in practical settings. As a consequence we cannot rely on perturbative arguments that focus on a small neighborhood of $\boldsymbol{\theta}^*$.

A substantial body of theoretical work studied the Lasso with fixed (non-random) designs $\boldsymbol{X}$ in the regime $s\log p/n = O(1)$ [11, 14, 38, 7]. These approaches require that λ be chosen so that λ is larger than $\sigma\|\boldsymbol{X}^\top\boldsymbol{z}\|_\infty$ or, more recently, the s^{th}-largest element of $\{\sigma|\breve{\boldsymbol{x}}_j^\top\boldsymbol{z}|\}_{j\le n}$ with high probability; they rely on restricted eigenvalue or similar compatibility conditions on the design matrix $\boldsymbol{X}$; and they control the Lasso estimation error up to constants. Unfortunately, these results provide limited insight on the distribution of the estimator $\widehat{\boldsymbol{\theta}}$.

A more recent line of research attempts to address these limitations by characterizing the distribution of $\widehat{\boldsymbol{\theta}}$ for design matrices $\boldsymbol{X}$ with i.i.d. Gaussian entries [6, 49, 35]. These analyses assume n, p and the number of non-zero coefficients $\|\boldsymbol{\theta}^*\|_0$ to be large and of the same order, and apply to any λ of the order of the typical size of $\sigma|\breve{\boldsymbol{x}}_j^\top\boldsymbol{z}|$. This covers the typical values of the regularization selected by standard procedures such as cross-validation [17, 35]. Under these assumptions, [6] first proved 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 $\boldsymbol{g}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_p)$, 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 $\eta(x;\zeta) := (|x|-\lambda/\zeta)_+\text{sign}(x)$ 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).

Both numerical simulations and universality arguments suggest that the results of [6, 49, 35] apply to independent but possibly non-Gaussian covariates (see [5, 39, 36] for rigorous universality results). However, these predictions are expected not to be asymptotically exact when covariates are correlated.

The present paper substantially generalizes this line of work by extending it to the case of correlated Gaussian designs with well-conditioned covariance. Namely we assume the covariates $(\boldsymbol{x}_i)_{i\le n}$ to be i.i.d. with $\boldsymbol{x}_i\sim\mathsf{N}(0,\boldsymbol{\Sigma}/n)$. Our results hold uniformly over covariances with eigenvalues in $[\kappa_{\min},\kappa_{\max}]$ for some $0<\kappa_{\min}<\kappa_{\max}<\infty$; regularization parameters $\lambda\in[\lambda_{\min},\lambda_{\max}]$ for some $0<\lambda_{\min}<\lambda_{\max}<\infty$; and signals $\boldsymbol{\theta}^*$ satisfying a suitable sparsity condition. The sparsity condition involves a modified Gaussian width of a certain convex cone in $\mathbb{R}^p$. We expect this condition to be often tight (in particular, it is for $\boldsymbol{\Sigma}=\mathbf{I}$). Assumptions on Gaussian widths have been used in the past to characterize noiseless and stable sparse recovery in the compressed sensing literature [16, 51]. Here we show that they also imply uniform approximation of the distribution of $\widehat{\boldsymbol{\theta}}$.

We next provide a succinct overview of our results.

Lasso estimator. We characterize the distribution of the Lasso estimator $\widehat{\boldsymbol{\theta}}$. As in the case $\boldsymbol{\Sigma} = \mathbf{I}_p$, this characterization involves observations $\boldsymbol{y}^f$ from a related statistical model:

$$\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau \boldsymbol{g}\,, \tag{3}$$

where $\boldsymbol{g} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$ and $\tau \geq 0$. We call this 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

$$\eta(\boldsymbol{y}^f, \zeta) := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}. \tag{4}$$

We show that, for any Lipschitz function $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, the value of $\phi(\widehat{\boldsymbol{\theta}}/\sqrt{p}, \boldsymbol{\theta}^*/\sqrt{p})$ concentrates uniformly over λ, and with probability close to 1 uniformly over a suitable sparsity class. The value of it concentrates on is the expected value of the corresponding quantity under the fixed design model; that is, $\mathbb{E}[\phi(\widehat{\boldsymbol{\theta}}^f/\sqrt{p}, \boldsymbol{\theta}^*/\sqrt{p})]$. The effective noise and threshold parameters τ^*, ζ^* are given as the unique solution of a pair of nonlinear equations introduced below.

In the case of uncorrelated covariates, the fixed design problem is particularly simple because the optimization problem (4) is separable, and $\eta(\boldsymbol{y}^f, \zeta)$ reduces to soft thresholding applied component-wise. For specific correlation structures $\boldsymbol{\Sigma}$, problem (4) can also be simplified, but we defer this to future work. More generally, it is simpler than the original problem since the objective in Eq. (4) is strongly convex, and hence more directly amenable to deriving explicit bounds.

Residuals and sparsity. In low-dimensional theory, the residuals vector $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}$ is roughly $\mathsf{N}(0, \sigma^2\mathbf{I}_n)$, a remark that provides the basis for classical F tests and for bootstrapping the residuals. We prove that in the high-dimensional setting the residuals are instead approximately $\mathsf{N}(0, (\tau^*\zeta^*)^2\mathbf{I}_n)$, suggesting that these methods should be revised in high-dimension.

We also estimate the sparsity of the lasso estimator, showing that it concentrates so that $\|\boldsymbol{\theta}^*\|_0 \approx n(1-\zeta^*)$. Notice that, together with the previous result, this implies that the parameters τ^*, ζ^* can be entirely estimated from the data. Since τ^* controls the noise in the fixed design model, its estimation is of particular interest. A simple method is to use the following degrees-of-freedom adjusted residuals

$$\widehat{\tau}(\lambda)^2 := \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)^2}\,. \tag{5}$$

It was already observed in [35] 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.

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:

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

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 $\boldsymbol{\Sigma}^{-1}$ and proved approximate normality when $\|\boldsymbol{\theta}^*\|_0 =: s_0 = o(\sqrt{n}/\log p)$ [54, 52, 27, 26, 28]. Later work considered the case of Gaussian covariates with known covariance, and set $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$. In this idealized setting, the sparsity condition was relaxed to

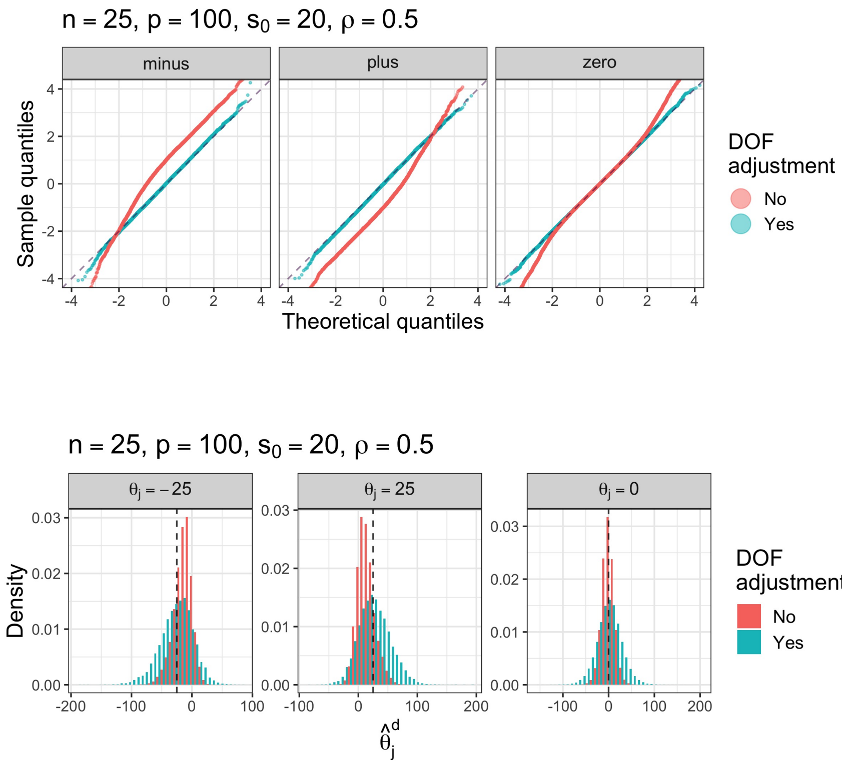


Fig 1. The debiased Lasso with and without degrees-of-freedom (DOF) adjustment. Here $p = 100$, $n = 25$, $s_0 = 20$, $\Sigma_{ij} = \rho^{|i-j|} = 0.5^{|i-j|}$, $\lambda = 4$, $\sigma = 1$. The coefficients vector $\boldsymbol{\theta}^*$ contains 10 entries $\theta_i^* = +25$, and 10 entries $\theta_i^* = -25$. Quantiles and densities are compared with the ones of the standard normal distribution.

$s_0 = o(n/(\log p)^2)$ for inference on a single coordinate [28] and $s_0 = o(n^{2/3}/\log(p/s_0)^{1/3})$ for a general linear functional of $\boldsymbol{\theta}^*$ [9].

The latter conditions turn out to be tight for the standard choice $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$. As shown in [27, 35] for uncorrelated designs and in [9, 10] for correlated designs with $n > p$, it is necessary to adjust the previous construction for the degrees of freedom by setting $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}/(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)$:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \frac{1}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) . \tag{6}$$

Here we establish approximate normality and unbiasedness of this construction for arbitrary aspect ratios n/p and arbitrary covariances. As a consequence, we construct confidence intervals with coverage guarantees on average across coordinates in the proportional regime. 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. In contrast, debiasing with degrees of freedom adjustment is successful already for problem dimensions on the order of 10s or 100s. (See Section 4.1 for details. Similar simulations at different model parameters is shown in Section D)

Inference on a single coordinate. Our results on $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ are not sharp enough to show that a fixed single coordinate $\widehat{\theta}^{\mathrm{d}}_j$ is asymptotically Gaussian, and hence we do not establish a per coordinate coverage guarantee of confidence intervals based on $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$. While it might be possible to leverage our results to get per-coordinate control following the strategy of [47], we adopt a simpler approach here. We use a leave-one-out method to construct confidence intervals for which we can prove asymptotic validity via a direct argument. Further, we prove that the length of these intervals is close to optimal.

We observe empirically that the confidence intervals produced by the leave-one-out method are very similar to the ones obtained using the debiased Lasso. We leave a rigorous study of this phenomenon to future work. 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).

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 $\mathrm{supp}(\boldsymbol{x}) := \{i \mid \boldsymbol{x}_i \neq 0\}$. 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 $rn^{1/q}$ and center $\boldsymbol{v}$, namely,

$$\mathsf{B}_q(\boldsymbol{v}; r) := \left\{ \boldsymbol{x} \in \mathbb{R}^p \;\middle|\; \frac{1}{p}\|\boldsymbol{x} - \boldsymbol{v}\|_q^q \leq r^q \right\} \text{ for } q > 0, \text{ and } \mathsf{B}_0(\nu) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \;\middle|\; \frac{\|\boldsymbol{\theta}\|_0}{p} \leq \nu \right\}.$$

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 $\mathbb{S}^n_{\geq 0}$ is used to denote the set of $n \times n$ positive semidefinite matrices.

2. Preliminaries. As mentioned above, our main result establishes an asymptotic equivalence between the undersampled linear model of Eq. (2) and the linear model with fixed design[1] $\boldsymbol{\Sigma}^{1/2}$ of Eq. (3). We define the prediction vector in the fixed-design model by $\widehat{\boldsymbol{y}}(\boldsymbol{y}^f, \zeta) := \boldsymbol{\Sigma}^{1/2}\eta(\boldsymbol{y}^f, \zeta)$.

Setting the stage, let the in-sample prediction risk and degrees-of-freedom of the Lasso estimator in the fixed-design model be

$$\begin{aligned}
\mathsf{R}(\tau^2, \zeta) &:= \frac{1}{n}\mathbb{E}\left[\|\widehat{\boldsymbol{y}}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}, \zeta) - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*\|_2^2\right], \\
\mathsf{df}(\tau^2, \zeta) &:= \frac{1}{n\tau}\mathbb{E}\left[\langle \widehat{\boldsymbol{y}}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}, \zeta), \boldsymbol{g}\rangle\right] \\
&= \frac{1}{n}\mathbb{E}\left[\|\eta(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}, \zeta)\|_0\right],
\end{aligned}$$

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 terminology "degrees-of-freedom" for the quantity df originated with [55], and its equivalence to the expected sparsity of the Lasso estimate holds, for example, by [55, Theorem 1].

Fixed point equations. Let τ^*, ζ^* be solutions to the system of equations

$$\tau^2 = \sigma^2 + \mathsf{R}(\tau^2, \zeta), \tag{8a}$$

$$\zeta = 1 - \mathsf{df}(\tau^2, \zeta). \tag{8b}$$

We refer to these equations as the *fixed point equations*. As it turns out, these solutions play an essential role in characterizing the distribution of the Lasso estimator. We start by showing

[1] We may take any square-root of the matrix $\boldsymbol{\Sigma}$. For simplicity, we will always assume we take a symmetric square-root.

the solutions τ^*, ζ^* are well-defined, as stated formally in the next theorem, whose proof postponed to Appendix A.3.

THEOREM 1. *If $\mathbf{\Sigma}$ is invertible and $\sigma^2 > 0$, then Eqs.* (8a) *and* (8b) *have a unique solution.*

Let us denote the Lasso estimator in the fixed-design model with noise variance ${\tau^*}^2$ and regularization ζ^* by

$$\widehat{\boldsymbol{\theta}}^f := \eta(\mathbf{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g},\ \zeta^*)\,, \tag{9}$$

where $\boldsymbol{g} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$. Our main results establish that the estimator $\widehat{\boldsymbol{\theta}}$ performs approximately like $\widehat{\boldsymbol{\theta}}^f$ and can therefore be understood via the behavior of $\widehat{\boldsymbol{\theta}}^f$. The quality of this approximation and bounds on the behavior of $\widehat{\boldsymbol{\theta}}^f$ depend, in part, on the complexity of the unknown parameter $\boldsymbol{\theta}^*$. The relevant measure of complexity, which we call $(s, \mathcal{G}^*, M)$*-approximate sparsity*, involves an interplay between a sparse approximation of $\boldsymbol{\theta}^*$, the ℓ_1-penalty, and the population covariance $\mathbf{\Sigma}$, which is made precise in the following.

Approximate sparsity. A vector $\boldsymbol{\theta}^*$ is referred to as $(\boldsymbol{x}, M)$*-approximately sparse* for $\boldsymbol{x} \in \{-1, 0, 1\}^p$ and $M > 0$ if there exists $\bar{\boldsymbol{\theta}}^* \in \mathbb{R}^p$ with $\frac{1}{p}\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1 \le M$ and $\boldsymbol{x} = \mathrm{sign}(\bar{\boldsymbol{\theta}}^*)$ (here the sign is taken in an entry-wise manner, with $\mathrm{sign}(0) = 0$). Thus, $(\boldsymbol{x}, M)$-approximate sparsity implies that $\boldsymbol{\theta}^*$ is well-approximated in an ℓ_1 sense by an $\|\boldsymbol{x}\|_0$-sparse or even sparser vector.

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 $(\mathbb{R}^p, \mathcal{B}, \gamma_p)$. Naturally, this space is equipped with the scalar product

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

and the corresponding norm $\|\boldsymbol{f}\|_{L^2}$. For $\boldsymbol{x} \in \{+1, 0, -1\}^p$, and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, define $F(\,\cdot\,; \boldsymbol{x}, \mathbf{\Sigma}) : \mathbb{R}^p \to \mathbb{R}$ via

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

We also denote by $\mathcal{K}(\boldsymbol{x}, \mathbf{\Sigma}) := \big\{\boldsymbol{v} \in \mathbb{R}^p : F(\boldsymbol{v}; \boldsymbol{x}, \mathbf{\Sigma}) \le 0\big\}$ the associated closed convex cone. We define the *functional Gaussian width* of $\mathcal{K}(\boldsymbol{x}, \mathbf{\Sigma})$ by

$$\mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma}) := \sup_{\boldsymbol{v} \in L^2} \Big\{ \frac{1}{p}\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2} : \ \ \|\boldsymbol{v}\|_{L^2} \le \sqrt{p}\,,\ \ \mathbb{E}[F(\boldsymbol{v}; \boldsymbol{x}, \mathbf{\Sigma})] \le 0 \Big\}\,, \tag{10}$$

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

DEFINITION 1. We say $\boldsymbol{\theta}^*$ is $(s, \mathcal{G}^*, M)$*-approximately sparse* for $\mathcal{G}^* > 0$ and $s \in \mathbb{Z}_{>0}$ if there exists $\boldsymbol{x} \in \{-1, 0, 1\}^p$ such that $\boldsymbol{\theta}^*$ is $(\boldsymbol{x}, M)$-approximately sparse and $\|\boldsymbol{x}\|_0 = s$, $\mathcal{G}(\boldsymbol{x}, \mathbf{\Sigma}) \le \mathcal{G}^*$.

We remark that the Gaussian width defined in (10) differs from the standard notion of Gaussian width which appears elsewhere in the literature (see, e.g., [23, 16, 51]). The latter can be defined as

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

Equivalently, recalling that $\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma})$ is the cone of vectors $\boldsymbol{v}$ such that $F(\boldsymbol{v};\boldsymbol{x},\boldsymbol{\Sigma}) \le 0$, one can instead write

$$\mathcal{G}_{\mathrm{std}}(\boldsymbol{x},\boldsymbol{\Sigma}) = \frac{1}{p}\mathbb{E}\Big[\max_{\substack{\boldsymbol{v}\in\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma})\\ \|\boldsymbol{v}\|_2^2/p\le 1}} \langle \boldsymbol{v},\boldsymbol{g}\rangle\Big]. \tag{12}$$

The definitions (10) and (11) immediately imply $\mathcal{G}_{\mathrm{std}}(\boldsymbol{x},\boldsymbol{\Sigma}) \le \mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})$.

On the other hand, we expect $\mathcal{G}_{\mathrm{std}}(\boldsymbol{x},\boldsymbol{\Sigma}) \ge (1-o_p(1))\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})$ in many cases of interest. This is indeed the case for $\boldsymbol{\Sigma} = \mathbf{I}_p$. More generally, if there exists a Lipschitz continuous $\boldsymbol{v}(\boldsymbol{g})$ nearly achieving the supremum in Eq. (10), then we expect the distance of $\boldsymbol{v}(\boldsymbol{g})$ from the cone $\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma})$ to concentrate around 0. In this case, projecting $\boldsymbol{v}(\boldsymbol{g})$ onto $\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma})$ will yield a lower bound on $\mathcal{G}_{\mathrm{std}}(\boldsymbol{x},\boldsymbol{\Sigma})$ which is close to $\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})$.

The distribution of the random design $\boldsymbol{X}$, response vector $\boldsymbol{y}$, and Lasso estimate $\widehat{\boldsymbol{\theta}}$ is determined by the tuple $(\boldsymbol{\theta}^*,\boldsymbol{\Sigma},\sigma,\lambda)$. Our results stated below hold uniformly over choices of $(\boldsymbol{\theta}^*,\boldsymbol{\Sigma},\sigma,\lambda)$ that satisfy the following conditions:

A1 There exist $0<\lambda_{\min}\le\lambda_{\max}<\infty$, $0<\kappa_{\min}\le\kappa_{\max}<\infty$, and $0<\sigma_{\min}\le\sigma_{\max}<\infty$ such that
  - (a) The Lasso regularization parameter λ is bounded $\lambda_{\min}\le\lambda\le\lambda_{\max}$.
  - (b) The singular values $\kappa_j(\boldsymbol{\Sigma})$ of the population covariance $\boldsymbol{\Sigma}$ are bounded $\kappa_{\min}\le\kappa_j(\boldsymbol{\Sigma})\le\kappa_{\max}$ for all j. We define $\kappa_{\mathrm{cond}} := \kappa_{\max}/\kappa_{\min}$.
  - (c) The noise variance σ^2 is bounded $\sigma^2_{\min}\le\sigma^2\le\sigma^2_{\max}$.

A2 There exist $0<\tau_{\min}\le\tau_{\max}<\infty$ and $0<\zeta_{\min}\le\zeta_{\max}<\infty$ such that the unique solution τ^*, ζ^* to the fixed point equations (8a) and (8b) are bounded $\tau_{\min}\le\tau^*\le\tau_{\max}$ and $\zeta_{\min}\le\zeta^*\le\zeta_{\max}$.

We denote the collections of constants appearing in assumptions A1 and A2 by

$$\mathcal{P}_{\mathrm{model}} := (\lambda_{\min},\lambda_{\max},\kappa_{\min},\kappa_{\max},\sigma_{\min},\sigma_{\max}), \quad \mathcal{P}_{\mathrm{fixPt}} := (\tau_{\min},\tau_{\max},\zeta_{\min},\zeta_{\max}).$$

In other words, any choice of the constants $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$ determines a uniformity class of parameters $(\boldsymbol{\theta}^*,\boldsymbol{\Sigma},\sigma,\lambda)$ within which the results stated below apply.

Uniqueness and boundedness guarantees. Checking assumption A2 requires solving (8a) and (8b), which can be a difficult task. However, it turns out that assumption A1 is sufficient to imply assumption A2 provided $\boldsymbol{\theta}^*$ is approximately sparse. We formulate this result in the theorem below and prove it in Section A.4.

THEOREM 2. *Under assumption A1 and if $\boldsymbol{\theta}^*$ is $(s,\sqrt{\delta}(1-\Delta_{\min}),M)$-approximately sparse for some $s/p\ge\nu_{\min}>0$ and $1\ge\Delta_{\min}>0$, then there exist $0<\tau_{\min}\le\tau_{\max}<\infty$ and $0<\zeta_{\min}\le\zeta_{\max}<\infty$ depending only on $\mathcal{P}_{\mathrm{model}}$, δ, $\nu_{\min}$, and $\Delta_{\min}$ such that the unique solution τ^*, ζ^* to Eqs. (8a) and (8b) satisfy $\tau_{\min}\le\tau^*\le\tau_{\max}$ and $\zeta_{\min}\le\zeta^*\le\zeta_{\max}$.*

Let us make a few remarks on this result. First, explicit expressions for $\tau_{\min},\tau_{\max},\zeta_{\min},\zeta_{\max}$ can be found in the proof of Theorem 2. Secondly, while Theorem 2 establishes a sufficient condition for assumption A2 to hold, the latter can hold even if $\boldsymbol{\theta}^*$ is not approximately sparse. For instance, this is the case if $\boldsymbol{\theta}^*$ has a fixed empirical distribution (with finite second moment) and $\boldsymbol{\Sigma}=\mathbf{I}_p$.

It is useful to compare the notion of $(s,\mathcal{G}^*,M)$-approximate sparsity introduced above to sparsity with respect to ℓ_q-norms. It follows from the definition that, for $\boldsymbol{\Sigma}=\mathbf{I}_p$, the Gaussian width depends on $\boldsymbol{x}$ only via $\varepsilon := \|\boldsymbol{x}\|_0/p$. We define

$$\omega^*(\varepsilon) := \mathcal{G}(\boldsymbol{x},\mathbf{I}_p) \text{ for any } \boldsymbol{x} \text{ with } \|\boldsymbol{x}\|_0/p=\varepsilon. \tag{13}$$

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

$$\omega^*(\varepsilon)^2 = \varepsilon + 2(1-\varepsilon)\Phi(-\alpha)\,,$$
$$\text{where } \alpha \text{ satisfies} \qquad \varepsilon = \frac{2[\varphi(\alpha) - \alpha\Phi(-\alpha)]}{\alpha + 2[\varphi(\alpha) - \alpha\Phi(-\alpha)]}\,.$$

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. It is easy to check that $\varepsilon \mapsto \omega^*(\varepsilon)$ is increasing and continuous in ε and goes to 0 and 1 as $\varepsilon \to 0$ and 1, respectively.

The following result formally clarifies the connections between the approximate sparsity (Definition 1) and the ℓ_q-norm for any $q > 0$.

PROPOSITION 3. *Suppose the covariance matrix $\mathbf{\Sigma}$ has singular values $0 < \kappa_{\min} \le \kappa_j(\mathbf{\Sigma}) \le \kappa_{\max} < \infty$ for all j, and let $\kappa_{\rm cond} = \kappa_{\max}/\kappa_{\min}$. For $\omega^*(\varepsilon)$ defined in Eq. (13), define*

$$\varepsilon^*(\kappa_{\rm cond}, \delta) := \sup\{\varepsilon \,|\, \omega^*(\varepsilon) \le \sqrt{\delta/\kappa_{\rm cond}}\}.$$

Then the following hold true:

(a) *If $\boldsymbol{\theta}^* \in \mathsf{B}_q(\nu)$ for $q, \nu > 0$, then for any $\delta > 0$,*

$$\boldsymbol{\theta}^* \text{ is } \Big(\lfloor p\varepsilon^*(\kappa_{\rm cond}, \delta/2)\rfloor,\ \sqrt{\delta/2},\ \nu\big(1 - \varepsilon^*(\kappa_{\rm cond}, \delta/2)\big)\Big)\text{-approximately sparse.}$$

(b) *If $\boldsymbol{\theta}^* \in \mathsf{B}_0(\varepsilon^*(\kappa_{\rm cond}, \alpha))$ for some $\alpha < \delta$, then*

$$\boldsymbol{\theta}^* \text{ is } \Big(\lfloor p\varepsilon^*(\kappa_{\rm cond}, \alpha)\rfloor,\ \sqrt{\alpha},\ 0\Big)\text{-approximately sparse.}$$

The proof of this result is given in Appendix C.1.

In particular, Proposition 3(a) implies that when $\boldsymbol{\theta}^* \in \mathsf{B}_q(\nu)$ for $q > 0$, the conditions of Theorem 2 are satisfied with $\nu_{\max} = \varepsilon^*(\kappa_{\rm cond}, \delta/2)$, any $\nu_{\min} < \varepsilon^*(\kappa_{\rm cond}, \delta/2) - 1/p$, $\mathcal{G}^* = \sqrt{\delta/2}$, and $M = \nu$. Proposition 3(b) implies that when $\boldsymbol{\theta}^* \in \mathsf{B}_0(\varepsilon^*(\kappa_{\rm cond}, \alpha))$, the conditions of Theorem 2 are satisfied with $\nu_{\max} = \varepsilon^*(\kappa_{\rm cond}, \delta/2)$, any $\nu_{\min} < \varepsilon^*(\kappa_{\rm cond}, \delta/2) - 1/p$, $\mathcal{G}^* = \sqrt{\alpha}$, and $M = 0$. Then, in both cases assumption A2 is satisfied. Therefore, our results below apply also to $\boldsymbol{\theta}^*$ in ℓ_q-balls for $q > 0$ or to sufficiently sparse $\boldsymbol{\theta}^*$.

3. 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.

3.1. *Control of the Lasso estimate.* Our first result controls the behavior of the Lasso estimate $\widehat{\boldsymbol{\theta}}$ in the random design model uniformly over $(\boldsymbol{\theta}^*, \mathbf{\Sigma}, \sigma, \lambda)$ satisfying assumptions A1 and A2.

THEOREM 4. *If assumptions A1 and A2 hold, then there exist constants $C, c, c', \gamma > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$ δ, such that for any 1-Lipschitz function $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\exists \boldsymbol{\theta} \in \mathbb{R}^p,\ \Big|\phi\Big(\frac{\boldsymbol{\theta}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^f}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big]\Big| > \epsilon \text{ and } \mathcal{R}(\boldsymbol{\theta}) \le \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{R}(\boldsymbol{\theta}) + \gamma\epsilon^2\right) \le \frac{C}{\epsilon^2} e^{-cn\epsilon^4}.$$

We provide the proof of Theorem 4 in Section B.1.

In words, Theorem 4 shows that—with high probability—any Lipschitz function of any approximate minimizer of the regularized risk $\mathcal{R}(\boldsymbol{\theta})$ concentrates around a deterministic value. This deterministic value is the expectation of the same Lipschitz function of the Lasso estimator in the associated fixed design model. The noise level and regularization in the latter are obtained from the fixed point equations (8a), (8b).

As may have been noticed, because the Lasso estimate $\widehat{\boldsymbol{\theta}}$ is the minimizer of the loss function $\mathcal{R}(\boldsymbol{\theta})$, it satisfies $\mathcal{R}(\widehat{\boldsymbol{\theta}}) \le \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{R}(\boldsymbol{\theta}) + \gamma\epsilon^2$ for all $\epsilon \ge 0$. Thus one can conclude the following corollary immediately.

COROLLARY 5. *Under assumptions A1 and A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that for any 1-Lipschitz function $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\Big| \phi\Big(\frac{\widehat{\boldsymbol{\theta}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^f}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big] \Big| > \epsilon \right) \le \frac{C}{\epsilon^2} e^{-cn\epsilon^4} .$$

Corollary 5 establishes the connection between the Lasso estimator in the random-design model $\widehat{\boldsymbol{\theta}}$ and the Lasso estimator in a fixed-design setting $\widehat{\boldsymbol{\theta}}^f$. Considering a uniform measure over the coordinates of these vectors, it also reveals that the joint empirical distribution of the coordinates of the Lasso estimator and the true parameters vector $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*, \widehat{\theta}_i}$ is close to the one in the fixed design model $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*, \widehat{\theta}_i^f}$ with uniformly high probability.

Simultaneous control over λ. As a matter of fact, one can further generalize the above result to achieve simultaneous control over the penalization parameter λ in the interval $[\lambda_{\min}, \lambda_{\max}]$. Simultaneous control over λ is particularly useful for hyper-parameter tuning.

THEOREM 6. *Assume assumption A1 holds and $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse for some $s/p \ge \nu_{\min} > 0$ and $1 \ge \Delta_{\min} > 0$. Then there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\nu_{\min}$, $\Delta_{\min}$, M, and δ such that the following holds: if $n \ge \sqrt{2}/\Delta_{\min}$, then for any 1-Lipschitz function $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\exists \lambda \in [\lambda_{\min}, \lambda_{\max}],\ \Big| \phi\Big(\frac{\widehat{\boldsymbol{\theta}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^f}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big] \Big| > \epsilon \right) \le \frac{C}{\epsilon^4} e^{-cn\epsilon^4} .$$

The proof of this result is presented in Section B.7.

Theorem 6 provides a sharp characterization of the Lasso estimator which holds simultaneously over all λ in a bounded interval $[\lambda_{\min}, \lambda_{\max}]$. 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 $\widehat{\boldsymbol{\theta}}_\lambda^f$ for the Lasso estimator and fixed-design estimator at regularization λ, we have

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

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

Recall that it is standard to choose λ on the order of the typical size of the $s^{\rm th}$ largest realized value of $\sigma|\breve{\boldsymbol{x}}_j^\top \boldsymbol{z}|$ over j, where s is the sparsity of $\boldsymbol{\theta}^*$ (see, e.g., [7]). In our model, this suggests the choice $\lambda_{\rm std} := \sigma|\breve{\boldsymbol{x}}_j^\top \boldsymbol{z}|$ is of order $\sqrt{\sigma^2 \operatorname{trace}(\boldsymbol{\Sigma}) \log(p/s)/n}$. Since p/s, σ, and trace$(\mathbf{\Sigma})/n$ are of order one, $\lambda_{\rm std} = \Theta(1)$ as well. As shown in [35], the choice $\lambda = \lambda_{\rm std}$ is in general suboptimal by a large factor. The above result controls the optimal error when λ varies in an interval $[c_1\lambda_{\rm std}, c_2\lambda_{\rm std}]$ for any universal constants c_1, c_2.

Control of the empirical distribution. Previous work on iid 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: $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$ to a limiting distribution in a certain sense (either weakly or in Wasserstein sense [6, 35], for example). When covariates are iid, the behavior of $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$ captures all non-trivial behavior of the distribution of $\widehat{\boldsymbol{\theta}}$: indeed, the exchangeability of the model implies that conditional on $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$, 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. However, in the case of correlated covariates, $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$ no longer captures all non-trivial information about the distribution of $\widehat{\boldsymbol{\theta}}$. Thus, Theorem 4, Corollary 5, and Theorem 6 involve general test functions which can probe this additional structure.

Nevertheless, the empirical distribution $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$ may be of interest, in part because it is easily interpretable. Thus, we also provide a result detailing the concentration behavior of this important object. The idea is that since the fixed-design model is well-conditioned, we can show that the empirical distribution $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i^f}$ concentrates. We then leverage Theorem 4 to establish the concentration of empirical distribution $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*,\widehat{\theta}_i}$ also in the random-design model. Precisely, our results are established in terms of a particular metrization of the weak-topology on the space of probability measures on $\mathbb{R}^2$, namely

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

Here $\{\phi_k\}$ denotes a countable dense subset of the 1-Lipschitz functions $\mathbb{R}^2 \mapsto \mathbb{R}$. The metric d_{w^*} metrizes weak convergence in the sense that $\mu_i \xrightarrow{\rm d} \mu$ if and only if $d_{w^*}(\mu_i,\mu) \to 0$.

COROLLARY 7. *There exists μ_*—a probability distribution on $\mathbb{R}^2$, and constants $C, C', c > 0$ depending only on $\mathcal{P}_{\rm model}$ and $\mathcal{P}_{\rm fixPt}$ such that*

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

and

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

Corollary 7 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 6, one can also control properties of μ_* such as its second moments in terms of $\mathcal{P}_{\rm model}$ and $\mathcal{P}_{\rm fixPt}$. We prove Corollary 7 in Appendix B.8.

3.2. *Control of the Lasso residual.* Thus far, we have characterized the distribution of the Lasso estimator in the random design model with general covariance structures. In this section, we aim to establish a control for the residual of the Lasso estimator. Informally, the Lasso residual behaves like a random vector which follows from a normal distribution with zero mean and covariance $(\tau^*\zeta^*)^2\mathbf{I}_n$. Similar to the results aforementioned, the quality of the approximation is controlled uniformly over the models and estimators satisfying assumptions A1 and A2.

Let us formally state our result.

THEOREM 8. *Under assumptions A1 and A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that for any 1-Lipschitz function $\phi : \mathbb{R}^p \to \mathbb{R}$, we have for all $\epsilon < c'$*

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

where $\boldsymbol{h} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. 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 8 is provided in Section B.2.

3.3. *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 around $(1-\zeta^*)$ (given in Eq. (8a) and (8b)), which is made precise in the following result.

THEOREM 9. *Under assumptions A1 and A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, 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 B.4. Recall that, by Theorem 2 and Proposition 3, assumption A1 is sufficient for this result to hold when $\boldsymbol{\theta}^*$ falls in ℓ_q-balls for $q > 0$ or for $\boldsymbol{\theta}^*$ sufficiently sparse.

We make a note that recently Bellec and Zhang [8, Section 3.4] establish that $\|\widehat{\boldsymbol{\theta}}\|_0/n$ concentrates around its expectation with deviations of order $O(n^{-1/2})$ using the second-order Stein's formula. We complement these results by showing that $\|\widehat{\boldsymbol{\theta}}\|_0/n$ has large-deviation probabilities which decay exponentially. Moreover, our result also implies that the value on which $\|\widehat{\boldsymbol{\theta}}\|_0/n$ concentrates is uniformly bounded away from 1 for given $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ.

Control of the subgradient. The proof of Theorem 9 is built upon controlling the vector

$$\widehat{\boldsymbol{t}} = \frac{1}{\lambda}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}), \tag{14}$$

which is a subgradient of the ℓ_1-norm at $\widehat{\boldsymbol{\theta}}$. Since controlling this subgradient may be of independent interest, we state our result formally below. Similarly, we prove that $\widehat{\boldsymbol{t}}$ behaves approximately like the corresponding subgradient in the fixed-design model

$$\widehat{\boldsymbol{t}}^f := \frac{\zeta^*}{\lambda}\,\boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f), \tag{15}$$

where $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$, $\widehat{\boldsymbol{\theta}}^f = \eta(\boldsymbol{y}^f, \zeta^*)$, and $\boldsymbol{g} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$. The quality of the approximation is controlled uniformly over models and estimators satisfying assumptions A1 and A2.

For any measurable set $D \subset \mathbb{R}^p$, define its ϵ-enlargement $D_\epsilon := \{\boldsymbol{x} \in \mathbb{R}^p \mid \inf_{\boldsymbol{x}' \in D} \|\boldsymbol{x} - \boldsymbol{x}'\|_2/\sqrt{p} \geq \epsilon\}$. The following result makes the connection between $\widehat{\boldsymbol{t}}$ and $\widehat{\boldsymbol{t}}^f$ precise.

LEMMA 10. *Under assumptions A1 and A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for any measurable set $D \subset \mathbb{R}^p$ and for all $\epsilon < c'$*

$$\mathbb{P}\left(\widehat{\boldsymbol{t}} \in D_\epsilon\right) \leq 2\mathbb{P}\left(\widehat{\boldsymbol{t}}^f \notin D\right) + \frac{C}{\epsilon^2} e^{-cn\epsilon^4}. \tag{16}$$

Consequently, there exist (possibly new) constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for any 1-Lipschitz function $\phi : \mathbb{R}^p \to \mathbb{R}$ and for $\epsilon < c'$

$$\mathbb{P}\left(\left|\phi\Big(\frac{\widehat{\boldsymbol{t}}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{t}}^f}{\sqrt{p}}\Big)\Big]\right| \geq \epsilon\right) \leq \frac{C}{\epsilon^2} e^{-cn\epsilon^4}. \tag{17}$$

We prove Lemma 10 in Section B.3.

3.4. *Control of the debiased Lasso.* Recall that the debiased Lasso with degrees-of-freedom adjustment is defined according to expression (6)

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

In this section, we aim to show that the debiased Lasso approximately follows a Gaussian distribution with mean $\boldsymbol{\theta}^*$ and covariance ${\tau^*}^2\boldsymbol{\Sigma}^{-1}$. The next theorem makes this statement precise.

THEOREM 11. *Under assumptions A1 and A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for any 1-Lipschitz $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\left|\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big]\right| > \epsilon\right) \leq \frac{C}{\epsilon^3} e^{-cn\epsilon^6},$$

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

We prove Theorem 11 in Section B.5.

Using a strategy like that in the proof of Corollary 7, one can show that the joint empirical distributions $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*, \widehat{\theta}_i^{\mathrm{d}}}$ and $p^{-1}\sum_{i=1}^p \delta_{\theta_i^*, \theta_i^* + \tau^*(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g})_i}$ both concentrate on the same distribution in the sense that they are close in weak-$*$ distance to the same distribution μ_* with high-probability.

Using Theorem 11, 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 Eq. (5). We refer the resulting intervals as the *debiased confidence intervals*.

COROLLARY 12. *Fix $q \in (0,1)$. For each $j \in [p]$, define the interval*

$$\mathsf{CI}_j^{\mathrm{d}} := \left[\widehat{\theta}_j^{\mathrm{d}} - \Sigma_{j|-j}^{-1/2}\widehat{\tau}(\lambda) z_{1-q/2},\ \widehat{\theta}_j^{\mathrm{d}} + \Sigma_{j|-j}^{-1/2}\widehat{\tau}(\lambda)\, z_{1-q/2}\right], \tag{18}$$

where $z_{1-q/2}$ is the $(1-q/2)$-quantile of the standard normal distribution, $\widehat{\tau}(\lambda)$ is given by Eq. (5), and

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

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 A2, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for all $\epsilon < c'$

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

We prove Corollary 12 in Section B.5.

It is worth emphasizing that the debiasing construction of Eq. (6) assumes that the population covariance $\mathbf{\Sigma}$ is known. In practice, $\mathbf{\Sigma}$ needs to be estimated from data. Accurate estimates can be produced under two scenarios: (i) When sufficiently strong information is known about the structure of $\mathbf{\Sigma}$ (for instance $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-1}$ are band diagonal or very sparse); (ii) When additional 'unlabeled' data $(\boldsymbol{x}_i')_{i\geq 1}$ is available.

REMARK 3.1. It is instructive to compare the degrees-of-freedom adjusted debiased Lasso of Eq. (6) with the more standard construction without adjustment [54, 52, 27, 26, 28]:

$$\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \mathbf{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}). \tag{19}$$

When $\|\widehat{\boldsymbol{\theta}}\|_0/n = o(1)$, the two constructions are comparable, namely $\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}} \approx \widehat{\boldsymbol{\theta}}^{\mathrm{d}}$. In this regime, the errors in estimating the nuisance $\boldsymbol{\theta}_{-j}^*$ negligibly degrade the precision of inference on θ_j^*.

In contrast, in the proportional asymptotic regime, it turns out that the errors in estimating $\boldsymbol{\theta}_{-j}^*$ do affect the precision of inference on θ_j^*. The denominator $1 - \|\widehat{\boldsymbol{\theta}}\|_0/n$ in Eq. (6) becomes crucial for correcting the bias induced by these errors.

3.5. *Confidence interval for a single coordinate.* While Theorem 11 and Corollary 12 establish coverage of the debiased confidence intervals $\mathsf{CI}_j^{\mathrm{d}}$ on average across coordinates, they do not provide sufficient control to establish the coverage of $\mathsf{CI}_j^{\mathrm{d}}$ for a fixed j. To illustrate the problem, we quantify the control Theorem 11 provides for a single coordinate of $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$. Theorem 11 implies that for any 1-Lipschitz $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, the difference $\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big]$ lies with high-probability in an interval of length $\tilde{O}(p^{-1/6})$, where $\tilde{O}$ hides factors which only depend on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ, or are poly-logarithmic in p. Applied to $\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) = (\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*)/\sqrt{p}$, this implies that the difference $\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*$ lies with high-probability in an interval of length $\tilde{O}(p^{2/3})$. Theorem 11 and Corollary 12 suggest that the typical fluctuations of $\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*$ are of order $O(1)$. Thus, the control of a single coordinate provided by Theorem 11 is at a larger scale than the scale of its typical fluctuations. A recent paper [10] controls a single coordinate at the relevant scale for $\delta > 1$, but leaves open the case $\delta \leq 1$. Addressing this case remains an open problem.

Instead, we provide a construction of confidence intervals for a single coordinate using a leave-one-out technique. We call these confidence intervals, defined below, the *leave-one-out* *confidence intervals*, denoted by $\mathsf{CI}_j^{\mathrm{loo}}$. We can write the observation vector $\boldsymbol{y}$ as

$$\boldsymbol{y} = \begin{pmatrix} \cdots & \breve{\boldsymbol{x}}_j & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ \theta_j^* \\ \vdots \end{pmatrix} + \sigma \boldsymbol{z} = \theta_j^* \breve{\boldsymbol{x}}_j + \boldsymbol{X}_{-j} \boldsymbol{\theta}_{-j}^* + \sigma \boldsymbol{z}\,, \tag{20}$$

where $\boldsymbol{X}_{-j} \in \mathbb{R}^{n\times(p-1)}$ denotes the original design matrix excluding the j-th column and $\breve{\boldsymbol{x}}_j$ denotes the j-th column. Define $\breve{\boldsymbol{x}}_j^{\perp} := \breve{\boldsymbol{x}}_j - \boldsymbol{X}_{-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j} \in \mathbb{R}^n$ so that $\breve{\boldsymbol{x}}_j^{\perp}$ is independent of $\boldsymbol{X}_{-j}$ (see Section C.2). According to decomposition (20),

$$\boldsymbol{y} = \boldsymbol{X}_{-j} \underbrace{(\boldsymbol{\theta}_{-j}^* + \theta_j^* \boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j})}_{=:\boldsymbol{\theta}_{\mathrm{loo}}^*} + \breve{\boldsymbol{x}}_j^{\perp}\theta_j^* + \sigma \boldsymbol{z}\,, \tag{21}$$

and

$$\breve{\boldsymbol{x}}_j^{\perp}\theta_j^* + \sigma\boldsymbol{z} \sim \mathsf{N}(0, \sigma_{\mathrm{loo}}^2 \mathbf{I}_n) \ \text{ with } \ \sigma_{\mathrm{loo}}^2 := \sigma^2 + \frac{\Sigma_{j|-j}\theta_j^{*2}}{n}\,,$$

where $\Sigma_{j|-j} = \Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}\boldsymbol{\Sigma}_{-j,j}$. Expression (21) can be viewed as defining a linear-model with $p-1$ covariates, with true parameter $\boldsymbol{\theta}_{\mathrm{loo}}^*$, and noise variance σ_{loo}^2. We call this the *leave-one-out model*. Let τ_{loo}^*, ζ_{loo}^* be the solution to the fixed point equations (8a) and (8b) in the leave-one-out model.

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

$$\xi_j := \frac{(\breve{\boldsymbol{x}}_j^{\perp})^\top(\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}})}{\Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}\|_0/n)}\,. \tag{22}$$

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., $\breve{\boldsymbol{x}}_j^{\perp}$), and a sample regression of the outcome $\boldsymbol{y}$ on the other features (i.e., $\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$). When $\theta_j^* = 0$, these residuals will be independent. Indeed, $\breve{\boldsymbol{x}}_j^{\perp}$ is independent of $(\boldsymbol{y}, \boldsymbol{X}_{-j})$, and because $\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$ is a function of $(\boldsymbol{y}, \boldsymbol{X}_{-j})$, $\breve{\boldsymbol{x}}_j^{\perp}$ is also independent of $\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$. The theory from the preceding sections allows us to quantify the distribution of the variable importance statistic ξ_j even when $\theta_j^* \neq 0$ and so permits the construction of confidence intervals.

Similarly to $\widehat{\tau}(\lambda)$ defined in Eq. (5), we estimate the effective noise level in the leave-one-out-model by

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

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

$$\mathsf{CI}_j^{\mathrm{loo}} := \left[\xi_j - \Sigma_{j|-j}^{-1/2}\widehat{\tau}_{\mathrm{loo}}^j z_{1-\alpha/2},\ \ \xi_j + \Sigma_{j|-j}^{-1/2}\widehat{\tau}_{\mathrm{loo}}^j z_{1-\alpha/2}\right]\,. \tag{23}$$

We demonstrate below that this confidence interval $\mathsf{CI}_j^{\mathrm{loo}}$ achieves approximate coverage for every fixed j, whose proof is provided in Section B.6.2.

THEOREM 13. *Assume $p \geq 2$. Let $\delta_{\mathrm{loo}} = n/(p-1)$. Assume λ, $\boldsymbol{\Sigma}$, and σ satisfy assumption A1, and that $\boldsymbol{\theta}_{-j}^*$ is $(s, \sqrt{\delta_{\mathrm{loo}}}(1-\Delta_{\min}), M)$-approximately sparse with respect to covariance $\boldsymbol{\Sigma}_{-j,-j}$ for some $s/(p-1) \geq \nu_{\min} > 0$ and $1 \geq \Delta_{\min} > 0$. Let $M' > 0$ be such that $|\theta_j^*| \leq M'(p-1)^{1/4}$.*

(a) (Coverage and power of the leave-one-out confidence interval) *There exist constants* $C, c, c' > 0$ *depending only on* $\mathcal{P}_{\mathrm{model}}$*,* $\nu_{\min}$*,* $\Delta_{\min}$*,* M*,* M'*, and* δ_{loo} *such that for all* $\epsilon < c'$*,*

(24)

$$\left|\mathbb{P}_{\theta_j^*}\left(\theta \notin \mathsf{CI}_j^{\mathrm{loo}}\right) - \mathbb{P}_{\theta_j^*}\left(|\theta_j^* + \tau_{\mathrm{loo}}^* G - \theta| > \tau_{\mathrm{loo}}^* z_{1-\alpha/2}\right)\right| \le C\left((1+|\theta_j^*|)\epsilon + \frac{1}{\epsilon^3}e^{-cn\epsilon^6} + \frac{1}{n\epsilon^2}\right),$$

where $G \sim \mathsf{N}(0,1)$*. (See discussion following 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* $\mathcal{P}_{\mathrm{model}}$*,* $\nu_{\min}$*,* $\Delta_{\min}$*,* M*,* M'*, and* δ_{loo} *such that for all* $\epsilon < c'$*,*

$$\mathbb{P}_{\theta_j^*}\left(\left|\frac{\widehat{\tau}_{\mathrm{loo}}^j}{\tau_{\mathrm{loo}}^*} - 1\right| > \epsilon\right) \le \frac{C}{\epsilon^3}e^{-cn\epsilon^6}. \tag{25}$$

Note that $\mathbb{P}\left(|\theta_j^* + \tau_{\mathrm{loo}}^* G - \theta| > \tau_{\mathrm{loo}}^* z_{1-\alpha/2}\right)$ is the power of the standard two-sided confidence interval under Gaussian observations $\theta_j^* + \tau_{\mathrm{loo}}^* G$ against alternative θ. The left-hand side of (24) does not depend on ϵ, so that the optimal bound is found by choosing $\epsilon < c'$ which minimizes the right-hand side. When $|\theta_j^*| = o(n^{1/6}/\log n)$, the right-hand side can be made small by for example, taking $n^{-1/6}\log n \ll \epsilon \ll \min\{c', 1/|\theta_j^*|\}$.

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. Towards this, it is useful to briefly recall this construction and discuss the relative merits of our approach.

In general, when the feature distribution is known, one can perform an exact test of

$$\boldsymbol{y} \perp\!\!\!\perp \breve{\boldsymbol{x}}_j \mid \boldsymbol{X}_{-j}, \tag{26}$$

even without Gaussianity or any assumption on the conditional distribution of the outcome $\boldsymbol{y}$ given the features $\boldsymbol{X}$ (see, e.g., [15, 30, 33]). 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 $T(\boldsymbol{y}, \boldsymbol{X}) \mid \boldsymbol{X}_{-j}$ is equal to the distribution of $T(\boldsymbol{y}, \boldsymbol{x}_j', \boldsymbol{X}_{-j})$ 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, 30, 33] for more details about how these observations lead to the construction of an exact test.

When the linear model is well-specified, the null (26) corresponds to $\theta_j^* = 0$, and our leave-one-out procedure implements the conditional randomization test under this null, as we now explain. The statistic ξ_j, defined in Eq. (22) 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. We make a few remarks. 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 $\breve{\boldsymbol{x}}_j^{\perp}$ is independent of $(\boldsymbol{y}, \boldsymbol{X}_{-j}, \widehat{\boldsymbol{\theta}}_{\mathrm{loo}})$ under the null $\theta_j^* = 0$, one has

$$\xi_j | \boldsymbol{y}, \boldsymbol{X}_{-j} \sim \mathsf{N}\left(0, \Sigma_{j|-j}^{-1}(\widehat{\tau}_{\mathrm{loo}}^j)^2\right).$$

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 $0 \notin \mathsf{CI}_j^{\mathrm{loo}}$ is exactly the conditional randomization test for the null (26) based on the variable importance statistic $|\xi_j|$.[2]

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

As a consequence, the leave-one-out confidence intervals have exact finite sample coverage under the null $\theta_j^* = 0$.

Moreover, Theorem 13 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.

The linearity assumption in our setting allows us to push this rationale further. For any ω, when $\theta_j^* = \omega$, the j^{th} residualized covariate $\breve{\boldsymbol{x}}_j^\perp$ is independent of the "pseudo-outcome" $\boldsymbol{y} - \omega \breve{\boldsymbol{x}}_j^\perp$ and $\boldsymbol{X}_{-j}$. By computing a Lasso estimate using this pseudo-outcome in place of $\boldsymbol{y}$, the statistician may perform an exact test of the null $\theta_j^* = \omega$. The inversion of this collection of tests, indexed by ω, produces an exact confidence interval. Details of this construction are provided in Appendix B.6.

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 $\boldsymbol{y} - \omega \breve{\boldsymbol{x}}_j^\perp$ for each value of ω. 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 4.2 show that coverage is good already for n, p, s on the order of 10s or 100s. An additional benefit of Theorem 13 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 13(a) applied under θ_j^* provides an estimate of the power of the conditional randomization test under alternative $\theta_j^* = \omega$.

We conjecture that the exact confidence intervals outlined above, the leave-one-out confidence intervals, and the debiased confidence intervals asymptotically agree up to negligible terms. Our simulations in Section 4.2 support this conjecture in the case of the equivalence of the leave-one-out confidence and the debiased confidence intervals.

4. Numerical simulations.

This section contains numerical experiments which *(i)* suggest that the large s, n, p behavior established in this paper is a good description of the debiased Lasso even for s, n, p on the order of 10s or 100s, *(ii)* demonstrate the importance of the degrees-of-freedom adjustment, *(iii)* provide evidence for the necessity of the approximate sparsity constraint in Claim 2 for the results of Theorems 4 and 8, and *(iv)* present numerical evidence that our results may hold for a broader spectrum of feature distributions going beyond Gaussian designs. We present here some representative simulations.

In our simulations, we adopt the normalization $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}/p)$ rather than $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}/n)$ as is adopted in the theoretical development of this paper. This amounts to a simple change of variables. We prefer the normalization $\boldsymbol{\Sigma}/p$ to make the dependence of performance metrics on n more interpretable: indeed, under this normalization increasing n for p fixed does not affect the normalization of each row of $\boldsymbol{X}$ and thus better models the collection of additional samples or measurements.

4.1. *Debiasing with degrees-of-freedom adjustment.*

We compare the degrees-of-freedom adjusted debiased Lasso of Eq. (6) with the standard unadjusted estimator of Eq. (19).

Figure 1 reports results on the distribution of the two estimators. We set $p = 100$, $n = 25$, and $s = 20$, and fix $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with $s/2$ coordinates equal to 25 and the rest equal to -25 chosen uniformly at random. We repeat the following $N_{\text{sim}} = 1000$ times. First, we generate data from the linear model (2) where $\boldsymbol{x}_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}/n)$, $\sigma = 1$ and $\boldsymbol{\Sigma}$ comes from the autoregressive model $\mathsf{AR}(0.5)$:

$$\Sigma_{ij} = 0.5^{|i-j|}.$$

In each simulation, we use the same $\boldsymbol{\theta}^*$ vector but independent draws of $\boldsymbol{X}, \boldsymbol{z}$. We compute for each $j \leq n$ the values $\frac{\Sigma_{j|-j}^{1/2}(1-\|\widehat{\boldsymbol{\theta}}\|_0/n)(\widehat{\theta}_j^{\mathrm{d}}-\theta_j^*)}{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}}$ and $\frac{\Sigma_{j|-j}^{1/2}(\widehat{\theta}_{0j}^{\mathrm{d}}-\theta_j^*)}{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}}$ corresponding to the debiased Lasso with and without degrees-of-freedom adjustment respectively. Aggregating over coordinates and simulations (giving $p \cdot N_{\mathrm{sim}} = 10^5$ observations of single coordinates), we plot histograms and quantile plots for all coordinates corresponding to $\theta_j = -25,\, 0,\, 25$ separately. In the quantile plots, the empirical quantiles are compared with the theoretical quantiles of the standard normal distribution $\mathsf{N}(0, 1)$.

Without the degrees-of-freedom adjustment, visible deviations from normality occur. For active coordinates, we observe bias and skew; for inactive coordinates, we observe tails which are too fat. The fattening of the tails occurs around and beyond the quantiles corresponding to two-sided confidence intervals constructed at the 0.05 level. Thus, failure to implement degrees-of-freedom adjustments will lead to under-coverage in standard statistical practice even prior to corrections for multiple testing. In contrast, with degrees-of-freedom adjustment, no obvious deviations from normality occur for either the inactive or active coordinates. Normality is retained well into the normal tail. Since we take $s = 20$, $n = 25$ and $p = 100$, our simulations suggest approximate normality already for s, n, p on the order of 10s and 100s. Our simulations are well outside the condition $s = \tilde{o}(n^{2/3})$ required by [9] or $n > p$ required by [9, 10].

The simulations presented Figure 1 are representative of simulations conducted at various parameter settings.

4.2. *Confidence interval for a single coordinate.* In this section we consider the behavior of the debiased confidence interval $\mathsf{CI}_j^{\mathrm{d}}$ (defined in Eq. (18)) and leave-one-out confidence interval $\mathsf{CI}_j^{\mathrm{loo}}$ (defined in Eq. (23)).

In Figure 2, we examine the coverage of the confidence interval for both an active coordinate and an inactive coordinate. As in Figure 1, we consider $p = 100$, $n = 25$, and $s = 20$, and fix $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with $s/2$ coordinates equal to 25 and the rest equal to -25. The locations of the active coordinates are chosen uniformly at random. We set the coordinate of interest to be θ_{50}. For each model specification, we perform the following $N_{\mathrm{sim}} = 1000$ times. First, we generate data from the linear model (2) with $\sigma = 1$ and $\boldsymbol{\Sigma}$ the $\mathsf{AR}(0.5)$ covariance $\Sigma_{ij} = 0.5^{|i-j|}$. We construct for $j = 50$ the $(1-\alpha)$-confidence intervals $\mathsf{CI}_j^{\mathrm{d}}$ and $\mathsf{CI}_j^{\mathrm{loo}}$ at level $\alpha = 0.05$. We also construct the following interval based on the debiased Lasso without degrees-of-freedom adjustment given by Eq. (19):

$$\mathsf{CI}_j^{\mathrm{d,noDOF}} := \left[\widehat{\theta}_{0j}^{\mathrm{d}} - \frac{\Sigma_{j|-j}^{-1/2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}} z_{1-\alpha/2},\;\; \widehat{\theta}_{0j}^{\mathrm{d}} + \frac{\Sigma_{j|-j}^{-1/2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}} z_{1-\alpha/2}\right].$$

The confidence intervals from the first 40 of the 1000 simulations are plotted in Figure 2 for the cases $\theta_{50}^* = 0$ and $\theta_{50}^* = 25$. Both the debiased Lasso and the leave-one-out confidence intervals achieve coverage, and these two confidence intervals approximately agree. In contrast, when $\theta_{50} = 25$, the confidence interval without degrees-of-freedom adjustment is uncentered and too narrow, leading to large under-coverage. When $\theta_j^* = 0$, the empirical coverage (for 1000 simulations) is 95.4% for the debiased Lasso with degrees-of-freedom adjustment, 95% for leave-one-out confidence interval, and 93.7% for the debiased Lasso without degrees-of-freedom adjustment. When $\theta_j^* = 25$, these coverages are 93.7%, 93.8%, and 78.3%, respectively.

These simulations provide evidence that the leave-one-out confidence intervals $\mathsf{CI}_j^{\mathrm{loo}}$ are valid for fixed coordinate j, already for moderate values of n, p, and not only for large n, p as guaranteed by Theorem 13. Further, the debiased confidence intervals $\mathsf{CI}_j^{\mathrm{d}}$ also appear

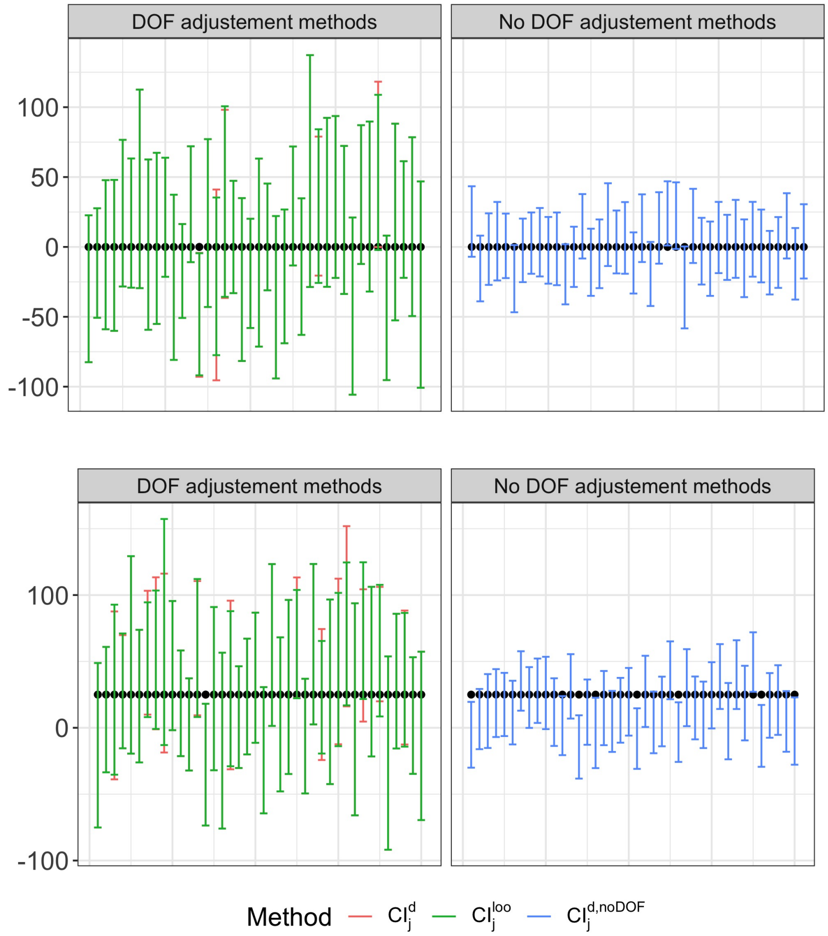


Fig 2. Confidence interval for a single coordinate. Here $p = 100$, $n = n = 25$, $s = 20$, $\Sigma_{ij} = 0.5^{|i-j|}$, $\lambda = 4$, $\sigma = 1$. In the top plots, the truth is $\theta_j^* = 0$, and in the bottom plots the truth is $\theta_j^* = 25$.

to achieve coverage per-coordinate and not only on average even though our theory does not establish this. Finally, these simulations support that $\mathsf{CI}_j^{\mathrm{d}}$ and $\mathsf{CI}_j^{\mathrm{loo}}$ are asymptotically equivalent.

4.3. *Approximate sparsity and Gaussian width.* Recall that our main results rely on assumption A2 on the solution (τ^*, ζ^*) of Eqs (8a), (8b) being uniformly bounded above and below. Theorem 2 establish that assumption A2 holds if $\boldsymbol{\theta}^*$ is $(\boldsymbol{x}, M)$-approximately sparse for some $\boldsymbol{x} \in \{+1, 0, -1\}^p$ and a constant M the corresponding Gaussian width satisfies $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \leq \sqrt{\delta}(1 - \Delta_{\min})$ for some $\Delta_{\min} > 0$ (plus some additional technical condition).

In Figure 3, we explore the role (and tightness) of this Gaussian width condition. Again, we let $\boldsymbol{\Sigma}$ be the AR(0.5) covariance matrix: $\Sigma_{ij} = 0.5^{|i-j|}$. We set $p = 1000$ and construct $\boldsymbol{\theta}^*$ as follows: we choose a support $S \subseteq [p]$ uniformly at random with $s_0 := |S| = 200$, and set $\theta_i^* = 0$ for $i \in [p] \setminus S$ and $\theta_i^* \sim \mathsf{Unif}(\{+\mu, -\mu\})$ for $i \in S$. We view this as a $(\boldsymbol{x}, M)$ approximately sparse vector with $\boldsymbol{x} = \mathrm{sign}(\boldsymbol{\theta}^*)$, and $M = 0$.

We approximate the Gaussian width $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma})$ by Monte Carlo sampling. In order to do that, we generate 500 realizations of the optimization problem (12) and obtain 500 estimates $\{\widehat{\mathcal{G}^*}\}_{i=1}^{500}$. We plot a histogram of $\{p\widehat{\mathcal{G}^*}^2\}_{i=1}^{500}$ in both plots in Figure 3 as well as the median of these values as a vertical dashed line.[3] On the same plots, we plot the logarithm of the ℓ_2-risk and of the sparsity $\|\widehat{\boldsymbol{\theta}}\|_0/n$ as a function of n for four different magnitudes μ of the active coordinates of $\boldsymbol{\theta}^*$. Each point on the curves is generated by taking the median over 50 simulations.

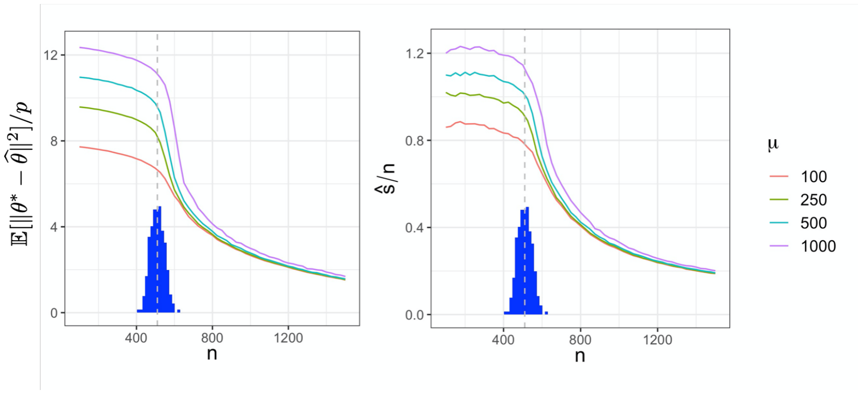


Fig 3. The Lasso risk and sparsity are uncontrolled when the $\boldsymbol{\theta}^*$ grows on a signed support set whose Gaussian width squared exceeds the aspect ratio n/p. Histogram shows standard Gaussian width $\mathcal{G}_{\mathrm{std}}(\boldsymbol{x}, \boldsymbol{\Sigma})$ as computed numerically over 500 trials in simulation. $p = 1000$, $s = 200$, $\Sigma_{ij} = .5^{|i-j|}$, $\lambda = 4$, $\sigma = 1$. The support set is chosen uniformly at random, and half of the active coordinates are positive, chosen uniformly at random.

For $n/p \leq \widehat{\mathcal{G}^*}^2$, the risk grows very rapidly with μ, whereas for $n/p \geq \widehat{\mathcal{G}^*}^2$, the risk grows only moderately with μ (if at all). There is a visually sharp transition in behavior at the threshold $n/p \approx \mathrm{med}(\widehat{\mathcal{G}^*})^2$. Similarly, when $n/p \leq \mathrm{med}(\widehat{\mathcal{G}^*})^2$, the sparsity $\|\widehat{\boldsymbol{\theta}}\|_0/n$ is equal to or greater than 1, whereas for $n/p \geq \mathrm{med}(\widehat{\mathcal{G}^*})^2$, this quantity is bounded away from 1 and does not substantially grow with μ. By the stationarity conditions for the Lasso, we know that $/\|\widehat{\boldsymbol{\theta}}\|_0/n \leq 1$ always. The observed value $\|\widehat{\boldsymbol{\theta}}\|_0/n > 1$ indicate that the Lasso in this regime is difficult to solve numerically[4].

Note that $\mathrm{med}(\widehat{\mathcal{G}^*})$ is an estimator standard Gaussian width $\mathcal{G}_{\mathrm{std}}(\boldsymbol{x}, \boldsymbol{\Sigma})$ (see Eq. (12)) instead of the functional Gaussian width (10) (see Eq. (10)) which enters our theory. However, these simulations should be interpreted in light of the conjecture that $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \approx \mathcal{G}_{\mathrm{std}}(\boldsymbol{x}, \boldsymbol{\Sigma})$.

4.4. *Non-Gaussian designs.* The results described in this work are proven under correlated Gaussian designs. When covariates are independent, numerical simulations and universality arguments in previous work suggest exact asymptotic characterizations still hold for independent but possibly non-Gaussian covariates (see e.g. [5, 39, 36] for rigorous universality results). Moreover, such universality phenomena are also expected to hold beyond the linear models: for instance, [47] (in Figure 9) present simulations for logistic regression with independent but non-Gaussian covariates whose behavior agrees with the corresponding asymptotic predictions for independent Gaussian covariates. Nevertheless, these predictions are incorrect when covariates are correlated. This suggests that the most severe limitation of the existing exact asymptotic theory is not the Gaussianity assumption but rather the independence assumption. It is this assumption that the current paper weakens.

[3]We normalize the height of the histogram to fit on our plots.

[4]We use the glmnet package for all Lasso simulations.

Here we provide some numerical evidence which suggests that our theory describes the behavior of the Lasso under some realistic data generating distributions (when the Gaussianity assumption breaks). Theoretical investigations of universality is left for future work. We consider the design matrix with covariates generated according to a hidden Markov model. Hidden Markov models are frequently used for modeling the covariates in genetics applications (see, e.g. [45]). The specification of the hidden Markov model used in our simulation is described in details in Appendix D. The specification is such that covariates with indices differing by approximately 10 or less have non-negligible correlation. The response is generated according to model (2), with $n = 1280,\ p = 2000,\ s_0 = 0.128p$, and $\sigma = 0.2$, and all active coordinates of $\boldsymbol{\theta}^*$ are set to 1. We run our debiasing procedure with degrees of freedom adjustment for $N_{\text{sim}} = 10$ independent generations of the data, with the knowledge of the underlying population covariance matrix for the covariates. We then aggregate the standardized and centered debiased Lasso estimates across coordinates and across simulations, separately for the inactive and active coordinates, and provide a qq-plot for each; the results are presented in Figure 4. It is worth noting that from the simulations, one can see the success of the debiasing procedure with degrees of freedom adjustment carries even into the tails of the distribution. This phenomenon cannot be justified using prior theory based on independent Gaussian covariates.

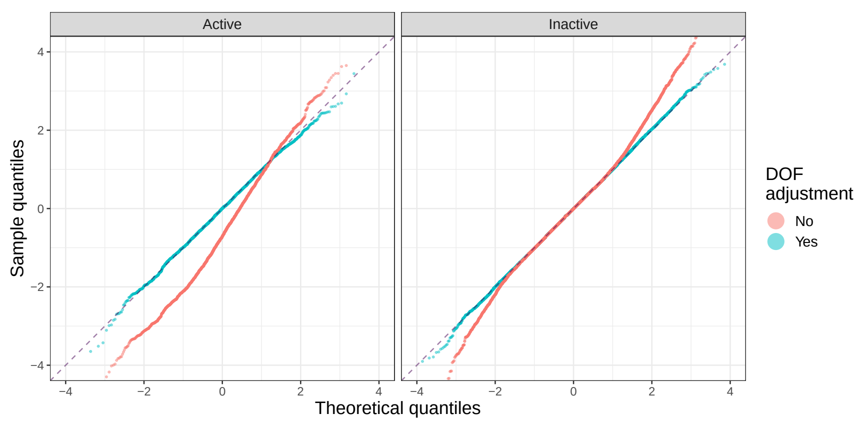


Fig 4. The debiased Lasso with and without degrees-of-freedom (DOF) adjustment for hidden Markov model features. Here $n = 1280, p = 2000, s_0 = .128 \cdot p$, and $\sigma = .2$, and all active coordinates of $\boldsymbol{\theta}^*$ equal to 1. Quantiles and densities are compared with the ones of the standard normal distribution.

5. Main proof ingredients. Our proofs are built upon a tight version of Gordon's min-max theorem for convex functions. Gordon's original theorem [24, 25] is a Gaussian comparison inequality for the minimization-maximization of two related Gaussian processes, and has several applications in random matrix theory and convex optimization [44, 41]. In a line of work initiated by [46] and formalized by [49], the comparison inequality was shown to be tight when the underlying Gaussian process is convex-concave. This observation has led to several works establishing exact asymptotics for high-dimensional convex procedures, including general penalized M-estimators in linear regression [49, 48] and binary classification [18, 37, 32]. (We also refer to [6, 2, 19, 20, 42, 3] for alternative proof techniques to obtain sharp results in high-dimensional regression models, in the proportional asymptotics.)

Earlier work has so far focused on the case of independent features or correlated features with unpenalized or ridge-penalized procedures. Analyzing the Lasso estimator under general Gaussian designs, however, requires overcoming several technical challenges, as the ℓ_1-penalty breaks the isometry underlying procedure. In this section, we summarize our proof strategy, emphasizing the technical innovations that are required in the context of general correlated designs. Our work builds on the approach of [35], which studied the Lasso and debiased Lasso estimators in the case $\mathbf{\Sigma} = \mathbf{I}_p$.

Control of the Lasso estimate. We find it useful to first rewrite the Lasso optimization objective as

$$\mathcal{C}(\boldsymbol{v}) := \frac{1}{2n}\|\sigma \boldsymbol{z} - \boldsymbol{X}\mathbf{\Sigma}^{-1/2}\boldsymbol{v}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{\theta}^* + \mathbf{\Sigma}^{-1/2}\boldsymbol{v}\|_1 .$$

Here we introduce the *prediction error* vector $\boldsymbol{v} := \mathbf{\Sigma}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. The variable $\boldsymbol{v}$ is used to whiten the design matrix and isolate the dependence of the objective on it. Indeed, $\boldsymbol{X}\mathbf{\Sigma}^{-1/2}$ has entries distributed i.i.d. from $\mathsf{N}(0, 1/n)$, and we have expanded $\boldsymbol{y}$ to reveal its dependence on $\boldsymbol{X}$. We denote by $\widehat{\boldsymbol{v}}$ the minimizer of $\mathcal{C}(\boldsymbol{v})$, i.e., $\widehat{\boldsymbol{v}} := \mathbf{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$. By a standard argument, Gordon's min-max theorem implies that the Lasso optimization behaves, in a certain sense, like the optimization of the simpler objective

$$\mathcal{L}(\boldsymbol{v}) := \frac{1}{2}\left(\sqrt{\sigma^2 + \frac{\|\boldsymbol{v}\|_2^2}{n}} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle}{n}\right)_+^2 + \frac{\lambda}{n}\left(\|\boldsymbol{\theta}^* + \mathbf{\Sigma}^{-1/2}\boldsymbol{v}\|_1 - \|\boldsymbol{\theta}^*\|_1\right),$$

which we call *Gordon's objective*. The precise statement is as follows.

LEMMA 5.1 (Gordon's lemma). *The following hold.*

(a) *Let $D \subset \mathbb{R}^p$ be a closed set. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{C}(\boldsymbol{v}) \le t\right) \le 2\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{L}(\boldsymbol{v}) \le t\right).$$

(b) *Let $D \subset \mathbb{R}^p$ be a closed, convex set. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{C}(\boldsymbol{v}) \ge t\right) \le 2\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{L}(\boldsymbol{v}) \ge t\right).$$

By studying Gordon's objective, and comparing the value of $\min_{\boldsymbol{v}\in D}\mathcal{L}(\boldsymbol{v})$ for suitable choices of the set D, we can extract properties of $\widehat{\boldsymbol{v}}$ and hence $\widehat{\boldsymbol{\theta}}$. In particular, in Theorem 6, we compare the value taken for $D = \mathbb{R}^p$ and

$$D = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \;\middle|\; \left|\phi\Big(\frac{\boldsymbol{\theta}^* + \mathbf{\Sigma}^{-1/2}\boldsymbol{v}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^f}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big]\right| > \epsilon\right\},$$

where $\widehat{\boldsymbol{\theta}}^f$ is defined by Eq. (9) with τ^*, ζ^* the unique solution to Eqs. (8a) and (8b). The argument is carried out in detail in Appendix B.1.

This discussion clarifies that we can control the behavior of the Lasso objective only insofar as we can control the behavior of Gordon's objective. The major technical challenge to apply this approach to general correlated designs is in relating the minimizer of Gordon's objective to the fixed design estimator $\widehat{\boldsymbol{\theta}}^f$. In particular, this requires showing that the solution (τ^*, ζ^*) of Eqs. (8a) and (8b) is unique and bounded in terms of simple model parameters (see Theorems 1 and 2).

Generalizing an idea introduced in [37], we control the solutions Eqs. (8a) and (8b) by showing that these equations are the KKT conditions for a certain convex optimization problem on the infinite dimensional Hilbert space $L^2(\mathbb{R}^p;\mathbb{R}^p)$. To be more specific, the optimization problem is

$$\min_{\boldsymbol{v}\in L^2} \mathscr{E}(\boldsymbol{v}) := \min_{\boldsymbol{v}\in L^2} \left\{ \frac{1}{2}\Big(\sqrt{\frac{\|\boldsymbol{v}\|_{L^2}^2}{n} + \sigma^2} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}}{n}\Big)_+^2 + \frac{\lambda}{n}\mathbb{E}\left[\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})\|_1 - \|\boldsymbol{\theta}^*\|_1\right]\right\}.$$

The objectives $\mathcal{L}$ and $\mathscr{E}$ are closely related, but their arguments belong to different spaces. The objective $\mathcal{L}$ takes vectorial arguments $\boldsymbol{v}\in\mathbb{R}^p$; the objective $\mathscr{E}$ takes functional arguments $\boldsymbol{v}:\mathbb{R}^p\to\mathbb{R}^p$. Both objectives are convex. In Appendix A.3, we show that $\boldsymbol{v}\in L^2$ is a minimizer of $\mathscr{E}$ if and only if $\boldsymbol{v}(\boldsymbol{g}) = \eta(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g};\zeta^*)$ for τ^*, ζ^* a solution to the fixed point equations. This follows from showing that Eqs. (8a) and (8b) correspond to KKT conditions for the minimization of $\mathscr{E}$. Further, we show that $\mathscr{E}$ diverges to infinity as $\|\boldsymbol{v}\|_{L^2}\to\infty$ and is strictly convex in a neighborhood of any minimizer, whence a minimizer exists, and it is unique (Theorem 1). We are then able to conclude that the fixed point equations also have a unique solution. We defer the details of this argument to Appendix A.3.

Controlling the size of the fixed point parameters (Theorem 2) relies on bounding the norm of the minimizer of $\mathscr{E}$. Again, our approach is geometric: rather than analyzing the fixed point equations directly, we study the growth of the objective $\mathscr{E}$ as $\|\boldsymbol{v}\|_{L^2}/\sqrt{n}$ diverges. The functional Gaussian width (10) controls this growth. This explains the centrality of the Gaussian width $\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})$ in our analysis. In fact, under only a sparsity constraint on $\boldsymbol{\theta}^*$, we can control the growth $\mathscr{E}$ in $\|\boldsymbol{v}\|_{L^2}/\sqrt{n}$ in an n-independent way only when $\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma}) < \sqrt{\delta}$ where $\boldsymbol{x}\in\partial\|\boldsymbol{\theta}^*\|_1$. Thus, we suspect our analysis based on the Gaussian width is in a certain sense tight, though we do not attempt to make this claim precise. The detailed argument bounding the fixed point parameters is in Appendix A.4.

The present approach is significantly more general both than the one of [35], which studies the Lasso for $\boldsymbol{\Sigma} = \mathbf{I}_p$, and of [37] which studies binary classification under a ridge-type regularization. When $\boldsymbol{\Sigma} = \mathbf{I}_p$, the Lasso estimator in the fixed-design model is separable, and Eqs. (8a) and (8b) simplify because

$$\mathsf{R}(\tau^2,\zeta) = \frac{1}{\delta}\mathbb{E}_{\Theta,G}[(\eta_{\text{soft}}(\Theta^* + \tau G, \lambda/\zeta) - \Theta^*)^2]\,,$$
$$\mathsf{df}(\tau^2,\zeta) = \frac{1}{\delta}\mathbb{P}(\eta_{\text{soft}}(\Theta^* + \tau G, \lambda/\zeta) \neq 0)\,,$$

where $\Theta^*\sim\frac{1}{p}\sum_{j=1}^p \delta_{\theta_j^*}$ independent of $G\sim\mathsf{N}(0,1)$, and $\eta_{\text{soft}}(y;\zeta) := (|y|-\zeta)_+\text{sign}(y)$. Hence—in that case—existence and uniqueness of the solution of Eqs. (8a) and (8b) can be proved by analyzing the explicit form of these equations.

Also, our approach is more general than the one of [37], which constructs a Hilbert-space optimization problem by taking the $n,p\to\infty$ limit of the Gordon's problem. In the present case, since we intend to establish a non-asymptotic control, for finite n,p there is no natural sequence of covariances in which to embed $\boldsymbol{\Sigma}$.

Furthermore, we generalize our result to achieve a uniform control over the penalization parameter $\lambda\in[\lambda_{\min},\lambda_{\max}]$ (see Theorem 6). The argument is based on a careful analysis of the sensitivity of the Lasso problem and its corresponding solution regarding the penalization parameter λ. More details can be found in Appendix B.7.

Control of the Lasso sparsity. It is not feasible to directly control quantity $\|\widehat{\boldsymbol{\theta}}\|_0/n$ using Theorem 4 with $\phi(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \|\boldsymbol{\theta}\|_0/n$, since this function is not Lipschitz or even continuous. Instead, we establish lower and upper bounds on the sparsity separately.

To explain the argument, define for any $\boldsymbol{\theta} \in \mathbb{R}^p$ the ϵ-*strongly active* coordinates of $\boldsymbol{\theta}$ to be $\{j \in [p] \mid |\theta_j| > \epsilon\}$. Likewise, for any $\boldsymbol{t} \in \mathbb{R}^p$ define the ϵ-*strongly inactive* coordinates of $\boldsymbol{t}$ to be $\{j \in [p] \mid |t_j| < 1-\epsilon\}$ (this definition is motivated by the fact that if $\boldsymbol{t}$ is the sub-gradient of the Lasso, if $|t_j| < 1-\epsilon$ then $\theta_j = 0$ and t_j would have to change by at least ϵ for θ_j to become active). Our argument relies on the following two facts (here $\widehat{\boldsymbol{\theta}}$ is, as always, the Lasso estimate, and $\widehat{\boldsymbol{t}}$ is the subgradient of Eq. (14)):

(27)
$$\text{if } \frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n} \le 1-\zeta^*-\epsilon, \text{ then } \inf_{\boldsymbol{\theta}} \left\{ \frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2 \;\Big|\; \frac{|\{j \mid |\theta_j| > \epsilon\}|}{n} > 1-\zeta^*-\frac{\epsilon}{2} \right\} \ge \sqrt{\frac{\delta\epsilon^3}{2}},$$

and

(28)
$$\text{if } \frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n} \ge 1-\zeta^*+\epsilon, \text{ then } \inf_{\boldsymbol{t}} \left\{ \frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{t}}-\boldsymbol{t}\|_2 \;\Big|\; \frac{|\{j \mid |t_j| < 1-\epsilon\}|}{n} > 1-\zeta^*-\frac{\epsilon}{2} \right\} \ge \sqrt{\frac{\delta\epsilon^3}{2}}.$$

The first argument holds because the vectors $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ differ by at least ϵ in $n\epsilon/2$ coordinates; namely, in those coordinates in which $\boldsymbol{\theta}$ is ϵ-strongly active and $\widehat{\boldsymbol{\theta}}$ is inactive. The second argument holds similarly. In words, vectors which are very sparse are separated in Euclidean distance from vectors with many ϵ-active coordinates; similarly, subgradients with many active coordinates are separated in Euclidean distance from vectors with many ϵ-inactive coordinates.

To proceed, we leverage the following fact: for any set $D \subset \mathbb{R}^p$ which contains the fixed-design Lasso estimate $\widehat{\boldsymbol{\theta}}^f$ with high-probability, the random design Lasso estimate $\widehat{\boldsymbol{\theta}}$ is close to D with high-probability. Similarly, for any set $D \subset \mathbb{R}^p$ which contains the fixed-design subgradient $\widehat{\boldsymbol{t}}^f$ with high-probability, the random-design subgradient $\widehat{\boldsymbol{t}}$ is close to D with high-probability. In the case of the subgradient, this is made precise in the statement of Lemma 10. A similar statement holds for the Lasso estimate, and developed in the proof of Theorem 4. Taking D to be the set over which the infimum in Eq. (27) (resp. Eq. (28)) is taken, we can conclude $\|\widehat{\boldsymbol{\theta}}\|_0/n > 1-\zeta^*-\epsilon$ (resp. $\|\widehat{\boldsymbol{\theta}}\|_0/n < 1-\zeta^*+\epsilon$) with high-probability as soon as we can show $\widehat{\boldsymbol{\theta}}^f \in D$ (resp. $\widehat{\boldsymbol{t}}^f \in D$) with high-probability. The details of this argument are carried out in Appendix B.4.

Control of the debiased Lasso. We may write the debiased Lasso as a function of the Lasso estimate $\widehat{\boldsymbol{\theta}}$, the subgradient $\widehat{\boldsymbol{t}}$, and the Lasso sparsity $\|\widehat{\boldsymbol{\theta}}\|_0/n$:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \frac{\lambda \boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{t}}}{1-\|\widehat{\boldsymbol{\theta}}\|_0/n}.$$

Because $1-\|\widehat{\boldsymbol{\theta}}\|_0/n$ concentrates on ζ^* by Theorem 9, the debiased Lasso is with high-probability close to

$$\widehat{\boldsymbol{\theta}} + \frac{\lambda}{\zeta^*}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{t}}. \tag{29}$$

Our goal is to show that $\widehat{\boldsymbol{\theta}} + \frac{\lambda}{\zeta^*}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{t}} - \boldsymbol{\theta}^*$ is approximately Gaussian noise with zero mean and covariance $\tau^{*2}\boldsymbol{\Sigma}^{-1}$. Heuristically, if we replace the Lasso estimate and subgradient by their fixed-design counterparts, we get

$$\widehat{\boldsymbol{\theta}}^f + \frac{\lambda}{\zeta^*}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{t}}^f - \boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}^f - \boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f) = \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g},$$

where in the first inequality we have used that $\frac{\lambda}{\zeta^*}\widehat{\boldsymbol{t}}^f = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}})$ by the KKT conditions for the optimization (4). Thus, we would like to justify the heuristic replacement of the random design quantities with their fixed-design counterparts.

It turns out that it is not straightforward to justify this heuristic. Theorem 4 and Lemma 10 compare the distributions of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{t}}$ to their fixed design counterparts individually but not jointly. That is, Theorem 4 compares the distribution of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^f$, and Lemma 10 compares the distribution of $\widehat{\boldsymbol{t}}$ and $\widehat{\boldsymbol{t}}^f$, but neither directly compares the joint distribution of $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{t}})$ and $(\widehat{\boldsymbol{\theta}}^f, \widehat{\boldsymbol{t}}^f)$. To conclude $\widehat{\boldsymbol{\theta}} + \frac{\lambda}{\zeta^*}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{t}}$ "behaves like" $\widehat{\boldsymbol{\theta}}^f + \frac{\lambda}{\zeta^*}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{t}}^f$, we require such a joint comparison. Unfortunately, the approach of [35], does not extend directly to general covariance structures other than $\mathbf{I}_p$. Indeed, the empirical distributions $\hat{\mu} = \frac{1}{p}\sum_{j=1}^p \delta_{\widehat{\theta}_j}$ and $\hat{\mu}' := \frac{1}{p}\sum_{j=1}^p \delta_{\widehat{t}_j}$ do not have a simple characterization for general $\boldsymbol{\Sigma}$.

To conquer this issue, we resort to a smoothing argument. For penalized regression estimators with differentiable penalties, the subgradient $\widehat{\boldsymbol{t}}$ is a function of the estimate $\widehat{\boldsymbol{\theta}}^f$. Indeed, $\widehat{\boldsymbol{t}} \in \partial\|\widehat{\boldsymbol{\theta}}\|_1$ does not identify $\widehat{\boldsymbol{t}}$ from $\widehat{\boldsymbol{\theta}}$ only due to the non-differentiability of the ℓ_1-norm at inactive coordinates. Thus, for smooth procedures, the expression corresponding to Eq. (29) is a deterministic[5] function only of the estimate. Thus, the replacement of the quantities in (29) by their fixed-design counterparts can be justified via analysis of the distribution of the estimate $\widehat{\boldsymbol{\theta}}$ individually. To leverage this simplification under smoothness, we introduce the α-smoothed Lasso, in which we replace the ℓ_1-penalty by a smooth approximation in the original Lasso objective (1). We prove a characterization of the α-smoothed Lasso analogous to Theorem 4, and use this to establish the success of the debiasing procedure corresponding to the smoothed estimator. Finally, we argue that the debiased Lasso estimate is well-approximated by the debiased α-smoothed Lasso estimate for small enough smoothing parameter, allowing us to conclude Theorem 11. The details of this argument are provided in Appendix B.5.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to 'The Lasso with general Gaussian designs with applications to hypothesis testing.'
(doi: COMPLETED BY THE TYPESETTER). The supplement contains proofs and technical details that were omitted from the main text.

REFERENCES

[1] R. Adler and J. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer New York, 2009.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[3] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[4] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[5] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[6] M. Bayati and A. Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[7] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: Improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642, 12 2018.

[5]In particular, it has no dependence on the design $\boldsymbol{X}$ except through $\widehat{\boldsymbol{\theta}}$.

[8] P. C. Bellec and C.-H. Zhang. Second order stein: Sure for sure and other applications in high-dimensional inference. 2018.
[9] P. C. Bellec and C.-H. Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *arXiv:1902.08885*, 2019.
[10] P. C. Bellec and C.-H. Zhang. Second order Poincaré inequalities and de-biasing arbitrary convex regularizers when $p/n \to gamma$. *arXiv:1912.11943*, 2019.
[11] P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
[12] J. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: How sharp is the restricted isometry property? *SIAM Review*, 53, 04 2010.
[13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.
[14] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
[15] E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.
[16] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12, 12 2010.
[17] D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso. *arXiv:1605.02214*, 2016.
[18] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv:1911.05822*, 2019.
[19] D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.
[20] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.
[21] N. El Karoui and E. Purdom. Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235, 2018.
[22] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.
[23] S. A. Geer and S. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
[24] Y. Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.
[25] Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in r^n. In J. Lindenstrauss and V. D. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.
[26] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
[27] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.
[28] A. Javanmard, A. Montanari, et al. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.
[29] S. M. Kakade and S. Shalev-Shwartz. On the duality of strong convexity and strong smoothness : Learning applications and matrix regularization. 2009.
[30] E. Katsevich and A. Ramdas. A theoretical treatment of conditional independence testing under model-x. *arXiv: Statistics Theory*, 2020.
[31] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.
[32] T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv:2002.01586*, 2020.
[33] M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Fast and powerful conditional randomization testing via distillation. *arXiv: Methodology*, 2020.
[34] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.
[35] L. Miolane and A. Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv:1811.01212*, 2018.
[36] A. Montanari and P.-M. Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

[37] A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*, 2020.

[38] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 11 2012.

[39] S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

[40] N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[41] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010.

[42] G. Reeves and H. D. Pfister. The replica-symmetric prediction for compressed sensing with gaussian matrices is exact. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 665–669. IEEE, 2016.

[43] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

[44] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: Extreme singular values. *Proceedings of the International Congress of Mathematicians 2010, ICM 2010*, 03 2010.

[45] M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 08 2018.

[46] M. Stojnic. A framework to characterize performance of LASSO algorithms, 2013.

[47] P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[48] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[49] C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

[50] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[51] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[52] S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[53] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 11 2010.

[54] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[55] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 10 2007.

Supplement to 'The Lasso with general Gaussian designs with applications to hypothesis testing.'

A. Preliminaries.

A.1. *A Gaussian width tradeoff.* It is convenient to define the descent cone

$$L^2 \supset \mathcal{D}(\boldsymbol{x}, \boldsymbol{\Sigma}) := \left\{ \boldsymbol{\Sigma}^{1/2}\boldsymbol{w} \;\Big|\; \mathbb{E}\Big[\sum_{j \in \mathrm{supp}(\boldsymbol{x})} x_j w_j(\boldsymbol{g}) + \|\boldsymbol{w}_{S^c}(\boldsymbol{g})\|_1 \Big] \le 0 \right\}. \tag{30}$$

In this section, we consider relaxations of the constraint $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{x}, \boldsymbol{\Sigma})$ appearing in Eq. (10). In particular, we quantify a tradeoff between the expected correlation $\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2}$ appearing in (10) and the expected growth of the ℓ_1 lower bound appearing in (30). Constraining this trade-off is the central tool in establishing bounds on the solutions to Eqs. (8a) and (8b) (see the proof of Theorem 4(a) in Section A.4).

LEMMA A.1. *Fix $\boldsymbol{\Sigma} \in \mathbb{S}^p_{\ge 0}$ with singular values bounded $0 < \kappa_{\min} \le \kappa_j(\boldsymbol{\Sigma}) \le \kappa_{\max} < \infty$ for all j. Define $\kappa_{\rm cond} := \kappa_{\max}/\kappa_{\min}$. Let $\boldsymbol{x} \in \{-1,0,1\}^p$ with $\|\boldsymbol{x}\|_0/p \ge \nu_{\min} > 0$. Let $S = \mathrm{supp}(\boldsymbol{x})$. Then, for any $\boldsymbol{v} \in L^2$ and any $\epsilon > 0$, we have either*

$$\frac{1}{p}\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2} \le \frac{\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) + \epsilon}{\sqrt{p}} \|\boldsymbol{v}\|_{L^2}, \tag{31}$$

or

$$\frac{1}{p}\mathbb{E}\Big[\sum_{j\in S} x_j w_j(\boldsymbol{g}) + \|\boldsymbol{w}(\boldsymbol{g})_{S^c}\|_1\Big] \ge \frac{\nu_{\min}^{1/2}\epsilon}{\sqrt{p}\kappa_{\max}^{1/2}(2+\kappa_{\rm cond})}\|\boldsymbol{v}(\boldsymbol{g})\|_{L^2}, \tag{32}$$

where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v} \in L^2$.

PROOF OF LEMMA A.1. We may alternatively write (10) as

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) = \sup_{\substack{\boldsymbol{w} \in \mathcal{D}(\boldsymbol{x}, \mathbf{I}_p) \\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_{L^2}^2/p \le 1}} \frac{1}{p}\mathbb{E}\Big[\boldsymbol{w}(\boldsymbol{g})^\top \tilde{\boldsymbol{g}}\Big],$$

where $\tilde{\boldsymbol{g}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}$ and $\boldsymbol{g}$ is interpreted as the identity function in L^2. The Lagrangian for this problem reads:

$$\mathcal{L}(\boldsymbol{w}; \kappa, \xi) := \frac{1}{p}\mathbb{E}\left[\boldsymbol{w}^\top \tilde{\boldsymbol{g}}\right] + \frac{\kappa}{2}\left(1 - \frac{1}{p}\mathbb{E}\left[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2^2\right]\right) - \frac{\xi}{p}\mathbb{E}\Big[\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\Big],$$

where the Langrange multipliers κ, ξ are restricted to be non-negative. First, we bound the dual optimal Lagrange multipliers. We bound

$$\frac{\kappa}{2} + \frac{1}{p}\mathbb{E}\left[\boldsymbol{w}^\top \tilde{\boldsymbol{g}} - \frac{\kappa\kappa_{\min}}{2}\|\boldsymbol{w}\|_2^2 - \xi\Big(\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\Big)\right] \ge \mathcal{L}(\boldsymbol{w}; \kappa, \xi)$$

$$\ge \frac{\kappa}{2} + \frac{1}{p}\mathbb{E}\left[\boldsymbol{w}^\top \tilde{\boldsymbol{g}} - \frac{\kappa\kappa_{\max}}{2}\|\boldsymbol{w}\|_2^2 - \xi\Big(\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\Big)\right]. \tag{33}$$

The expected value appearing in the upper bound is maximized by maximizing the integrand for each value of $\tilde{\boldsymbol{g}}$. Because the integrand is separable across coordinates, we may do this explicitly. The maximal value of the integrand at fixed $\tilde{\boldsymbol{g}}$ is

$$\frac{\kappa}{2}+\frac{1}{2p\kappa\kappa_{\min}}\sum_{j\in S}(\tilde{g}_j-\xi x_j)^2+\frac{1}{p\kappa\kappa_{\min}}\sum_{j\in S^c}\left(\frac{\tilde{g}_j^2}{2}-\xi\mathsf{M}_\xi(\tilde{g}_j)\right),$$

where $\mathsf{M}_\xi(\tilde{g}_j)$ is the Moreau envelope of the ℓ_1-norm

$$\mathsf{M}_\xi(y):=\inf_{x\in\mathbb{R}}\left\{\frac{1}{2\xi}(y-x)^2+|x|\right\}.$$

Because $\xi\mathsf{M}_\xi(\tilde{g}_j)\geq 0$, we have $\mathbb{E}[\tilde{g}_j^2/2-\xi\mathsf{M}_\xi(\tilde{g}_j)]\leq\mathbb{E}[\tilde{g}_j^2/2]\leq\mathbb{E}[(\tilde{g}_j-\xi x)^2/2]\leq(\kappa_{\max}+\xi^2)/2$ whenever $x=\pm1$. Thus,

$$\frac{\kappa}{2}+\frac{1}{\kappa}\frac{\kappa_{\mathrm{cond}}+\xi^2/\kappa_{\min}}{2}\geq\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa,\xi).$$

This further implies that

$$\inf_{\kappa,\xi\geq0}\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa,\xi)\leq\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};1,\kappa_{\min}^{1/2})=1+\frac{\kappa_{\mathrm{cond}}}{2}.$$

Similarly, maximizing the right-hand side of Eq. (33) explicitly and using $\tilde{g}_j^2/2-\xi\mathsf{M}_\xi(\tilde{g}_j)\geq 0$,

$$\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa,\xi)\geq\frac{\kappa}{2}+\frac{1}{\kappa}\frac{|S|(1/\kappa_{\mathrm{cond}}+\xi^2/\kappa_{\max})}{2p}.$$

If either $\kappa/2>1+\kappa_{\mathrm{cond}}/2$ or $\xi^2/\kappa_{\max}>4(1+\kappa_{\mathrm{cond}}/2)^2/(|S|/p)$, then $\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa,\xi)>1+\kappa_{\mathrm{cond}}/2$. Combining the previous two displays, we conclude that $\inf_{\kappa,\xi\geq0}\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa,\xi)$ is achieved at some

$$\kappa^*\leq 2+\kappa_{\mathrm{cond}}\quad\text{and}\quad\xi^*\leq\frac{\kappa_{\max}^{1/2}(2+\kappa_{\mathrm{cond}})}{(|S|/p)^{1/2}}.\tag{34}$$

Since the constraints on $\boldsymbol{w}$ are strictly feasible, strong duality holds:

$$\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa^*,\xi^*)=\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma}).$$

The dual optimal variable ξ^* quantifies the tradeoff we seek to control, as we now show. For any function $\boldsymbol{w}:\mathbb{R}^p\to\mathbb{R}^p$, let $\bar{\boldsymbol{w}}:\mathbb{R}^p\to\mathbb{R}^p$ be defined by $\bar{\boldsymbol{w}}(\boldsymbol{g})=\sqrt{p}\boldsymbol{w}(\boldsymbol{g})/\mathbb{E}[\|\boldsymbol{w}(\boldsymbol{g})\|_{\boldsymbol{\Sigma}}^2]^{1/2}$, where $\|\boldsymbol{w}\|_{\boldsymbol{\Sigma}}^2=\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}$. Then

$$\begin{aligned}&\frac{1}{p}\langle\bar{\boldsymbol{w}},\tilde{\boldsymbol{g}}\rangle_{L^2}-\frac{\xi^*}{p}\mathbb{E}\left[\sum_{j\in S}x_j\bar{w}_j(\boldsymbol{g})+\|\bar{\boldsymbol{w}}_{S^c}(\boldsymbol{g})\|_1\right]\\&\qquad=\frac{1}{p}\mathbb{E}[\bar{\boldsymbol{w}}(\boldsymbol{g})^\top\tilde{\boldsymbol{g}}]+\frac{\kappa^*}{2}\left(1-\frac{1}{p}\mathbb{E}[\|\bar{\boldsymbol{w}}\|_{\boldsymbol{\Sigma}}^2]\right)-\frac{\xi^*}{p}\mathbb{E}\left[\sum_{j\in S}x_j\bar{w}_j(\boldsymbol{g})+\|\bar{\boldsymbol{w}}_{S^c}(\boldsymbol{g})\|_1\right]\\&\qquad\leq\sup_{\boldsymbol{w}\in L^2}\mathcal{L}(\boldsymbol{w};\kappa^*,\xi^*)=\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma}),\end{aligned}$$

where in the first equality we used that $\mathbb{E}[\|\bar{\boldsymbol{w}}(\boldsymbol{g})\|_{\boldsymbol{\Sigma}}^2]/p=1$. We conclude that for any $\epsilon>0$,

$$\text{either}\quad\frac{1}{p}\langle\bar{\boldsymbol{w}},\tilde{\boldsymbol{g}}\rangle_{L^2}\leq\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})+\epsilon\quad\text{or}\quad\frac{1}{p}\mathbb{E}\left[\sum_{j\in S}x_j\bar{w}_j(\tilde{\boldsymbol{g}})+\|\bar{\boldsymbol{w}}(\boldsymbol{g})_{S^c}\|_1\right]\geq\frac{\epsilon}{\xi^*}.$$

Plugging in $\boldsymbol{w} = \mathbb{E}[\|\boldsymbol{w}(\boldsymbol{g})\|_{\boldsymbol{\Sigma}}^2]^{1/2}\bar{\boldsymbol{w}}/\sqrt{p}$ and the upper bound on ξ^* in (34), the lemma follows. $\square$

A.2. *The α-smoothed Lasso.* Controlling the debiased Lasso (Theorem 11) will require a smoothing argument in which we replace the ℓ_1-penalty by a differentiable approximation. In anticipation of this, we study the smoothed and non-smoothed Lasso in a unified way. Results about the Lasso estimate and residual will be instances of these general results.

For $\alpha > 0$, define the Moreau envelope of the ℓ_1-norm

$$\mathsf{M}_\alpha(\boldsymbol{\theta}) := \inf_{\boldsymbol{b}\in\mathbb{R}^p}\left\{\frac{1}{2\alpha}\|\boldsymbol{\theta}-\boldsymbol{b}\|_2^2 + \|\boldsymbol{b}\|_1\right\}, \tag{35}$$

and define $\mathsf{M}_0(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. Notice that this coincides with the Hüber loss. In particular, for all $\boldsymbol{\theta}\in\mathbb{R}^p$,

$$\|\boldsymbol{\theta}\|_1 - \frac{p\alpha}{2} \le \mathsf{M}_\alpha(\boldsymbol{\theta}) \le \|\boldsymbol{\theta}\|_1 . \tag{36}$$

For all $\alpha \ge 0$, define the *α-approximate Lasso* in the random-design model

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_\alpha &:= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}) - \|\boldsymbol{\theta}^*\|_1)\right\}\\ &=: \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\mathcal{R}_\alpha(\boldsymbol{\theta}),\end{aligned}\tag{37}$$

where the term $-\|\boldsymbol{\theta}^*\|_1$ is added to the definition of $\mathcal{R}_\alpha(\boldsymbol{\theta})$ for future convenience. Define the α-approximate Lasso in the fixed-design model

$$\begin{aligned}\eta_\alpha(\boldsymbol{y}^f,\zeta) &:= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\mathsf{M}_\alpha(\boldsymbol{\theta})\right\},\\ \widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta) &:= \boldsymbol{\Sigma}^{1/2}\eta_\alpha(\boldsymbol{y}^f,\zeta).\end{aligned}\tag{38}$$

Denote the in-sample prediction risk and degrees-of-freedom of the α-smoothed Lasso in the fixed-design model by

$$\begin{aligned}\mathsf{R}_\alpha(\tau^2,\zeta) &:= \frac{1}{n}\mathbb{E}\left[\|\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g},\zeta) - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*\|_2^2\right],\\ \mathsf{df}_\alpha(\tau^2,\zeta) &:= \frac{1}{n\tau}\mathbb{E}\left[\langle\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*+\tau\boldsymbol{g},\zeta),\boldsymbol{g}\rangle\right]\\ &= \frac{1}{n}\mathbb{E}[\operatorname{div}\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*+\tau\boldsymbol{g})],\end{aligned}\tag{39}$$

where the expectation is over $\boldsymbol{g}\sim\mathsf{N}(\mathbf{0}_p,\mathbf{I}_p)$. Let $\tau_\alpha^*,\zeta_\alpha^*$ be solutions to the system of equations

$$\tau_\alpha^2 = \sigma^2 + \mathsf{R}_\alpha(\tau_\alpha^2,\zeta_\alpha), \tag{40a}$$

$$\zeta_\alpha = 1 - \mathsf{df}_\alpha(\tau_\alpha^2,\zeta_\alpha). \tag{40b}$$

We refer to these equations as the *α-smoothed fixed point equations*. For $\alpha = 0$, these definitions agree with the corresponding definitions for the Lasso. The solutions τ_α^*, ζ_α^* are well-defined.

LEMMA A.2. *For all $\alpha\ge 0$, if $\boldsymbol{\Sigma}$ is invertible and $\sigma^2 > 0$, then Eqs.* (40a) *and* (40b) *have a unique solution.*

In the following sections, we prove Lemma A.2 and control the behavior of the α-smoothed Lasso using the solutions τ_α^*, ζ_α^* to the α-smoothed fixed point equations.

A.3. *The fixed point equations have a unique solution: proofs of Theorem 1 and Lemma A.2.* Theorem 1 is the $\alpha = 0$ instance of Lemma A.2.

PROOF OF LEMMA A.2. Define functions $\mathcal{T}, \mathcal{Z} : L^2(\mathbb{R}^p; \mathbb{R}^p) \to \mathbb{R}$ by

$$\mathcal{T}(\boldsymbol{v})^2 := \sigma^2 + \frac{1}{n} \|\boldsymbol{v}\|_{L^2}^2 ,$$
$$\mathcal{Z}(\boldsymbol{v}) := \left(1 - \frac{1}{n\mathcal{T}(\boldsymbol{v})} \langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}\right)_+ ,$$

where $\boldsymbol{g}$ is interpreted as the identity function in L^2. Define $\mathscr{E}_\alpha : L^2(\mathbb{R}^p; \mathbb{R}^p) \to \mathbb{R}$ by

(41)
$$\begin{aligned}\mathscr{E}_\alpha(\boldsymbol{v}) &:= \frac{1}{2}\Big(\sqrt{\frac{\|\boldsymbol{v}\|_{L^2}^2}{n} + \sigma^2} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}}{n}\Big)_+^2 + \frac{\lambda}{n}\mathbb{E}\Big\{\big(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})) - \|\boldsymbol{\theta}^*\|_1\big)\Big\} \\ &=: \mathscr{F}(\boldsymbol{v}) + \frac{\lambda}{n}\mathbb{E}\Big\{\big(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})) - \|\boldsymbol{\theta}^*\|_1\big)\Big\} .\end{aligned}$$

Let us emphasize the argument of $\mathscr{E}_\alpha$ is not a vector but a function $\boldsymbol{v} : \mathbb{R}^p \to \mathbb{R}^p$.

Each of the two terms in the definition of $\mathscr{E}_\alpha$ are convex and continuous. Moreover, for all $\boldsymbol{g}$ we have, by Eq. (36),

$$\begin{aligned}\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})) &\geq \|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})\|_1 - \frac{p\alpha}{2} \\ &\geq \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})\|_1 - \|\boldsymbol{\theta}^*\|_1 - \frac{p\alpha}{2} \geq \kappa_{\max}^{-1/2}\|\boldsymbol{v}(\boldsymbol{g})\|_2 - \|\boldsymbol{\theta}^*\|_1 - \frac{p\alpha}{2} .\end{aligned}$$

For any $M > 0$,

$$\begin{aligned}|\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2}| &= |\mathbb{E}[\langle \boldsymbol{v}(\boldsymbol{g}), \boldsymbol{g}\mathbf{1}_{\|\boldsymbol{g}\|_2 > M}\rangle] + \mathbb{E}[\langle \boldsymbol{v}(\boldsymbol{g}), \boldsymbol{g}\mathbf{1}_{\|\boldsymbol{g}\|_2 \leq M}\rangle]| \\ &\leq \|\boldsymbol{v}\|_{L^2}\mathbb{E}[\|\boldsymbol{g}\|_2^2\mathbf{1}_{\|\boldsymbol{g}\|_2 > M}]^{1/2} + M\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{g})\|_2].\end{aligned}$$

Take M large enough that $\mathbb{E}[\|\boldsymbol{g}\|^2\mathbf{1}_{\|\boldsymbol{g}\| > M}] < n/2$. Then

$$\begin{aligned}\mathscr{E}_\alpha(\boldsymbol{v}) &\geq \frac{1}{2}\left(\frac{1}{2}\frac{\|\boldsymbol{v}\|_{L^2}}{\sqrt{n}} - \frac{M}{n}\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{g})\|_2]\right)_+^2 + \frac{\lambda\kappa_{\max}^{-1/2}}{n}\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{g})\|_2] - \frac{2\lambda}{n}\|\boldsymbol{\theta}^*\|_1 - \frac{\alpha}{2\delta} \\ &\geq \min\left\{\frac{1}{32n}\|\boldsymbol{v}\|_{L^2}^2, \frac{\lambda\kappa_{\max}^{-1/2}}{4Mn^{1/2}}\|\boldsymbol{v}\|_{L^2}\right\} - \frac{2\lambda}{n}\|\boldsymbol{\theta}^*\|_1 - \frac{\alpha}{2\delta} ,\end{aligned}$$

where the second inequality holds by considering the cases that $\|\boldsymbol{v}\|_{L^2}/(4\sqrt{n})$ is no smaller and no larger than $M\mathbb{E}[\|\boldsymbol{v}(\boldsymbol{g})\|_2]/n$, respectively. We see that $\mathscr{E}_\alpha(\boldsymbol{v}) \to \infty$ as $\|\boldsymbol{v}\|_{L^2} \to \infty$, whence by [4, Theorem 11.9] $\mathscr{E}_\alpha$ has a minimizer. Let $\boldsymbol{v}_\alpha^*$ be one such minimizer.

Consider the following convex function in L^2 parameterized by $\tau, \zeta \geq 0$:

$$\begin{aligned}\tilde{\mathscr{E}}_\alpha(\boldsymbol{v}; \zeta, \tau) &:= \frac{\zeta}{2n}\big\|\boldsymbol{v} - \tau\boldsymbol{g}\big\|_{L^2}^2 + \frac{\lambda}{n}\mathbb{E}\Big\{\big(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})) - \|\boldsymbol{\theta}^*\|_1\big)\Big\} \\ &= \mathbb{E}\left\{\frac{\zeta}{2n}\|\boldsymbol{v}(\boldsymbol{g}) - \tau\boldsymbol{g}\|_2^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})) - \|\boldsymbol{\theta}^*\|_1)\right\} .\end{aligned}$$

For fixed $\zeta, \tau \geq 0$, the function $\boldsymbol{v}_\alpha^*$ minimizes $\tilde{\mathscr{E}}_\alpha$ if and only if $\boldsymbol{v}_\alpha^*(\boldsymbol{g})$ minimizes the objective inside the expectation for almost every $\boldsymbol{g}$. That is, if and only if

(42)
$$\boldsymbol{v}_\alpha^* = \boldsymbol{\Sigma}^{1/2}(\eta_\alpha(\boldsymbol{\theta}^* + \tau\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}; \zeta) - \boldsymbol{\theta}^*) \text{ almost surely.}$$

For any $\boldsymbol{v}_0, \boldsymbol{v}_1 \in L^2$ fixed, we have by differentiation of $\mathscr{F}$ with respect to $\varepsilon \in \mathbb{R}$ that

$$\begin{aligned}\tilde{\mathscr{E}}_\alpha(\boldsymbol{v}_0+\varepsilon\boldsymbol{v}_1;\mathcal{Z}(\boldsymbol{v}_0),\mathcal{T}(\boldsymbol{v}_0))-\mathscr{E}_\alpha(\boldsymbol{v}_0+\varepsilon\boldsymbol{v}_1) &= \frac{\mathcal{Z}(\boldsymbol{v}_0)}{2n}\|\boldsymbol{v}_0+\varepsilon\boldsymbol{v}_1-\mathcal{T}(\boldsymbol{v}_0)\boldsymbol{g}\|_{L^2}^2-\mathscr{F}(\boldsymbol{v}_0+\varepsilon\boldsymbol{v}_1)\\ &= \tilde{\mathscr{E}}_\alpha(\boldsymbol{v}_0;\mathcal{Z}(\boldsymbol{v}_0);\mathcal{T}(\boldsymbol{v}_0))-\mathscr{E}_\alpha(\boldsymbol{v}_0)+O(\epsilon^2)\,.\end{aligned}$$

Thus, $\boldsymbol{v}_1$ is a descent direction of $\boldsymbol{v}\mapsto\tilde{\mathscr{E}}_\alpha(\boldsymbol{v};\mathcal{Z}(\boldsymbol{v}_0),\mathcal{T}(\boldsymbol{v}_0))$ at $\boldsymbol{v}_0$ if and only if it is also a descent direction of $\boldsymbol{v}\mapsto\mathscr{E}_\alpha(\boldsymbol{v})$ at $\boldsymbol{v}_0$. In particular, $\boldsymbol{v}_0$ minimizes $\mathscr{E}_\alpha$ if and only if it minimizes $\tilde{\mathscr{E}}(\boldsymbol{v};\zeta,\tau)$ for $\zeta=\mathcal{Z}(\boldsymbol{v}_0)$ and $\tau=\mathcal{T}(\boldsymbol{v}_0)$. By (42), we conclude that $\boldsymbol{v}_\alpha^*$ is a minimizer of $\mathscr{E}_\alpha$ if and only if

$$\boldsymbol{v}_\alpha^*(\boldsymbol{g})=\boldsymbol{\Sigma}^{1/2}(\eta_\alpha(\boldsymbol{\theta}^*+\mathcal{T}(\boldsymbol{v}_\alpha^*)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g};\mathcal{Z}(\boldsymbol{v}_\alpha^*))-\boldsymbol{\theta}^*)\ \text{almost surely.}\tag{43}$$

That is, if and only if $\tau_\alpha^*=\mathcal{T}(\boldsymbol{v}_\alpha^*)$, $\zeta_\alpha^*=\mathcal{Z}(\boldsymbol{v}_\alpha^*)$ is a solution to equations (40a) and (40b). Because $\mathscr{E}_\alpha$ has minimizers, solutions to equations (40a) and (40b) exist.

To complete the proof, we only need to show that the minimizer $\boldsymbol{v}_\alpha^*$ of $\mathscr{E}_\alpha$ is unique. First, we claim $\mathcal{Z}(\boldsymbol{v}_\alpha^*)>0$ for all minimizers $\boldsymbol{v}_\alpha^*$. Assume otherwise that $\mathcal{Z}(\boldsymbol{v}_\alpha^*)=0$ for some minimizer $\boldsymbol{v}_\alpha^*$. Then, by property (43),

$$\boldsymbol{v}_\alpha^*=\boldsymbol{\Sigma}^{1/2}(\eta_\alpha(\boldsymbol{\theta}^*+\mathcal{T}(\boldsymbol{v}_\alpha^*)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g};0)-\boldsymbol{\theta}^*)=-\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*.$$

Thus, we have $\mathcal{Z}(\boldsymbol{v}_\alpha^*)=\Big(1-\frac{1}{n\mathcal{T}(\boldsymbol{v}_\alpha^*)}\langle\boldsymbol{g},-\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*\rangle_{L^2}\Big)_+=1$, a contradiction. We conclude $\mathcal{Z}(\boldsymbol{v}_\alpha^*)>0$ for all minimizers $\boldsymbol{v}_\alpha^*$ of $\mathscr{E}_\alpha$.

The function $\mathscr{E}_\alpha$ is strictly convex on $\mathcal{Z}(\boldsymbol{v})>0$. Indeed, for any $\boldsymbol{v}\neq\boldsymbol{v}'$, the function

$$t\mapsto\sqrt{\frac{\|(1-t)\boldsymbol{v}+t\boldsymbol{v}'\|_{L^2}^2}{n}+\sigma^2}=\sqrt{\frac{\|\boldsymbol{v}\|_{L^2}^2-2t\langle\boldsymbol{v},\boldsymbol{v}-\boldsymbol{v}'\rangle_{L^2}+t^2\|\boldsymbol{v}'\|_{L^2}^2}{n}+\sigma^2}$$

is strictly convex by univariate calculus. Because $x\mapsto x_+^2$ is convex and strictly increasing on $x>0$, strict convexity of $\mathscr{E}_\alpha$ on $\mathcal{Z}(\boldsymbol{v})>0$ follows. Because all minimizers $\boldsymbol{v}_\alpha^*$ satisfy $\mathcal{Z}(\boldsymbol{v}_\alpha^*)>0$, strict convexity on $\boldsymbol{v}_\alpha^*>0$ implies the minimizer is unique. □

A.4. *Uniform bounds on fixed point solutions: proof of Theorem 2.* In the context of the α-smoothed Lasso, we replace assumption A2 with the following assumption A2$_\alpha$. As before, our results below hold uniformly over families of instances $(\boldsymbol{\theta}^*,\boldsymbol{\Sigma},\sigma,\lambda)$ that satisfy such condition.

A2$_\alpha$ There exist $0<\tau_{\min}\le\tau_{\max}<\infty$ and $0<\zeta_{\min}\le\zeta_{\max}<\infty$ such that the unique solution τ_α^*, ζ_α^* to the fixed point equations (40a) and (40b) are bounded: $\tau_{\min}\le\tau_\alpha^*\le\tau_{\max}$ and $\zeta_{\min}\le\zeta_\alpha^*\le\zeta_{\max}$.

Theorem 2 is the $\alpha_{\max}=0$ instance of the following lemma.

LEMMA A.3. *Consider $\alpha_{\max}\ge0$. Under assumption A1 and if $\boldsymbol{\theta}^*$ is $(s,\sqrt{\delta}(1-\Delta_{\min}),M)$-approximately sparse for some $s/p\ge\nu_{\min}>0$ and $1\ge\Delta_{\min}>0$ and if $\alpha\le\alpha_{\max}$, then there exist $0<\tau_{\min}\le\tau_{\max}<\infty$ and $0<\zeta_{\min}\le\zeta_{\max}<\infty$ depending only on $\mathcal{P}_{\mathrm{model}}$, δ, $\nu_{\min}$, $\Delta_{\min}$, and $\alpha_{\max}$ such that the unique solution τ_α^*, ζ_α^* to Eqs. (40a) and (40b) satisfies $\tau_{\min}\le\tau_\alpha^*\le\tau_{\max}$ and $\zeta_{\min}\le\zeta_\alpha^*\le\zeta_{\max}$.*

The proof of Lemma A.3 relies on controlling the degrees of freedom of the α-smoothed Lasso in terms of its prediction risk in the fixed-design model. We establish this control in the next lemma.

LEMMA A.4. *For any $\tau, \zeta, \delta > 0$ and $\alpha \geq 0$ and if the eigenvalues of $\mathbf{\Sigma}$ are bounded as $0 < \kappa_{\min} \leq \kappa_j(\mathbf{\Sigma}) \leq \kappa_{\max} < \infty$, then*

$$\frac{\kappa_{\max}^{1/2}}{\delta^{1/2}}\left(\tau\kappa_{\text{cond}}^{1/2} + \sqrt{\tau^2\kappa_{\text{cond}} + \delta \mathsf{R}_\alpha(\tau^2,\zeta)}\right) \geq \frac{\lambda}{\zeta}\mathsf{df}_\alpha(\tau^2,\zeta) - \frac{\alpha\kappa_{\max}}{\delta}\,.$$

We prove Lemma A.4 at the end of this section.

PROOF OF LEMMA A.3. By general properties of proximal operators [4], the Jacobian matrix $\boldsymbol{D}\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta)$ of $\widehat{\boldsymbol{y}}_\alpha(\,\cdot\,,\zeta)$ is positive-semidefinite. Therefore $\mathsf{df}_\alpha(\tau_\alpha^{*2},\zeta_\alpha^*) \geq 0$ and $\zeta_\alpha^* \leq 1$ is immediate from Eq. (40b). Further, $\tau_\alpha^* \geq \sigma_{\min}$ is immediate from Eq. (40a). We may take

$$\tau_{\min} = \sigma_{\min} \;\text{ and }\; \zeta_{\max} = 1\,.$$

Establishing the bound $\tau_{\max}^2$. Because $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse, there exists $\boldsymbol{x} \in \{-1,0,1\}^p$ such that $\|\boldsymbol{x}\|_0 = s$ and $\mathcal{G}(\boldsymbol{x},\mathbf{\Sigma}) \leq \sqrt{\delta}(1-\Delta_{\min})$, and $\bar{\boldsymbol{\theta}}^* \in \mathbb{R}^p$ such that $\frac{1}{p}\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1 \leq M$ and $\boldsymbol{x} \in \partial\|\bar{\boldsymbol{\theta}}^*\|_1$. Denote $S = \operatorname{supp}(\boldsymbol{x}) \subset [p]$.

We may equivalently write the objective in (41) as a function of $\boldsymbol{w}(\tilde{\boldsymbol{g}}) := \mathbf{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})$ where $\tilde{\boldsymbol{g}} := \mathbf{\Sigma}^{1/2}\boldsymbol{g}$. Note that

$$\begin{aligned}\frac{1}{n}\big(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{w}(\tilde{\boldsymbol{g}})) - \|\boldsymbol{\theta}^*\|_1\big) &\geq \frac{1}{n}\left(\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{w}(\tilde{\boldsymbol{g}})\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1\right) - \frac{2}{n}\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1 - \frac{\alpha_{\max}}{2\delta}\\ &\geq \frac{1}{\delta p}\left(\sum_{j\in S} x_j w(\boldsymbol{g})_j + \|\boldsymbol{w}(\boldsymbol{g})_{S^c}\|_1\right) - \frac{2M}{\delta} - \frac{\alpha_{\max}}{2\delta}\,,\end{aligned} \tag{44}$$

where the first inequality uses the relation (36). Plugging in $\epsilon = \frac{\sqrt{\delta}-\mathcal{G}(\boldsymbol{x},\mathbf{\Sigma})}{2}$ in Lemma A.1, Eqs. (31) and (32), we have that either

$$\frac{1}{p}\mathbb{E}[\langle \boldsymbol{w}(\tilde{\boldsymbol{g}}), \tilde{\boldsymbol{g}}\rangle] \leq \frac{\sqrt{\delta} + \mathcal{G}(\boldsymbol{x},\mathbf{\Sigma})}{2\sqrt{p}}\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\mathbf{\Sigma}}^2]^{1/2}\,,$$

or

$$\frac{1}{p}\mathbb{E}\left[\sum_{j\in S} x_j w_j(\tilde{\boldsymbol{g}}) + \|\boldsymbol{w}(\tilde{\boldsymbol{g}})_{S^c}\|_1\right] \geq \frac{\sqrt{\delta} - \mathcal{G}(\boldsymbol{x},\mathbf{\Sigma})}{2\xi^*\sqrt{p}}\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\mathbf{\Sigma}}^2]^{1/2}\,,$$

where $\xi^* = \kappa_{\max}^{1/2}(2+\kappa_{\text{cond}})/\nu_{\min}^{1/2}$. Then, Eq. (44) gives

$$\begin{aligned}&\mathscr{E}_\alpha(\boldsymbol{v}) = \frac{1}{2}\left(\sqrt{\frac{\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\mathbf{\Sigma}}^2]}{n} + \sigma^2} - \frac{\mathbb{E}[\langle \tilde{\boldsymbol{g}}, \boldsymbol{w}(\tilde{\boldsymbol{g}})\rangle]}{\delta p}\right)_+^2 + \frac{\lambda}{n}\mathbb{E}\left\{\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{w}(\tilde{\boldsymbol{g}})) - \|\boldsymbol{\theta}^*\|_1\right\}\\ &\geq \min\left\{\frac{1}{2}\left(\frac{1-\mathcal{G}(\boldsymbol{x},\mathbf{\Sigma})/\sqrt{\delta}}{2}\right)_+^2\frac{\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\mathbf{\Sigma}}^2]}{n}, \frac{\lambda(1-\mathcal{G}(\boldsymbol{x},\mathbf{\Sigma})/\sqrt{\delta})\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\mathbf{\Sigma}}^2]^{1/2}}{2\xi^*\sqrt{n}}\right\}\\ &\qquad - \frac{2\lambda M}{\delta} - \frac{\alpha_{\max}\lambda}{2\delta}\,.\end{aligned}$$

As in the proof of Lemma A.2 (see Eq. (42)), let $\boldsymbol{v}_\alpha^*$ be the minimizer of $\mathscr{E}_\alpha$. Because $\mathsf{M}_\alpha(\boldsymbol{\theta}^*) \leq \|\boldsymbol{\theta}^*\|_1$, we bound $\sigma^2/2 = \mathscr{E}_0(\mathbf{0}) \geq \mathscr{E}_\alpha(\mathbf{0}) \geq \mathscr{E}_\alpha(\boldsymbol{v}_\alpha^*)$. Combining this bound with the previous display applied at $\boldsymbol{v}(\tilde{\boldsymbol{g}}) = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}_\alpha^*$, some algebra yields

$$\frac{\|\boldsymbol{v}_\alpha^*\|_{L^2}^2}{n} = \frac{\mathbb{E}[\|\boldsymbol{w}(\tilde{\boldsymbol{g}})\|_{\boldsymbol{\Sigma}}^2]}{n}$$

$$\le \max\left\{\frac{8(\sigma^2/2 + 2\lambda_{\max} M/\delta + \alpha_{\max}\lambda_{\max}/(2\delta))}{(1-\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})/\sqrt{\delta})^2}, \frac{(\sigma^2/\lambda_{\min} + 4M/\delta + \alpha_{\max}/\delta)^2\xi^{*2}}{(1-\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})/\sqrt{\delta})^2}\right\}$$

$$\le \max\left\{\frac{8(\sigma_{\max}^2/2 + 2\lambda_{\max} M/\delta + \alpha_{\max}/(2\delta))}{\Delta_{\min}^2}, \frac{(\sigma_{\max}^2/\lambda_{\min} + 4M/\delta + \alpha_{\max}/\delta)^2\kappa_{\max}(2+\kappa_{\text{cond}})^2}{\Delta_{\min}^2\nu_{\min}}\right\}.$$

Recalling the fixed point equation (40a), we may set $\tau_{\max}^2$ to be the sum of σ^2 and the right-hand side above.

Establishing the bound $\zeta_{\min}$. If $\mathsf{df}_\alpha(\tau_\alpha^{*2}, \zeta_\alpha^*) \le 1/2$, then by Eq. (39), $\zeta_\alpha^* \ge 1/2$. Alternatively, if $\mathsf{df}_\alpha({\tau^*}_\alpha^2, \zeta_\alpha^*) \ge 1/2$, then by Lemma A.4, it is guaranteed that

$$\frac{\kappa_{\max}^{1/2}\tau_{\max}}{\delta^{1/2}}\left(\kappa_{\text{cond}}^{1/2} + \sqrt{\kappa_{\text{cond}} + \delta}\right) \ge \frac{\lambda_{\min}}{2\zeta_\alpha^*} - \frac{\alpha_{\max}\kappa_{\max}}{\delta},$$

where we have used that $\tau_\alpha^* \le \tau_{\max}$ (as established above) and by (40b) that $\mathsf{R}_\alpha(\tau_\alpha^*, \zeta_\alpha^*) \le \tau_\alpha^{*2}$. Rearranging terms, we conclude

$$\zeta_\alpha^* \ge \frac{\lambda_{\min}\delta}{2(\kappa_{\max}^{1/2}\tau_{\max}\delta^{1/2}(\kappa_{\text{cond}}^{1/2} + \sqrt{\kappa_{\text{cond}}+\delta}) + \alpha_{\max}\kappa_{\max})}.$$

Thus, we may set

$$\zeta_{\min} = \frac{1}{2}\min\left\{1, \frac{\lambda_{\min}\delta}{\kappa_{\max}^{1/2}\tau_{\max}\delta^{1/2}(\kappa_{\text{cond}}^{1/2} + \sqrt{\kappa_{\text{cond}}+\delta}) + \alpha_{\max}\kappa_{\max}}\right\}.$$

The proof is complete. □

PROOF OF LEMMA A.4. The KKT conditions for the α-smoothed Lasso in the fixed-design model (38) are

$$\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f, \zeta) - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* = \tau\boldsymbol{g} - \frac{\lambda}{\zeta}\boldsymbol{\Sigma}^{-1/2}\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f, \zeta)). \tag{45}$$

where $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}$. Therefore,

$$\frac{1}{n}\|\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta) - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*\|_2^2 \ge \frac{\lambda^2}{\zeta^2\kappa_{\max}}\frac{\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta)\|_2^2}{n} - \frac{2\lambda\tau}{\zeta\kappa_{\min}^{1/2}}\frac{\|\boldsymbol{g}\|_2\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\|_2}{n}.$$

Taking expectations and applying Cauchy-Schwartz yields

$$\mathsf{R}_\alpha(\tau^2,\zeta) \ge \frac{\lambda^2}{\zeta^2\kappa_{\max}}\frac{\mathbb{E}[\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta)\|_2^2]}{n} - \frac{2\lambda\tau}{\zeta\delta^{1/2}\kappa_{\min}^{1/2}}\frac{\mathbb{E}[\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta)\|_2^2]^{1/2}}{\sqrt{n}},$$

Solving the resulting quadratic equation for $\frac{\lambda\mathbb{E}[\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta)\|_2^2]^{1/2}}{\zeta\sqrt{n}}$, we conclude

$$\frac{\lambda}{\zeta}\frac{\mathbb{E}[\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta)\|_2^2]^{1/2}}{\sqrt{n}} \le \frac{\kappa_{\max}^{1/2}}{\delta^{1/2}}\left(\tau\kappa_{\text{cond}}^{1/2} + \sqrt{\tau^2\kappa_{\text{cond}} + \delta\mathsf{R}_\alpha(\tau^2,\zeta)}\right). \tag{46}$$

We compute

$$\nabla \mathsf{M}_\alpha(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \eta_{\rm soft}(\boldsymbol{\theta},\alpha))/\alpha\,, \qquad \nabla^2 \mathsf{M}_\alpha(\boldsymbol{\theta}) = \mathrm{diag}((\mathbf{1}_{|\theta_j|\le\alpha})_j)/\alpha\,.$$

Because $\theta - \eta_{\rm soft}(\theta,\alpha))/\alpha = 1$ for $|\theta| \ge \alpha$, we bound

$$\|\nabla \mathsf{M}_\alpha(\boldsymbol{\theta})\|_2^2 \ge |\{j\in[p] \mid |\theta_j| \ge \alpha\}|\,.$$

The KKT condition (45) gives

$$\zeta \boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta)) = \lambda \nabla \mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\,.$$

Differentiating with respect to $\boldsymbol{y}^f$,

$$\zeta\boldsymbol{\Sigma}^{1/2} - \zeta\boldsymbol{\Sigma}^{1/2}\nabla\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f;\zeta) = \lambda\nabla^2\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\boldsymbol{\Sigma}^{-1/2}\nabla\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f;\zeta)\,.$$

(More precisely, $\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta)$ and $\eta_\alpha(\boldsymbol{y}^f,\zeta)$ are continuous and piecewise linear in $\boldsymbol{y}^f$, and the above identity holds in the interior of each linear region.) We therefore get

$$\begin{aligned}\nabla\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta) &= \left(\mathbf{I}_p + \frac{\lambda}{\zeta}\boldsymbol{\Sigma}^{-1/2}\nabla^2\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\boldsymbol{\Sigma}^{-1/2}\right)^{-1} = \left(\mathbf{I}_p + \frac{\lambda}{\alpha\zeta}(\boldsymbol{\Sigma}^{-1/2})_{\cdot,S^c}(\boldsymbol{\Sigma}^{-1/2})_{S^c,\cdot}\right)^{-1}\\ &= \mathbf{I}_p - \frac{\lambda}{\alpha\zeta}(\boldsymbol{\Sigma}^{-1/2})_{\cdot,S^c}\left(\mathbf{I}_{|S^c|} + \frac{\lambda}{\alpha\zeta}(\boldsymbol{\Sigma}^{-1/2})_{S^c,\cdot}(\boldsymbol{\Sigma}^{-1/2})_{\cdot,S^c}\right)^{-1}(\boldsymbol{\Sigma}^{-1/2})_{S^c,\cdot}\,,\end{aligned}$$

where $S = \{j\in[p] \mid |\eta_\alpha(\boldsymbol{y}^f,\zeta)| \ge \alpha\}$. Thus,

$$\mathrm{div}\,\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta) = \mathrm{trace}(\nabla\widehat{\boldsymbol{y}}_\alpha(\boldsymbol{y}^f,\zeta)) \le p - \frac{|S^c|}{1+\alpha\zeta\kappa_{\max}/\lambda} \le p - \frac{p - \|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\|_2^2}{1+\alpha\zeta\kappa_{\max}/\lambda}\,.$$

Rearranging and taking expectations,

$$\frac{\mathbb{E}[\|\nabla\mathsf{M}_\alpha(\eta_\alpha(\boldsymbol{y}^f,\zeta))\|_2^2]}{n} \ge \left(1 + \frac{\alpha\zeta\kappa_{\max}}{\lambda}\right)\mathsf{df}_\alpha(\tau^2,\zeta) - \frac{\alpha\zeta\kappa_{\max}}{\lambda\delta}\,.$$

Combining with Eq. (46),

$$\frac{\kappa_{\max}^{1/2}}{\delta^{1/2}}\left(\tau\kappa_{\rm cond}^{1/2} + \sqrt{\tau^2\kappa_{\rm cond} + \delta\mathsf{R}_\alpha(\tau^2,\zeta)}\right) \ge \left(\frac{\lambda}{\zeta} + \alpha\kappa_{\max}\right)\mathsf{df}_\alpha(\tau^2,\zeta) - \frac{\alpha\kappa_{\max}}{\delta}\,.$$

We may ignore the non-negative term $\alpha\kappa_{\max}$ in parentheses. The proof is complete. □

As a consequence, one arrives at the following result.

COROLLARY 14. *Under assumptions Lemma A.3, $\mathsf{df}_\alpha(\tau_\alpha^{*2},\zeta_\alpha^*)$ is uniformly bounded away from one. Namely*

$$\mathsf{df}_\alpha(\tau_\alpha^{*2},\zeta_\alpha^*) = 1 - \zeta^* \le 1 - \zeta_{\min}\,.$$

with $\zeta_{\min}$ depending uniquely on the constants $\mathcal{P}_{\rm model}$, δ, $\nu_{\min}$, $\Delta_{\min}$, and $\alpha_{\max}$.

A.5. *Continuity of fixed point solutions in smoothing parameter.*

LEMMA A.5. *If assumptions A1 and A2 hold, then there exist constants $\alpha_{\max}$, L_τ, and L_ζ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that for $\alpha \le \alpha_{\max}$,*

$$|\tau_0^* - \tau_\alpha^*| \le L_\tau\sqrt{\alpha}\,, \qquad |\zeta_0^* - \zeta_\alpha^*| \le L_\zeta\sqrt{\alpha}\,.$$

We emphasize that the assumptions in Lemma A.3 are made about the Lasso fixed point parameters rather than the α-smoothed Lasso fixed point parameters.

PROOF OF LEMMA A.5. The function

$$f : L^2 \to \mathbb{R}\,, \quad \boldsymbol{v} \mapsto \sqrt{\frac{\|\boldsymbol{v}\|_{L^2}^2}{n} + \sigma^2} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}}{n}\,,$$

is $(1/\sqrt{n} + 1/\sqrt{\delta n})$-Lipschitz. Evaluated at the minimizer $\boldsymbol{v}_0^*$ of $\mathscr{E}_0$ defined in (41), $f(\boldsymbol{v}_0^*)$ is equal to $\tau_0^*\zeta_0^* \geq \tau_{\min}\zeta_{\min}$ by the proof of Lemma A.2 in Section A.3. Thus, for $\|\boldsymbol{v} - \boldsymbol{v}_0^*\|_{L^2}/\sqrt{n} \leq \tau_{\min}\zeta_{\min}/(2(1+\delta^{-1/2}))$, it is guaranteed that

$$f(\boldsymbol{v}) \geq \frac{\tau_{\min}\zeta_{\min}}{2}\,.$$

Let $r := \min\left\{\frac{\tau_{\min}\zeta_{\min}}{2(1+\delta^{-1/2})}, \frac{\tau_{\min}}{2}\right\}$. By differentiation along affine paths, the function

$$\frac{1}{2}f(\boldsymbol{v})_+^2 \quad \text{is} \quad \frac{\sigma_{\min}^2 \inf_{\boldsymbol{v}\in B} f(\boldsymbol{v})_+}{n(\sup_{\boldsymbol{v}\in B}\|\boldsymbol{v}\|_{L^2}^2/n + \sigma_{\min}^2)^{3/2}} \quad \text{strongly convex on } \boldsymbol{v} \in B \text{ for any } B \subset L^2.$$

Thus, $\mathscr{E}_0$ is $\frac{\sigma^2\tau_{\min}\zeta_{\min}/2}{n(R^2+\sigma^2)^{3/2}}$-strongly convex on $\|\boldsymbol{v} - \boldsymbol{v}_0^*\|_{L^2}/\sqrt{n} \leq r$, where $R = \tau_{\max} + r$. Denote this strong convexity parameter by a/n.

By Eq. (36), for any $\boldsymbol{v} \in L^2$ and $\alpha \geq 0$, $\mathbb{E}[\mathsf{M}_0(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g}))] \geq \mathbb{E}[\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g}))] \geq \mathbb{E}[\mathsf{M}_0(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g}))] - p\alpha/2$. Thus, $\mathscr{E}_\alpha(\boldsymbol{v}_0^*) \leq \mathscr{E}_0(\boldsymbol{v}_0^*)$ and for $\|\boldsymbol{v} - \boldsymbol{v}_0^*\|_{L^2}/\sqrt{n} \leq r$, $\mathscr{E}_\alpha(\boldsymbol{v}) \geq \mathscr{E}_0(\boldsymbol{v}) - \lambda\alpha/(2\delta) \geq \mathscr{E}_0(\boldsymbol{v}_0^*) + a\|\boldsymbol{v} - \boldsymbol{v}_0^*\|_{L^2}^2/(2n) - \lambda\alpha/(2\delta)$. Thus, for $\sqrt{\frac{\lambda_{\max}\alpha}{a\delta}} \leq r$, we have $\frac{\|\boldsymbol{v}_\alpha^* - \boldsymbol{v}_0^*\|_{L^2}}{\sqrt{n}} \leq \sqrt{\frac{\lambda_{\max}\alpha}{a\delta}}$. Since, by the proof of Lemma A.2, $\tau_\alpha^* = \sqrt{\sigma^2 + \|\boldsymbol{v}_\alpha^*\|_{L^2}^2/n}$ and $\zeta_\alpha^* = (1 - \langle \boldsymbol{g}, \boldsymbol{v}_\alpha^*\rangle_{L^2}/n)$, we conclude

$$|\tau_0^* - \tau_\alpha^*| \leq \sqrt{\frac{\lambda_{\max}\alpha}{a\delta}}\,, \quad |\zeta_0^* - \zeta_\alpha^*| \leq \frac{1}{\delta^{1/2}}\sqrt{\frac{\lambda_{\max}\alpha}{a\delta}} \quad \text{for} \quad \alpha \leq \frac{r^2 a\delta}{\lambda_{\max}}\,.$$

The proof is complete. □

A.6. *The fixed point solutions as a saddle point.* A crucial role in our analysis is played by the max-min problem

$$\max_{\beta>0}\min_{\tau\geq\sigma}\psi_\alpha(\tau,\beta)\,, \tag{47}$$

$$\psi_\alpha(\tau,\beta) := -\frac{1}{2}\beta^2 - \frac{1-\delta}{2\delta}\tau\beta + \frac{\sigma^2\beta}{2\tau} + \frac{1}{n}\mathbb{E}\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\beta}{2\tau}\|\boldsymbol{\theta} - \boldsymbol{\theta}^* - \tau\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}\|_{\boldsymbol{\Sigma}}^2 + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}) - \|\boldsymbol{\theta}^*\|_1)\right\}\,,$$

where the expectation is taken over $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_p)$. We establish that Eqs. (40a) and (40b) are first-order conditions for the solution to this max-min problem, and in the non-smoothed ($\alpha = 0$) case, Eqs. (8a) and (8b) are first-order conditions for the solution to this max-min problem.

LEMMA A.6. *Let τ_α^*, ζ_α^* be the unique solution to Eqs.* (40a) *and* (40b)*, and let $\beta_\alpha^* = \tau_\alpha^*\zeta_\alpha^*$. Then $(\tau_\alpha^*, \beta_\alpha^*)$ is a saddle point for the max-min value in Eq.* (47)*. Namely, for all $\beta > 0$, $\tau \geq \sigma$,*

$$\psi_\alpha(\tau_\alpha^*, \beta) \leq \psi_\alpha(\tau_\alpha^*, \beta_\alpha^*) \leq \psi_\alpha(\tau, \beta_\alpha^*)\,, \tag{48}$$

$$\psi_\alpha(\tau_\alpha^*, \beta_\alpha^*) = \max_{\beta>0}\min_{\tau\geq\sigma}\psi_\alpha(\tau,\beta) = \min_{\tau\geq\sigma}\max_{\beta>0}\psi_\alpha(\tau,\beta)\,. \tag{49}$$

PROOF OF LEMMA A.6. Let us define function

$$\Xi_\alpha(\tau,\beta) := -\frac{1}{2}\beta^2 - \frac{1-\delta}{2\delta}\tau\beta + \frac{\sigma^2\beta}{2\tau} + \frac{1}{n}\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\beta}{2\tau}\|\boldsymbol{\theta}-\boldsymbol{\theta}^* - \tau\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}\|_{\boldsymbol{\Sigma}}^2 + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}) - \|\boldsymbol{\theta}^*\|_1)\right\},$$

so that $\psi_\alpha(\tau,\beta) = \mathbb{E}_{\boldsymbol{g}}\Xi_\alpha(\tau,\beta)$. It is easily seen that Ξ_α is convex-concave in (τ,β) for $\tau,\beta > 0$ because prior to the minimization over $\boldsymbol{\theta}$ it is jointly convex in $(\tau,\boldsymbol{\theta})$ and concave in β. By the envelope theorem [34, Theorem 1],

$$\begin{aligned}
\frac{\partial \Xi_\alpha}{\partial\beta} &= -\beta - \frac{1-\delta}{2\delta}\tau + \frac{\sigma^2}{2\tau} + \frac{1}{2\tau n}\|\eta_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}, \beta/\tau) - \boldsymbol{\theta}^* - \tau\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}\|_{\boldsymbol{\Sigma}}^2\,,\\
\frac{\partial \Xi_\alpha}{\partial\tau} &= -\frac{1-\delta}{2\delta}\beta - \frac{\sigma^2\beta}{2\tau^2} - \frac{\beta}{2\tau^2 n}\|\eta_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau\boldsymbol{g}, \beta/\tau) - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{\beta\|\boldsymbol{g}\|_2^2}{2n}\,.
\end{aligned}$$

Taking expectations with respect to $\boldsymbol{g}$, exchanging expectations and derivatives by dominated convergence, and expanding the square in the first line, we conclude

$$\begin{aligned}
\frac{\partial\psi_\alpha(\tau,\beta)}{\partial\beta} &= -\beta + \frac{\tau}{2} + \frac{\sigma^2}{2\tau} + \frac{1}{2\tau}\mathsf{R}_\alpha(\tau^2,\beta/\tau) - \tau\,\mathsf{df}_\alpha(\tau^2,\beta/\tau)\\
&= \tau\left(-\frac{\beta}{\tau} + 1 - \mathsf{df}_\alpha(\tau^2,\beta/\tau)\right) + \frac{1}{2\tau}\left(-\tau^2 + \sigma^2 + \mathsf{R}_\alpha(\tau^2,\beta/\tau)\right),\\
\frac{\partial\psi_\alpha(\tau,\beta)}{\partial\tau} &= \frac{\beta}{2} - \frac{\sigma^2\beta}{2\tau^2} - \frac{\beta}{2\tau^2}\mathsf{R}_\alpha(\tau^2,\beta/\tau) = \frac{\beta}{2\tau^2}(\tau^2 - \sigma^2 - \mathsf{R}_\alpha(\tau^2,\beta/\tau))\,.
\end{aligned}$$

Thus, if $(\tau_\alpha^*,\zeta_\alpha^*) = (\tau_\alpha^*,\beta_\alpha^*/\tau_\alpha^*)$ solves Eqs. (40a) and (40b), the derivatives in the preceding display are 0. Because $\psi_\alpha(\tau,\beta)$ is convex-concave in (τ,β), we conclude that, for any $\tau,\beta > 0$, Eq. (48) holds. Thus, $(\tau_\alpha^*,\beta_\alpha^*)$ is a saddle-point of ψ_α (see, e.g., [43, pg. 380]). By [43, Lemma 36.2], the max-min value of (47) is achieved at $(\tau_\alpha^*,\beta_\alpha^*)$, and the maximization and minimization may be exchanged as in Eq. (49). □

B. Proofs of main results.

B.1. *Control of α-smoothed Lasso estimate and proof of Theorem 4.* The following theorem controls the behavior of the α-smoothed lasso.

THEOREM B.1. *If assumptions A1 and A2$_\alpha$ hold, then there exist constants $C, c, c', \gamma > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for any 1-Lipschitz function $\phi : \mathbb{R}^p \to \mathbb{R}$, we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\exists\boldsymbol{\theta}\in\mathbb{R}^p,\ \left|\phi\Big(\frac{\boldsymbol{\theta}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}_\alpha^f}{\sqrt{p}}\Big)\Big]\right| > \epsilon \text{ and } \mathcal{R}_\alpha(\boldsymbol{\theta}) \le \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\mathcal{R}_\alpha(\boldsymbol{\theta}) + \gamma\epsilon^2\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}.$$

Theorem 4 is an immediate corollary of Theorem B.1.

PROOF OF THEOREM 4. Because $\boldsymbol{\theta}^*$ is deterministic, $\boldsymbol{\theta}/\sqrt{p} \mapsto \phi(\boldsymbol{\theta}/\sqrt{p}, \boldsymbol{\theta}^*/\sqrt{p})$ is a 1-Lipschitz function. Apply Theorem B.1 with $\alpha = 0$. □

Define the error vectors of the α-smoothed Lasso in the random-design model,

$$\widehat{\boldsymbol{w}}_\alpha := \widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*\,,\quad \widehat{\boldsymbol{v}}_\alpha := \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*)\,, \tag{50}$$

where $\widehat{\boldsymbol{\theta}}_\alpha$ is defined by (37). The error vector $\widehat{\boldsymbol{v}}_\alpha$ is the minimizer of the reparameterized objective

(51)
$$\begin{aligned}\mathcal{C}_\alpha(\boldsymbol{v}) &:= \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v} - \sigma\boldsymbol{z}\|_2^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1)\\ &= \max_{\boldsymbol{u}\in\mathbb{R}^n}\left\{\frac{1}{n}\boldsymbol{u}^\top(\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v} - \sigma\boldsymbol{z}) - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1)\right\} =: \max_{\boldsymbol{u}\in\mathbb{R}^n} C_\alpha(\boldsymbol{v},\boldsymbol{u})\,.\end{aligned}$$

We also define the error vector of the α-smoothed Lasso in the fixed-design model

$$\widehat{\boldsymbol{v}}_\alpha^f := \boldsymbol{\Sigma}^{1/2}(\eta_\alpha(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau_\alpha^*\boldsymbol{g}, \zeta_\alpha^*) - \boldsymbol{\theta}^*)\,. \tag{52}$$

We control the behavior of α-smoothed Lasso error $\widehat{\boldsymbol{v}}_\alpha$ in the random-design model using Gordon's minimax theorem [49, 35]. Define Gordon's objective by

(53)
$$\begin{aligned}\mathcal{L}_\alpha(\boldsymbol{v}) &:= \frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\,\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)_+^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1)\\ &= \max_{\boldsymbol{u}\in\mathbb{R}^n}\left\{-\frac{1}{n^{3/2}}\|\boldsymbol{u}\|_2\boldsymbol{g}^\top\boldsymbol{v} + \frac{1}{n}\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\,\boldsymbol{h}^\top\boldsymbol{u} - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1)\right\}\\ &=: \max_{\boldsymbol{u}\in\mathbb{R}^n} L_\alpha(\boldsymbol{v},\boldsymbol{u})\,,\end{aligned}$$

where $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}_p, \mathbf{I}_p)$, $\boldsymbol{h} \sim \mathsf{N}(\boldsymbol{0}_n, \mathbf{I}_n)$ and $\xi \sim \mathsf{N}(0,1)$ are all independent. Gordon's lemma compares the (possibly constrained) minimization of $\mathcal{C}_\alpha(\boldsymbol{v})$ with the corresponding minimization of $\mathcal{L}_\alpha(\boldsymbol{v})$.

LEMMA B.2 (Gordon's lemma). *The following hold.*

(a) *Let $D \subset \mathbb{R}^p$ be a closed set. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{C}_\alpha(\boldsymbol{v}) \le t\right) \le 2\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{L}_\alpha(\boldsymbol{v}) \le t\right)\,.$$

(b) *Let $D \subset \mathbb{R}^p$ be a closed, convex set. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{C}_\alpha(\boldsymbol{v}) \ge t\right) \le 2\mathbb{P}\left(\min_{\boldsymbol{v}\in D}\mathcal{L}_\alpha(\boldsymbol{v}) \ge t\right)\,.$$

We prove Lemma B.2 later in this section.

PROOF OF THEOREM B.1. For any set D, define $D_\epsilon := \{\boldsymbol{x} \in \mathbb{R}^p \mid \inf_{\boldsymbol{x}'\in D}\|\boldsymbol{x} - \boldsymbol{x}'\|_2/\sqrt{p} \ge \epsilon\}$. Denote $L_\alpha^* := \psi_\alpha(\tau_\alpha^*, \beta_\alpha^*)$ where $\tau_\alpha^*, \beta_\alpha^*$ are as in Lemma A.6. To control $\widehat{\boldsymbol{v}}_\alpha$ using Gordon's lemma, we show that with high probability the minimal value of $\mathcal{L}_\alpha$ is close to L_α^*, and that if D contains $\widehat{\boldsymbol{v}}_\alpha^f$ with high probability, the objective $\mathcal{L}_\alpha$ is uniformly sub-optimal on D_ϵ with high probability. We need the following lemma.

LEMMA B.3. *There exist constants $C, c, c', \gamma > 0$, depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ, such that for $\epsilon \in (0, c')$, we have*

$$\min_{\boldsymbol{v}\in\mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f;\epsilon/2)}\mathcal{L}_\alpha(\boldsymbol{v}) > L_\alpha^* + 2\gamma\epsilon^2\,, \qquad |\min_{\boldsymbol{v}\in\mathbb{R}^p}\mathcal{L}_\alpha(\boldsymbol{v}) - L_\alpha^*| \le \gamma\epsilon^2\,, \tag{54}$$

with probability at least $1 - \frac{C}{\epsilon^2}\exp(-cn\epsilon^4)$.

We prove Lemma B.3 at the end of this section.

With $C, c, c', \gamma > 0$ as in Lemma B.3, we have for $\epsilon < c'$

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D_{\epsilon/2}} \mathcal{C}_\alpha(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) + \gamma\epsilon^2\right)$$

$$\le \mathbb{P}\left(\min_{\boldsymbol{v}\in D_{\epsilon/2}} \mathcal{C}_\alpha(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) + \gamma\epsilon^2 \text{ and } \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) \le L_\alpha^* + \gamma\epsilon^2\right) + \mathbb{P}\left(\min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) > L_\alpha^* + \gamma\epsilon^2\right)$$

$$\le \mathbb{P}\left(\min_{\boldsymbol{v}\in D_{\epsilon/2}} \mathcal{C}_\alpha(\boldsymbol{v}) \le L_\alpha^* + 2\gamma\epsilon^2\right) + \mathbb{P}\left(\min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) > L_\alpha^* + \gamma\epsilon^2\right)$$

$$\le 2\mathbb{P}\left(\min_{\boldsymbol{v}\in D_{\epsilon/2}} \mathcal{L}_\alpha(\boldsymbol{v}) \le L_\alpha^* + 2\gamma\epsilon^2\right) + 2\mathbb{P}\left(\min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) > L_\alpha^* + \gamma\epsilon^2\right)$$

$$\le 2\mathbb{P}\left(\widehat{\boldsymbol{v}}_\alpha^f \notin D\right) + 2\mathbb{P}\left(\min_{\boldsymbol{v}\in\mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f;\epsilon/2)} \mathcal{L}_\alpha(\boldsymbol{v}) \le L_\alpha^* + 2\gamma\epsilon^2\right) + 2\mathbb{P}\left(\min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) > L_\alpha^* + \gamma\epsilon^2\right)$$

(55)
$$\le 2\mathbb{P}\left(\widehat{\boldsymbol{v}}_\alpha^f \notin D\right) + \frac{4C}{\epsilon^2} e^{-cn\epsilon^4},$$

where the third-to-last inequality holds by Gordon's Lemma (Lemma B.2); the second to last inequality holds because either $\widehat{\boldsymbol{v}}_\alpha^f \notin D$ or $D_{\epsilon/2} \subset \mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f;\epsilon/2)$; and the last inequality holds by Lemma B.3.

Define $\tilde{\phi}\left(\frac{\boldsymbol{v}}{\sqrt{p}}\right) := \kappa_{\min}^{1/2}\phi\left(\frac{\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}}{\sqrt{p}}\right)$ (recall that $\boldsymbol{\theta}^*$ is deterministic), with ϕ as in the statement of Theorem B.1. Define the set

$$D := \left\{\boldsymbol{v} \in \mathbb{R}^p \;\middle|\; \left|\tilde{\phi}\left(\frac{\boldsymbol{v}}{\sqrt{p}}\right) - \mathbb{E}\left[\tilde{\phi}\left(\frac{\widehat{\boldsymbol{v}}_\alpha^f}{\sqrt{p}}\right)\right]\right| \le \frac{\epsilon}{2}\right\}.$$

By Eq. (52) and recalling that $\beta_\alpha^* = \zeta_\alpha^*\tau_\alpha^*$, we have

$$\widehat{\boldsymbol{v}}_\alpha^f = \arg\min_{\boldsymbol{v}\in\mathbb{R}^p}\left\{\frac{\beta_\alpha^*}{2\tau_\alpha^*}\|\boldsymbol{v} - \tau_\alpha^*\boldsymbol{g}\|_2^2 + \lambda\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\|_1\right\}. \tag{56}$$

Thus, $\widehat{\boldsymbol{v}}_\alpha^f$ as a function of $\tau_\alpha^*\boldsymbol{g}$ is a proximal operator, whence $\widehat{\boldsymbol{v}}_\alpha^f$ is a τ_α^*-Lipschitz function of $\boldsymbol{g}$ [40, pg. 131]. Gaussian concentration of Lipschitz functions [13, Theorem 5.6] guarantees that

$$\mathbb{P}\left(\widehat{\boldsymbol{v}}_\alpha^f \notin D\right) \le 2\exp\left(-\frac{p\epsilon^2}{8\tau_\alpha^{*2}}\right) = 2\exp\left(-\frac{n\epsilon^2}{8\tau_\alpha^{*2}\delta}\right) \le 2\exp\left(-\frac{n\epsilon^2}{8\tau_{\max}^2\delta}\right).$$

Combined with Eq. (55) and appropriately adjusting constants, for $\epsilon < c'$

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in D_{\epsilon/2}} \mathcal{C}_\alpha(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_\alpha(\boldsymbol{v}) + \gamma\epsilon^2\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}.$$

Because $\mathcal{C}_\alpha$ is a reparameterization of the α-smoothed Lasso objective, the preceding display is equivalent to

$$\mathbb{P}\left(\exists\boldsymbol{\theta}\in\mathbb{R}^p, \left|\phi\left(\frac{\boldsymbol{\theta}}{\sqrt{p}}\right) - \mathbb{E}\left[\phi\left(\frac{\widehat{\boldsymbol{\theta}}_\alpha^f}{\sqrt{p}}\right)\right]\right| > \kappa_{\min}^{-1/2}\epsilon \text{ and } \mathcal{R}_\alpha(\boldsymbol{\theta}) \le \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\mathcal{R}_\alpha(\boldsymbol{\theta}) + \gamma\epsilon^2\right) \le \frac{4C}{\epsilon^2}e^{-cn\epsilon^4}.$$

Theorem B.1 follows by a change of variables. $\square$

PROOF OF LEMMA B.2. Because $\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) \to \infty$ as $\|\boldsymbol{v}\|_2 \to \infty$,

$$\min_{\boldsymbol{v}\in D} \mathcal{C}_\alpha(\boldsymbol{v}) = \lim_{R\to\infty} \min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \mathcal{C}_\alpha(\boldsymbol{v})\,.$$

Note that $\arg\max_{\boldsymbol{u}\in\mathbb{R}^n} C_\alpha(\boldsymbol{v},\boldsymbol{u}) = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v} - \sigma\boldsymbol{z}$ has ℓ_2-norm no larger than $\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}\|\boldsymbol{v}\|_2 + \sigma\|\boldsymbol{z}\|_2$. In particular, for any realization of $\boldsymbol{X},\boldsymbol{z}$, we have for R sufficiently large that $\|\boldsymbol{v}\|_2 \le R$ implies $\|\arg\max_{\boldsymbol{u}\in\mathbb{R}^n} C_\alpha(\boldsymbol{v},\boldsymbol{u})\|_2 \le R^2$. In particular, for any realization of $\boldsymbol{X},\boldsymbol{z}$

$$\min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \mathcal{C}_\alpha(\boldsymbol{v}) = \min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} C_\alpha(\boldsymbol{v},\boldsymbol{u}) \quad \text{for } R \text{ sufficiently large,}$$

where "sufficiently large" can depend on $\boldsymbol{X},\boldsymbol{z}$. Thus, almost surely

$$\min_{\boldsymbol{v}\in D} \mathcal{C}_\alpha(\boldsymbol{v}) = \lim_{R\to\infty} \min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} C_\alpha(\boldsymbol{v},\boldsymbol{u})\,.$$

An equivalent argument shows that almost surely

$$\min_{\boldsymbol{v}\in D} \mathcal{L}_\alpha(\boldsymbol{v}) = \lim_{R\to\infty} \min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} L_\alpha(\boldsymbol{v},\boldsymbol{u})\,.$$

Because $\sqrt{n}\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ has iid standard Gaussian entries, by Gordon's min-max lemma (see, e.g., [35, Corollary G.1]), for any finite R and closed D

$$\mathbb{P}\left(\min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} C_\alpha(\boldsymbol{v},\boldsymbol{u}) < t\right) \le 2\mathbb{P}\left(\min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} L_\alpha(\boldsymbol{v},\boldsymbol{u}) < t\right)\,,$$

and if D is also convex

$$\mathbb{P}\left(\min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} C_\alpha(\boldsymbol{v},\boldsymbol{u}) > t\right) \le 2\mathbb{P}\left(\min_{\substack{\boldsymbol{v}\in D\\ \|\boldsymbol{v}\|_2\le R}} \max_{\|\boldsymbol{u}\|_2\le R^2} L_\alpha(\boldsymbol{v},\boldsymbol{u}) > t\right)\,.$$

Although [35, Corollary G.1]) states Gordon's lemma with weak inequalities inside the probabilities, strict inequalities follow by applying [35, Corollary G.1]) with $t' \uparrow t$and $t' \downarrow t$ in the previous two displays respectively. Taking $R \to \infty$, we conclude that the previous two displays hold without norm bounds on for R sufficiently large$\boldsymbol{v}$ and $\boldsymbol{u}$. The strict inequalities can be made weak by applying the result with $t' \downarrow t$ and $t' \uparrow t$ respectively. □

PROOF OF LEMMA B.3. Recall by Lemma A.6 that the max-min value of (47) is achieved at τ_α^*, β_α^*. We have $\beta_{\min} \le \beta_\alpha^* \le \beta_{\max}$, where $\beta_{\min} := \tau_{\min}\zeta_{\min}$ and $\beta_{\max} := \tau_{\max}\zeta_{\max}$. Let $t = \min(\beta_{\min}/16, \sigma_{\min})$. Define events

$$\mathcal{A}_1 := \left\{\ \|\boldsymbol{g}\|_2 \le 2\sqrt{p},\ \left(1 - \frac{\beta_{\min}}{8\tau_{\max}}\right) \le \frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} \le 2\right\}\,,$$
$$\mathcal{A}_2 := \left\{\left|\frac{\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2}{n} - \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2}{n}\right]\right| \le t^2,\ \frac{\boldsymbol{g}^\top\widehat{\boldsymbol{v}}_\alpha^f}{n} \le \mathbb{E}\left[\frac{\boldsymbol{g}^\top\widehat{\boldsymbol{v}}_\alpha^f}{n}\right] + t\right\}\,.$$

There exist $r, a > 0$, depending only on $\beta_{\min}, \beta_{\max}, \tau_{\min}, \tau_{\max}$ such that on the event $\mathcal{A}_1 \cap \mathcal{A}_2$ the objective $\mathcal{L}_\alpha$ is a/n-strongly convex on $\mathsf{B}_2(\widehat{\boldsymbol{v}}_\alpha^f; r)$. This follows verbatim from the proof of Theorem B.1 in [35] up to the first display on pg. 28.

Let $R = \tau_{\max} + r$, $\gamma = a/(96\delta)$, $c' = \sqrt{ar^2/(24\gamma)}$, and $\epsilon \in (0, c')$. Define events

$$\mathcal{A}_3 := \left\{ \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}) \ge L_\alpha^* - \gamma\epsilon^2 \right\},$$

$$\mathcal{A}_4 := \left\{ \mathcal{L}_\alpha(\widehat{\boldsymbol{v}}_\alpha^f) \le L_\alpha^* + \gamma\epsilon^2 \right\}.$$

On event $\mathcal{A}_3 \cap \mathcal{A}_4$,

$$\mathcal{L}_\alpha(\widehat{\boldsymbol{v}}_\alpha^f) \le \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}) + 2\gamma\epsilon^2 < \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}) + 3\gamma\epsilon^2.$$

Because $3\gamma\epsilon^2 < ar^2/8$, the previous display corresponds to (49) of [35]. Thus, the last paragraph verbatim of the proof of Theorem B.1 in [35] implies that on $\bigcap_{i=1}^4 \mathcal{A}_i$,

$$\min_{\boldsymbol{v} \in \mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f; \epsilon/2)} \mathcal{L}_\alpha(\boldsymbol{v}) = \min_{\boldsymbol{v} \in \mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f; \sqrt{8\delta \cdot 3\gamma\epsilon^2/a})} \mathcal{L}_\alpha(\boldsymbol{v}) \ge \min_{\boldsymbol{v} \in \mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) + 3\gamma\epsilon^2.$$

Moreover, the last line of the proof of Theorem B.1 in [35] shows that the preceding display follows from Lemma B.1 in [35], whose statement additionally implies

$$\min_{\boldsymbol{v} \in \mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) = \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}).$$

We conclude that on $\bigcap_{i=1}^4 \mathcal{A}_i$,

$$\min_{\boldsymbol{v} \in \mathsf{B}_2^c(\widehat{\boldsymbol{v}}_\alpha^f; \epsilon/2)} \mathcal{L}_\alpha(\boldsymbol{v}) \ge \min_{\boldsymbol{v} \in \mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) + 3\gamma\epsilon^2 = \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}) + 3\gamma\epsilon^2 \ge L_\alpha^* + 2\gamma\epsilon^2,$$

and

$$L_\alpha^* + \gamma\epsilon^2 \ge \mathcal{L}_\alpha(\widehat{\boldsymbol{v}}_\alpha^f) \ge \min_{\boldsymbol{v} \in \mathbb{R}^p} \mathcal{L}_\alpha(\boldsymbol{v}) = \min_{\|\boldsymbol{v}\|_2 \le \sqrt{n} R} \mathcal{L}_\alpha(\boldsymbol{v}) \ge L_\alpha^* - \gamma\epsilon^2.$$

Lemma B.3 follows as soon as we show there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that for $\epsilon < c'$ we have $\mathbb{P}(\cap_{i=1}^4 \mathcal{A}_i) \ge 1 - \frac{C}{\epsilon^2} \exp(c\epsilon^4)$.

Now to complete the proof of Lemma B.3, it is only left for us to control the probability of each $\mathcal{A}_i$ respectively.

Event $\mathcal{A}_1$ occurs with high probability depending on $\beta_{\min}, \tau_{\max}, \delta$. Because $\boldsymbol{g} \mapsto \|\boldsymbol{g}\|_2$ and $\boldsymbol{h} \mapsto \|\boldsymbol{h}\|_2$ are Lipschitz functions of standard Gaussian random vectors, there exist C, c depending only on $\beta_{\min}, \tau_{\max}, \delta$ such that

$$\mathbb{P}(\mathcal{A}_1) \ge 1 - C\exp(-cn).$$

Event $\mathcal{A}_2$ occurs with high probability depending on $\sigma_{\min}, \beta_{\min}, \tau_{\max}.$. The function $\boldsymbol{g} \mapsto n^{-1/2}\|\widehat{\boldsymbol{v}}_\alpha^f\|_2$ is $n^{-1/2}\tau_{\max}$-Lipschitz because $\widehat{\boldsymbol{v}}_\alpha^f$ is a proximal operator applied to $\tau_\alpha^* \boldsymbol{g}$ by Eq. (56) [40, pg. 131]. By Gaussian concentration of Lipschitz functions, $n^{-1/2}\|\widehat{\boldsymbol{v}}_\alpha^f\|_2$ is $\tau_{\max}^2/n$-sub-Gaussian. By the fixed point equations (40a), we bound its expectation $\mathbb{E}[n^{-1/2}\|\widehat{\boldsymbol{v}}_\alpha^f\|_2] \le n^{-1/2}\mathbb{E}[\|\widehat{\boldsymbol{v}}_\alpha^f\|^2]^{1/2} \le \tau_{\max}$. Combining its sub-Gaussianity and bounded expectation, we conclude by Proposition G.5 of [35] that

$$\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2/n \text{ is } (C/n, C/n)\text{-sub-Gamma for some } C \text{ depending only on } \tau_{\max}.$$

Write

$$\tau_{\max}\boldsymbol{g}^\top \widehat{\boldsymbol{v}}_\alpha^f/n = (\|\widehat{\boldsymbol{v}}_\alpha^f - \tau_{\max}\boldsymbol{g}\|_2^2 - \|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2 - \tau_{\max}^2\|\boldsymbol{g}\|_2^2)/(2n).$$

Because $\boldsymbol{g} \mapsto \widehat{\boldsymbol{v}}_\alpha^f - \tau_{\max}\boldsymbol{g}$ is $2\tau_{\max}$-Lipschitz, the first term is $(C/n, C/n)$-sub-Gamma for some C depending only on $\tau_{\max}$. We conclude[6]

$$\tau_{\max}\boldsymbol{g}^\top \widehat{\boldsymbol{v}}_\alpha^f/n \text{ is } (C/n, C/n)\text{-sub-Gamma for some } C \text{ depending only on } \tau_{\max}.$$

By standard bounds on the tails of sub-Gamma random variables, we deduce that there exist $C, c > 0$ depending only on $\tau_{\max}$, such that

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2}{n} - \mathbb{E}\left[\frac{\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2}{n}\right]\right| > \epsilon\right) \le C\exp\big(-cn(\epsilon^2 \vee \epsilon)\big)\,, \\
&\mathbb{P}\left(\left|\frac{\boldsymbol{g}^\top\widehat{\boldsymbol{v}}_\alpha^f}{n} - \mathbb{E}\left[\frac{\boldsymbol{g}^\top\widehat{\boldsymbol{v}}_\alpha^f}{n}\right]\right| > \epsilon\right) \le C\exp\big(-cn(\epsilon^2 \vee \epsilon)\big)\,.
\end{aligned} \tag{57}$$

Because t depends only on $\sigma_{\min}, \beta_{\min}$, there exists $C, c > 0$ depending only on $\sigma_{\min}, \beta_{\min}, \tau_{\max}$ such that

$$\mathbb{P}(\mathcal{A}_2) \ge 1 - C\exp(-cn)\,.$$

Event $\mathcal{A}_3$ occurs with high probability depending on $\sigma_{\max}$, δ, $\beta_{\min}$, $\beta_{\max}$, $\tau_{\min}$, $\tau_{\max}$. Our control on the probability of $\mathcal{A}_3$ closely follows the proof of Proposition B.2 in [35]. Consider for any $\epsilon > 0$ the event

$$\mathcal{A}_3^{(1)} := \left\{\left|\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - 1\right| \le \epsilon\right\}. \tag{58}$$

By Gaussian concentration of Lipschitz functions, $\mathbb{P}(\mathcal{A}_3^{(1)}) \ge Ce^{-cn\epsilon^2}$ for all $\epsilon \ge 0$.

By maximizing over over $\boldsymbol{u}$ for which $\|\boldsymbol{u}\|/\sqrt{n} = \beta$ in Eq. (53), we compute

$$\begin{aligned}
\mathcal{L}_\alpha(\boldsymbol{v}) &= \max_{\beta\ge 0}\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)\beta - \frac{1}{2}\beta^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \\
&=: \max_{\beta\ge 0}\ell_\alpha(\boldsymbol{v}, \beta)\,.
\end{aligned}$$

Consider the slightly modified objective

$$\ell_\alpha^0(\boldsymbol{v}, \beta) := \left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)\beta - \frac{1}{2}\beta^2 + \frac{\lambda}{n}(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1)\,.$$

On the event (58), for every $\|\boldsymbol{v}\|_2/\sqrt{n} \le R$ and $\beta \in [0, \beta_{\max}]$,

$$|\ell_\alpha(\boldsymbol{v}, \beta) - \ell_\alpha^0(\boldsymbol{v}, \beta)| \le \beta_{\max}(R^2 + \sigma^2)^{1/2}\epsilon\,.$$

Thus, on the event (58),

$$\begin{aligned}
\min_{\|\boldsymbol{v}\|_2/\sqrt{n}\le R}\mathcal{L}_\alpha(\boldsymbol{v}) &= \min_{\|\boldsymbol{v}\|_2/\sqrt{n}\le R}\max_{\beta\ge 0}\ell_\alpha(\boldsymbol{v}, \beta) \ge \min_{\|\boldsymbol{v}\|_2/\sqrt{n}\le R}\ell_\alpha(\boldsymbol{v}, \beta_\alpha^*) \\
&\ge \min_{\|\boldsymbol{v}\|_2/\sqrt{n}\le R}\ell_\alpha^0(\boldsymbol{v}, \beta_\alpha^*) - \beta_{\max}(R^2 + \sigma^2)^{1/2}\epsilon\,.
\end{aligned} \tag{59}$$

---

[6]We remark taht the argument establishing that $\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2/n$ is sub-Gamma is exactly as it occurs in the proof of Lemma F.1 of [35]. The argument establishing $\|\widehat{\boldsymbol{v}}_\alpha^f\|_2^2/n$ requires a slightly modified argument to that appearing in the proof of Lemma F.1 in [35] due to the presense of the matrix $\boldsymbol{\Sigma}$.

For $\|\boldsymbol{v}\|_2/\sqrt{n} \le R$,

$$\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2} = \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} \left\{ \frac{\|\boldsymbol{v}\|_2^2/n + \sigma^2}{2\tau} + \frac{\tau}{2} \right\}.$$

Thus, we obtain that

$$\ell^0_\alpha(\boldsymbol{v}, \beta^*_\alpha) = \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} \left\{ \left(\frac{\|\boldsymbol{v}\|_2^2/n + \sigma^2}{2\tau} + \frac{\tau}{2} \right) \beta^*_\alpha - \frac{\boldsymbol{g}^\top \boldsymbol{v}}{n} \beta^*_\alpha - \frac{1}{2} \beta^{*\,2}_\alpha + \frac{\lambda}{n} (\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\},$$

which further implies that

$$\begin{aligned}
&\min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \ell^0_\alpha(\boldsymbol{v}, \beta^*_\alpha) \\
&= \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} \left\{ \frac{\beta^*_\alpha}{2} \left(\frac{\sigma^2}{\tau} + \tau \right) - \frac{1}{2} \beta^{*\,2}_\alpha + \frac{1}{n} \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \left\{ \frac{\beta^*_\alpha}{2\tau} \|\boldsymbol{v}\|_2^2 - \beta^*_\alpha \boldsymbol{g}^\top \boldsymbol{v} + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\} \right\} \\
&=: \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} F(\tau, \boldsymbol{g}).
\end{aligned}$$

We claim that $F(\tau, \boldsymbol{g})$ concentrates around its expectation. In order to see this, first note that for every $\tau \in [\sigma, \sqrt{\sigma^2+R^2}]$, the function

$$\boldsymbol{g} \mapsto \frac{1}{n} \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \left\{ \frac{\beta^*_\alpha}{2\tau} \|\boldsymbol{v}\|_2^2 - \beta^*_\alpha \boldsymbol{g}^\top \boldsymbol{v} + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\}$$

is $\beta_{\max} R/\sqrt{n}$-Lipschitz, whence $\boldsymbol{g} \mapsto F(\tau, \boldsymbol{g})$ is as well. By Gaussian concentration of Lipschitz functions [13, Theorem 5.6],

$$\mathbb{P}(|F(\tau, \boldsymbol{g}) - \mathbb{E}[F(\tau, \boldsymbol{g})]| > \epsilon) \le 2e^{-cn\epsilon^2},$$

for $c = 1/(2{\beta_{\max}}^2 R^2)$. Because $\tau \ge \tau_{\min} > 0$, for all $\boldsymbol{g}$ the function $\tau \mapsto F(\tau, \boldsymbol{g})$ is $(\beta_{\max} + \beta_{\max} R^2/(2\tau_{\min}^2))$-Lipschitz on $[\sigma, \sqrt{\sigma^2+R^2}]$, so that by an ϵ-net argument, we conclude that for C, c depending only on $R, \beta_{\max}, \tau_{\min}$ that

$$\mathbb{P}\left(\mathcal{A}_3^{(2)}\right) := \mathbb{P}\left(\sup_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} |F(\tau, \boldsymbol{g}) - \mathbb{E}[F(\tau, \boldsymbol{g})]| \le \epsilon \right) \ge 1 - \frac{C}{\epsilon} e^{-cn\epsilon^2}.$$

On $\mathcal{A}_3^{(2)}$,

$$\min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \ell^0_\alpha(\boldsymbol{v}, \beta_{\max}) = \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} F(\tau, \boldsymbol{g}) \ge \min_{\tau \in [\sigma, \sqrt{\sigma^2+R^2}]} \mathbb{E}[F(\tau, \boldsymbol{g})] - \epsilon. \tag{60}$$

We compute

$$\begin{aligned}
F(\tau, \boldsymbol{g}) &= \frac{\beta^*_\alpha}{2} \left(\frac{\sigma^2}{\tau} + \tau \right) - \frac{1}{2} \beta^{*\,2}_\alpha + \frac{1}{n} \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \left\{ \frac{\beta^*_\alpha}{2\tau} \|\boldsymbol{v}\|_2^2 - \beta^*_\alpha \boldsymbol{g}^\top \boldsymbol{v} + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\} \\
&= \frac{\beta^*_\alpha}{2} \left(\frac{\sigma^2}{\tau} + \tau \right) - \frac{1}{2} \beta^{*\,2}_\alpha - \frac{\beta^*_\alpha \tau \|\boldsymbol{g}\|_2^2}{2n} + \frac{1}{n} \min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \left\{ \frac{\beta^*_\alpha}{2\tau} \|\boldsymbol{v} - \tau \boldsymbol{g}\|_2^2 + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\} \\
&\ge \frac{\beta^*_\alpha}{2} \left(\frac{\sigma^2}{\tau} + \tau \right) - \frac{1}{2} \beta^{*\,2}_\alpha - \frac{\beta^*_\alpha \tau \|\boldsymbol{g}\|_2^2}{2n} + \frac{1}{n} \min_{\boldsymbol{v} \in \mathbb{R}^p} \left\{ \frac{\beta^*_\alpha}{2\tau} \|\boldsymbol{v} - \tau \boldsymbol{g}\|_2^2 + \lambda(\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \|\boldsymbol{\theta}^*\|_1) \right\}.
\end{aligned}$$

Taking expectations, we have for any $\tau \ge 0$

$$\mathbb{E}[F(\tau, \boldsymbol{g})] = \psi_\alpha(\tau, \beta^*_\alpha), \tag{61}$$

where ψ_α is defined in (47).

Combining Eqs. (59), (60), and (61), we conclude that on $\mathcal{A}_3^{(1)} \cap \mathcal{A}_3^{(2)}$

$$\min_{\|\boldsymbol{v}\|_2/\sqrt{n} \le R} \mathcal{L}_\alpha(\boldsymbol{v}) \ge \psi_\alpha(\tau_\alpha^*, \beta_\alpha^*) - K\epsilon\,,$$

with $K = \beta_{\max}\sqrt{R^2 + \sigma^2} + 1$. By a change of variables and applying the probability bounds on $\mathcal{A}_3^{(1)}$ and $\mathcal{A}_3^{(2)}$ establishes

$$\mathbb{P}(\mathcal{A}_3) \ge \mathbb{P}(\mathcal{A}_3^{(1)} \cap \mathcal{A}_3^{(2)}) \ge 1 - \frac{C}{\epsilon^2}\exp\left(-cn\epsilon^4\right)\,,$$

for some C, c depending only on R, $\sigma_{\max}$, $\beta_{\max}$, $\tau_{\min}$, and γ. Because $R = \tau_{\max} + r$ and r depends only on $\beta_{\min}$, $\beta_{\max}$, $\tau_{\min}$, $\tau_{\max}$, and $\gamma = a/(96\delta)$ and a depeneds only on $\beta_{\min}$, $\beta_{\max}$, $\tau_{\min}$, $\tau_{\max}$, the constants C, c depend only on $\sigma_{\max}$, δ, $\beta_{\min}$, $\beta_{\max}$, $\tau_{\min}$, $\tau_{\max}$.

Event $\mathcal{A}_4$ occurs with high probability depending on $\gamma, \sigma_{\min}, \kappa_{\min}, \delta, \beta_{\min}, \beta_{\max}, \tau_{\min}, \tau_{\max}$. There exist $C, c > 0$ depending only on $\kappa_{\min}, \delta, \tau_{\max}$ such that for $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}_\alpha^f)}{n} - \mathbb{E}\left[\frac{\mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}_\alpha^f)}{n}\right] > \epsilon\right) \le C\exp(-cn\epsilon^2)\,,$$

because $\boldsymbol{g} \mapsto \mathsf{M}_\alpha(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}_\alpha^f)/n$ is $\kappa_{\min}^{-1/2}\delta^{-1/2}\tau_{\max}/\sqrt{n}$-Lipschitz. For any $x_0 \ge 0$, note that $x \mapsto x_+^2$ is locally Lipschitz in any ball around x_0 with Lipschitz constant and ball radius depending only on an upper bound on $|x_0|$. Thus, considering $x_0 = \beta_\alpha^*$, there exists $L, c' > 0$ depending only on $\beta_{\max}$ such that for $\epsilon < c'$, if $\mathcal{A}_2$ and $\mathcal{A}_3^{(1)}$ occur, then

$$\left|\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)_+^2 - \beta_\alpha^{*\,2}\right| \le L\epsilon\,.$$

Using the probability bounds on $\mathcal{A}_2$ and $\mathcal{A}_3^{(1)}$ and absorbing L into constants, we may find $C, c, c' > 0$ depending only on $\gamma, \sigma_{\min}, \kappa_{\min}, \delta, \beta_{\min}, \tau_{\max}$ such that for $\epsilon < c'$.

$$\mathbb{P}(\mathcal{A}_4) \ge 1 - C\exp(-cn\epsilon^4)\,.$$

Because γ depends only on δ and a, and a depends only on $\beta_{\min}$, $\beta_{\max}$, $\tau_{\min}$, $\tau_{\max}$, the constants $C, c, c' > 0$ depend only on $\gamma, \sigma_{\min}, \kappa_{\min}, \delta, \beta_{\min}, \beta_{\max}, \tau_{\min}, \tau_{\max}$.

Lemma B.3 is established now follows by combining the probability bounds on $\mathcal{A}_i$ for $1 \le i \le 4$. □

B.2. *Control of Lasso residual: proof of Theorem 8.* Same as the proof of Theorem 4, the proof of Theorem 8 uses Gordon's lemma. Specifically, denote

$$\widehat{\boldsymbol{u}} := \boldsymbol{X}\widehat{\boldsymbol{w}} - \sigma\boldsymbol{z} = \boldsymbol{X}\widehat{\boldsymbol{\theta}} - \boldsymbol{y},$$

where $\widehat{\boldsymbol{w}} := \widehat{\boldsymbol{w}}_0 = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ as defined in Eq. (50). Then $\widehat{\boldsymbol{u}}$ is the unique maximizer of

$$\boldsymbol{u} \mapsto \min_{\boldsymbol{w}\in\mathbb{R}^p}\left\{\frac{1}{n}\langle\boldsymbol{X}\boldsymbol{w} - \sigma\boldsymbol{z},\, \boldsymbol{u}\rangle - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}(\|\boldsymbol{w} + \boldsymbol{\theta}^*\|_1 - \|\boldsymbol{\theta}^*\|_1)\right\},$$

where the function on the right hand side (before minimizing over $\boldsymbol{w}$) is defined as $C_0(\boldsymbol{v}, \boldsymbol{u}) =: C(\boldsymbol{v}, \boldsymbol{u})$ in expression (51) with re-parametrization $\boldsymbol{v} := \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}$. Compared with the analysis in Theorem 4 which focuses on the behavior of $\widehat{\boldsymbol{v}}$, the focus of this section is the behavior of $\widehat{\boldsymbol{u}}$.

Study of the corresponding Gordon's problem. Recall Gordon's optimization problem defined in expression (53) with $\alpha = 0$ and $\mathsf{M}_\alpha(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. For every $(\boldsymbol{v}, \boldsymbol{u})$, we have (defining $L(\boldsymbol{v}, \boldsymbol{u}) = L_0(\boldsymbol{v}, \boldsymbol{u})$, cf. Eq (53)):

$$L(\boldsymbol{v}, \boldsymbol{u}) := -\frac{1}{n^{3/2}}\|\boldsymbol{u}\|_2 \boldsymbol{g}^\top \boldsymbol{v} + \frac{1}{n}\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2} \cdot \boldsymbol{h}^\top \boldsymbol{u} - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}(\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v} + \boldsymbol{\theta}^*\|_1 - \|\boldsymbol{\theta}^*\|_1).$$

Denote $U(\boldsymbol{u}) = \min_{\boldsymbol{v}\in\mathbb{R}^p} L(\boldsymbol{v}, \boldsymbol{u})$ and $\widetilde{U}(\boldsymbol{u}) = L(\widehat{\boldsymbol{v}}^f, \boldsymbol{u})$ where $\widehat{\boldsymbol{v}}^f$ is defined in expression (52) with $\alpha = 0$, namely

$$\widehat{\boldsymbol{v}}^f := \boldsymbol{\Sigma}^{1/2}\left[\eta\left(\boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g},\, \frac{\beta^*}{\tau^*}\right) - \boldsymbol{\theta}^*\right].$$

By definition, $U(\boldsymbol{u}) \le \widetilde{U}(\boldsymbol{u})$. From direct calculations, the maximizer of $\widetilde{U}(\boldsymbol{u})$ is

$$\boldsymbol{u} = \left(\sqrt{\frac{\|\widehat{\boldsymbol{v}}^f\|_2^2}{n} + \sigma^2}\|\boldsymbol{h}\|_2 - \frac{\boldsymbol{g}^\top\widehat{\boldsymbol{v}}^f}{\sqrt{n}}\right)_+ \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}.$$

Let us define quantity $\widetilde{\boldsymbol{u}} := \tau^*\zeta^*\boldsymbol{h}$. By the concentration of $\widehat{\boldsymbol{v}}^f$ (given by inequality (57)) and the definition of the (τ^*, ζ^*) in (8a) and (8b), $\widetilde{\boldsymbol{u}}$ is ϵ-close to the maximizer of $\widetilde{U}(\boldsymbol{u})$ (in the sense that $\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}^*\|_2/\sqrt{n} \le \epsilon$). In particular, Lemma D.1 [35] holds verbatim here.

Define the set

$$D_\epsilon := \left\{\boldsymbol{u} \in \mathbb{R}^p \;\middle|\; \left|\phi\Big(\frac{\boldsymbol{u}}{\sqrt{n}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\tau^*\zeta^*\boldsymbol{h}}{\sqrt{n}}\Big)\Big]\right| > \epsilon\right\}.$$

The probability $\mathbb{P}(\widehat{\boldsymbol{u}} \in D_\epsilon)$ can be controlled as

$$\begin{aligned}
\mathbb{P}(\widehat{\boldsymbol{u}} \in D_\epsilon) &= \mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon}\min_{\boldsymbol{w}} C(\boldsymbol{v}, \boldsymbol{u}) \ge \max_{\boldsymbol{u}}\min_{\boldsymbol{w}} C(\boldsymbol{v}, \boldsymbol{u}))\\
&\le \mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon}\min_{\boldsymbol{v}} C(\boldsymbol{v}, \boldsymbol{u}) \ge L^* - \epsilon^2) + \mathbb{P}(\max_{\boldsymbol{u}}\min_{\boldsymbol{v}} C(\boldsymbol{v}, \boldsymbol{u}) \le L^* - \epsilon)\\
&\le 2\mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon}\min_{\boldsymbol{v}} L(\boldsymbol{v}, \boldsymbol{u}) \ge L^* - \epsilon^2) + 2\mathbb{P}(\max_{\boldsymbol{u}}\min_{\boldsymbol{v}} L(\boldsymbol{v}, \boldsymbol{u}) \le L^* - \epsilon^2),
\end{aligned}$$

where the last inequality follows by Gordon's lemma (Lemma B.2). The second term in the last expression is upper bounded $\frac{C}{\epsilon^2}e^{-cn\epsilon^4}$ using same argument as in Theorem 4, more concretely, in Lemma B.3. The first term is upper bounded as

$$2\mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon}\min_{\boldsymbol{v}} L(\boldsymbol{v}, \boldsymbol{u}) \ge L^* - \epsilon^2) = 2\mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon} U(\boldsymbol{u}) \ge L^* - \epsilon^2) \le 2\mathbb{P}(\max_{\boldsymbol{u}\in D_\epsilon} \widetilde{U}(\boldsymbol{u}) \ge L^* - \epsilon^2).$$

We control the right-hand side following verbatim from the proof of Theorem D.1 and Lemma D.1 of [35]. Putting the details above together yields $\mathbb{P}(\widehat{\boldsymbol{u}} \in D_\epsilon) \le \frac{C}{\epsilon^2}\exp(-cn\epsilon^4)$.

B.3. *Control of the subgradient: proof of Lemma 10.* The proof relies on concentration results established in Lemma B.3. To begin with, define for $\|\boldsymbol{t}\|_\infty \le 1$

$$\mathcal{V}(\boldsymbol{t}) := \min_{\boldsymbol{w}\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \sigma\boldsymbol{z}\|_2^2 + \frac{\lambda}{n}\boldsymbol{t}^\top(\boldsymbol{\theta}^* + \boldsymbol{w}) - \frac{\lambda}{n}\|\boldsymbol{\theta}^*\|_1\right\} =: \min_{\boldsymbol{w}\in\mathbb{R}^p} V(\boldsymbol{w}, \boldsymbol{t}).$$

Define, for $\boldsymbol{g} \sim \mathsf{N}(0, \mathbf{I}_p)$, $\boldsymbol{h} \sim \mathsf{N}(0, \mathbf{I}_n)$,

$$\begin{aligned}
\mathcal{T}(\boldsymbol{t}) &:= \min_{\boldsymbol{v}\in\mathbb{R}^p}\left\{\frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)_+^2 + \frac{\lambda}{n}\boldsymbol{t}^\top(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \frac{\lambda}{n}\|\boldsymbol{\theta}^*\|_1\right\}\\
&=: \min_{\boldsymbol{v}\in\mathbb{R}^p} T(\boldsymbol{v}, \boldsymbol{t}).
\end{aligned}$$

We may compare the maximization of $\mathcal{V}(\boldsymbol{t})$ with the maximization of $\mathcal{T}(\boldsymbol{t})$ using Gordon's lemma.

LEMMA B.4. *Let $D \subset \{\boldsymbol{t} \in \mathbb{R}^p \mid \|\boldsymbol{t}\|_\infty \leq 1\}$ be a closed set.*

(a) *For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\max_{\boldsymbol{t} \in D} \mathcal{V}(\boldsymbol{t}) \geq t\right) \leq 2\mathbb{P}\left(\max_{\boldsymbol{t} \in D} \mathcal{T}(\boldsymbol{t}) \geq t\right). \tag{62}$$

(b) *If D is also convex, then for any $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\max_{\boldsymbol{t} \in D} \mathcal{V}(\boldsymbol{t}) \leq t\right) \leq 2\mathbb{P}\left(\max_{\boldsymbol{t} \in D} \mathcal{T}(\boldsymbol{t}) \leq t\right). \tag{63}$$

We prove Lemma B.4 at the end of this section. The maximization of $\mathcal{T}(\boldsymbol{t})$ can be controlled because $\mathcal{T}(\boldsymbol{t})$ is strongly-concave with high probability. We first establish this strong-concavity.

LEMMA B.5. *Under assumption A1, the objective $\mathcal{T}(\boldsymbol{t})$ is $\frac{c_0\lambda_{\min}^2}{n\kappa_{\max}}$-strongly concave on the event*

$$\left\{\frac{\|\boldsymbol{h}\|_2^2}{n} \leq 2,\ \frac{\|\boldsymbol{g}\|_2^2}{n} \leq \frac{2}{\delta}\right\},$$

where $c_0 > 0$ is a constant depending only on δ.

We prove Lemma B.5 at the end of this section. We are ready to prove Lemma 10.

Consider $\alpha = 0$, and let $\boldsymbol{v}^* \in \mathbb{R}^p$ be a minimizer of $\mathcal{L}(\boldsymbol{v}) := \mathcal{L}_{\alpha=0}(\boldsymbol{v})$ defined in (53). Let

$$\begin{aligned}\boldsymbol{t}^* &:= -\frac{n}{\lambda}\boldsymbol{\Sigma}^{1/2}\nabla\left(\boldsymbol{v} \mapsto \frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}\right)_+^2\right)\Bigg|_{\boldsymbol{v}=\boldsymbol{v}^*} \\ &= -\frac{1}{\lambda}\boldsymbol{\Sigma}^{1/2}\left(\sqrt{\frac{\|\boldsymbol{v}^*\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}^*}{n}\right)_+\left(\frac{\|\boldsymbol{h}\|_2/\sqrt{n}}{\sqrt{\|\boldsymbol{v}^*\|_2^2/n + \sigma^2}}\boldsymbol{v}^* - \boldsymbol{g}\right).\end{aligned}$$

By the KKT conditions, $\frac{\lambda}{n}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{t}^* \in \frac{\lambda}{n}\partial(\boldsymbol{v} \mapsto \|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}\|_1)$ at $\boldsymbol{v} = \boldsymbol{v}^*$. With this definition, $\mathbf{0}_p$ is in the subdifferential with respect to $\boldsymbol{v}$ of $T(\boldsymbol{v}, \boldsymbol{t})$ at $(\boldsymbol{v}^*, \boldsymbol{t}^*)$. Moreover, $t_j^* = 1$ whenever $(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}^*)_j > 0$ and $t_j^* = -1$ whenever $(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}^*)_j < 0$, whence $\boldsymbol{t}^* \in \operatorname{argmax}_{\|\boldsymbol{t}\|_\infty \leq 1} T(\boldsymbol{v}^*, \boldsymbol{t})$. Because T is convex-concave, we have

$$T(\boldsymbol{v}^*, \boldsymbol{t}) \leq T(\boldsymbol{v}^*, \boldsymbol{t}^*) \leq T(\boldsymbol{v}, \boldsymbol{t}^*),$$

for all $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{t}\|_\infty \leq 1$. Thus, $(\boldsymbol{v}^*, \boldsymbol{t}^*)$ is a saddle-point and by [43, pg. 380]

$$\max_{\|\boldsymbol{t}\|_\infty \leq 1} \min_{\boldsymbol{v} \in \mathbb{R}^p} T(\boldsymbol{v}, \boldsymbol{t}) = \min_{\boldsymbol{v} \in \mathbb{R}^p} \max_{\|\boldsymbol{t}\|_\infty \leq 1} T(\boldsymbol{v}, \boldsymbol{t}),$$

and

$$\boldsymbol{t}^* \in \arg\max_{\|\boldsymbol{t}\|_\infty \leq 1} \mathcal{T}(\boldsymbol{t}).$$

Fix $\epsilon > 0$. Define the events

$$\mathcal{A}_1 := \left\{\widehat{\boldsymbol{t}}^f \in D\right\}, \qquad \mathcal{A}_2 := \left\{\frac{\|\boldsymbol{t}^* - \widehat{\boldsymbol{t}}^f\|_2}{\sqrt{p}} \leq \frac{\epsilon}{2}\right\},$$

$$\mathcal{A}_3 := \left\{\frac{\|\boldsymbol{h}\|_2^2}{n} \leq 2,\ \frac{\|\boldsymbol{g}\|_2^2}{n} \leq \frac{2}{\delta}\right\}, \qquad \mathcal{A}_4 := \left\{|\mathcal{T}(\boldsymbol{t}^*) - L^*| \leq \frac{c_0\lambda_{\min}^2}{16\delta\kappa_{\max}}\epsilon^2\right\}.$$

We claim that on the event $\bigcap_{a=1}^4 \mathcal{A}_a$,

$$\max_{\boldsymbol{t}\in D_\epsilon} \mathcal{T}(\boldsymbol{t}) \le L^* - \frac{c_0\lambda_{\min}^2}{16\delta\kappa_{\max}}\epsilon^2 \,. \tag{64}$$

Indeed, because $\mathcal{A}_1$ occurs, $\boldsymbol{t} \in D_\epsilon$ implies $\frac{\|\boldsymbol{t}-\widehat{\boldsymbol{t}}^f\|_2}{\sqrt{p}} \ge \epsilon$. Because $\mathcal{A}_2$ occurs, also $\frac{\|\boldsymbol{t}-\boldsymbol{t}^*\|_2}{\sqrt{p}} \ge \frac{\epsilon}{2}$. Because $\mathcal{A}_3$ occurs, $\mathcal{T}(\boldsymbol{t})$ is $\frac{c_0\lambda_{\min}^2}{n\kappa_{\max}}$-strongly concave by Lemma B.5, whence because $\boldsymbol{t}^*$ maximizes $\mathcal{T}$

$$\mathcal{T}(\boldsymbol{t}) \le \mathcal{T}(\boldsymbol{t}^*) - \frac{1}{2}\frac{c_0\lambda_{\min}^2}{n\kappa_{\max}}\frac{\epsilon^2}{4}\,.$$

Because $\mathcal{A}_4$ occurs, we conclude Eq. (64).

By Gordon's lemma for the subgradient (Lemma B.4) and because D_ϵ is closed,

$$\mathbb{P}\left(\max_{\boldsymbol{t}\in D_\epsilon} \mathcal{V}(\boldsymbol{t}) \ge L^* - \frac{c_0\lambda_{\min}^2}{16\delta\kappa_{\max}}\epsilon^2\right) \le 2\left(1 - \mathbb{P}\left(\bigcap_{a=1}^4 \mathcal{A}_a\right)\right) \le 2\sum_{a=1}^4 \mathbb{P}(\mathcal{A}_a^c)\,. \tag{65}$$

We control the probabilities in the sum one at a time.

Event $\mathcal{A}_2$ occurs with high probability depending on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ. By Lemma B.3, there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\epsilon \in (0, c')$ we have

$$\mathbb{P}\left(\frac{\|\boldsymbol{v}^* - \widehat{\boldsymbol{v}}^f\|_2}{\sqrt{p}} > \frac{\epsilon}{2}\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}.$$

Indeed, the event in the preceding display occurs when the two conditions in Eq. (54) are met. Also, $\|\widehat{\boldsymbol{v}}^f\|_2^2/n + \sigma^2$, $\boldsymbol{g}^\top\widehat{\boldsymbol{v}}^f/n$, and $\|\boldsymbol{h}\|_2/\sqrt{n}$ concentrate on τ^{*2}, $\tau^*(1-\zeta^*)$, and 1 at sub-Gamma or sub-Gaussian rates depending only on $\tau_{\max}$ (see, e.g., Eq. (57) in the proof of Lemma B.3). Combining this with the previous display and updating constants appropriately, we conclude there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\epsilon \in (0, c')$ we have

$$\mathbb{P}\left(\frac{1}{\sqrt{p}}\left\|\boldsymbol{t}^* - \frac{1}{\lambda}\boldsymbol{\Sigma}^{1/2}\left(\tau^* - \tau^*(1-\zeta^*)\right)\left(\frac{\widehat{\boldsymbol{v}}^f}{\tau^*} - \boldsymbol{g}\right)\right\|_2 > \frac{\epsilon}{2}\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}.$$

By the definition of $\widehat{\boldsymbol{t}}^f$ (Eq. (15)) and of $\widehat{\boldsymbol{v}}$ (Eq. (50)), the preceding display is equivalent to

$$\mathbb{P}\left(\mathcal{A}_2^c\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}\,.$$

Event $\mathcal{A}_3$ occurs with high probability depending on δ. By Gaussian concentration of Lipschitz functions, $\mathbb{P}(\mathcal{A}_3) \le Ce^{-cn}$ for some C, c depending only on δ.

Event $\mathcal{A}_4$ occurs with high probability depending on δ. Observe that $\mathcal{T}(\boldsymbol{t}^*) = \mathcal{L}(\boldsymbol{v}^*)$. Then, by Lemma B.3 there exist constants $C, c, c' > 0$, depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ, such that for $\epsilon \in (0, c')$,

$$\mathbb{P}(\mathcal{A}_4^c) = \mathbb{P}\left(|\mathcal{T}(\boldsymbol{t}^*) - L^*| > \epsilon\right) = \mathbb{P}\left(\left|\max_{\|\boldsymbol{t}\|_\infty \le 1} \mathcal{T}(\boldsymbol{t}) - L^*\right| > \epsilon\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}\,. \tag{66}$$

Combining the established probability bounds on $\mathcal{A}_i$, $i = 2, 3, 4$, Eq. (65) implies that for all $\epsilon < c'$,

$$\mathbb{P}\left(\max_{\boldsymbol{t}\in D_\epsilon} \mathcal{V}(\boldsymbol{t}) \ge L^* - \frac{3}{2}\gamma\epsilon\right) \le 2\mathbb{P}\left(\widehat{\boldsymbol{t}}^f \notin D\right) + \frac{C}{\epsilon^2}e^{-cn\epsilon^4} \text{ and } \mathbb{P}\left(\max_{\|\boldsymbol{t}\|_\infty \le 1} \mathcal{V}(\boldsymbol{t}) < L^* - \gamma\epsilon\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}\,,$$

where the second probability bound holds by Eq. (66). Thus, $\mathbb{P}\left(\widehat{\boldsymbol{t}} \in D_\epsilon\right) \le 2\mathbb{P}\left(\widehat{\boldsymbol{t}}^f \notin D\right) + \frac{C}{\epsilon^2}e^{-cn\epsilon^4}$. Using the definition of D_ϵ and a change of variables (which absorbs certain constants into c), we conclude that Eq. (16) holds.

To complete the proof of Lemma 10, we prove Eq. (17). Define

$$D = \left\{ \boldsymbol{t} \in \mathbb{R}^p \;\middle|\; \left|\phi\Big(\frac{\boldsymbol{t}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{t}}^f}{\sqrt{p}}\Big)\Big]\right| \le \epsilon \right\}.$$

By Eq. (15), $\widehat{\boldsymbol{t}}^f$ is $\frac{\tau_{\max}\zeta_{\max}\kappa_{\max}^{1/2}}{\lambda_{\min}}$-Lipschitz in $\boldsymbol{g}$, whence

$$\mathbb{P}\left(\widehat{\boldsymbol{t}}^f \notin D\right) \le 2\exp\left(-\frac{3\gamma n}{c_0\tau_{\max}\zeta_{\max}}\epsilon^2\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4},$$

where the last inequality holds for $\epsilon < c'$ with $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ. Eq. (17) is then a special case of Eq. (16). The proof of Lemma 10 is complete. □

PROOF OF LEMMA B.4. Fix $R > 0$. The function $\boldsymbol{t} \mapsto \min_{\|\boldsymbol{w}\|_2 \le R} V(\boldsymbol{w}, \boldsymbol{t})$ is concave and continuous and is defined on a compact set D. Moreover, $\min_{\|\boldsymbol{w}\|_2 \le R} V(\boldsymbol{w}, \boldsymbol{t})$ is non-increasing in R. Because the maximum of a non-increasing limit of continuous functions defined on a compact set is equal to the limit of the maxima of these functions,

$$\max_{\boldsymbol{t}\in D} \mathcal{V}(\boldsymbol{t}) = \max_{\boldsymbol{t}\in D} \lim_{R\to\infty} \min_{\|\boldsymbol{w}\|_2 \le R} V(\boldsymbol{w}, \boldsymbol{t}) = \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \min_{\|\boldsymbol{w}\|_2 \le R} V(\boldsymbol{w}, \boldsymbol{t}).$$

We may write

$$V(\boldsymbol{w}, \boldsymbol{t}) := \max_{\|\boldsymbol{u}\|_2 \le R'} \breve{V}(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{u}),$$

for any $R' > \|\boldsymbol{X}\|_{\text{op}}\|\boldsymbol{w}\|_2 + \sigma\|\boldsymbol{z}\|_2$, where

$$\breve{V}(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{u}) = \frac{1}{n}\boldsymbol{u}^\top(\boldsymbol{X}\boldsymbol{w} - \sigma\boldsymbol{z}) - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}\boldsymbol{t}^\top(\boldsymbol{\theta}^* + \boldsymbol{w}) - \frac{\lambda}{n}\|\boldsymbol{\theta}^*\|_1.$$

Because almost surely $R^2 > \|\boldsymbol{X}\|_{\text{op}}R + \sigma\|\boldsymbol{z}\|_2$ for sufficiently large R, we conclude

$$\begin{aligned}\max_{\boldsymbol{t}\in D} \mathcal{V}(\boldsymbol{t}) &= \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \min_{\|\boldsymbol{w}\|_2 \le R} \max_{\|\boldsymbol{u}\|_2 \le R^2} \breve{V}(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{u}) \\ &= \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \max_{\|\boldsymbol{u}\|_2 \le R^2} \min_{\|\boldsymbol{w}\|_2 \le R} \breve{V}(\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{u}), \end{aligned} \tag{67}$$

almost surely, where we may exchange minimization and maximization because they are taken over compact sets and $\breve{V}$ is convex-concave and continuous.

Similarly,

$$\max_{\boldsymbol{t}\in D} \mathcal{T}(\boldsymbol{t}) = \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \min_{\|\boldsymbol{v}\|_2 \le R} T(\boldsymbol{v}, \boldsymbol{t}).$$

We may write

$$T(\boldsymbol{v}, \boldsymbol{t}) = \max_{\|\boldsymbol{u}\|_2 \le R'} \breve{T}(\boldsymbol{v}, \boldsymbol{t}, \boldsymbol{u}),$$

for any $R' > \sqrt{n}\left(\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} + \frac{\|\boldsymbol{g}\|_2\|\boldsymbol{v}\|_2}{n}\right)$, where

$$\breve{T}(\boldsymbol{v}, \boldsymbol{t}, \boldsymbol{u}) := -\frac{1}{n^{3/2}}\|\boldsymbol{u}\|_2\boldsymbol{g}^\top\boldsymbol{v} + \frac{1}{n}\sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\cdot\boldsymbol{h}^\top\boldsymbol{u} - \frac{1}{2n}\|\boldsymbol{u}\|_2^2 + \frac{\lambda}{n}\boldsymbol{t}^\top(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}) - \frac{\lambda}{n}\|\boldsymbol{\theta}^*\|_1.$$

Because almost surely $R^2 > \sqrt{n}\left(\sqrt{\frac{R^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} + \frac{\|\boldsymbol{g}\|R}{n}\right)$ for sufficiently large R, we conclude

$$\begin{aligned}\max_{\boldsymbol{t}\in D} \mathcal{T}(\boldsymbol{t}) &= \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \min_{\|\boldsymbol{v}\|_2\le R} \max_{\|\boldsymbol{u}\|_2\le R^2} \breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u}) \\ &= \lim_{R\to\infty} \max_{\boldsymbol{t}\in D} \max_{\|\boldsymbol{u}\|_2\le R^2} \min_{\|\boldsymbol{v}\|_2\le R} \breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u}), \end{aligned} \tag{68}$$

where the second equality holds by the following argument.[7] For fixed $\boldsymbol{t}, \boldsymbol{u}$, the function $\breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u})$ depends on $\boldsymbol{v}$ only through $\boldsymbol{g}^\top\boldsymbol{v}$, $\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}$, and $\|\boldsymbol{v}\|_2$. Moreover, $\breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u})$ is convex in the triple $(\boldsymbol{g}^\top\boldsymbol{v}, \boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}, \|\boldsymbol{v}\|_2)$ and $\{(\boldsymbol{g}^\top\boldsymbol{v}, \boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}, \|\boldsymbol{v}\|_2) \mid \|\boldsymbol{v}\|_2 \le R\}$ is a compact, convex set. Similarly, for fixed $\boldsymbol{t}, \boldsymbol{v}$, the function $\breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u})$ depends on $\boldsymbol{u}$ only through $\boldsymbol{h}^\top\boldsymbol{u}$, $\|\boldsymbol{u}\|_2$. Moreover, $\breve{T}(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u})$ is convex in the pair $(\boldsymbol{h}^\top\boldsymbol{u}, \|\boldsymbol{u}\|_2)$ and $\{(\boldsymbol{h}^\top\boldsymbol{u}, \|\boldsymbol{u}\|_2) \mid \|\boldsymbol{u}\|_2 \le R^2\}$ is a compact, convex set. Thus, the exchange of minimization and maximization in the preceding display is justified.

By Gordon's Lemma (see [49, Theorem 3]), for any finite $R > 0$ and any $t \in \mathbb{R}$

$$\mathbb{P}\left(\max_{\boldsymbol{t}\in D} \max_{\|\boldsymbol{u}\|_2\le R^2} \min_{\|\boldsymbol{w}\|_2\le R} \breve{V}(\boldsymbol{w},\boldsymbol{t},\boldsymbol{u}) > t\right) \le 2\mathbb{P}\left(\max_{\boldsymbol{t}\in D} \max_{\|\boldsymbol{u}\|_2\le R^2} \min_{\|\boldsymbol{w}\|_2\le R} \breve{T}(\boldsymbol{w},\boldsymbol{t},\boldsymbol{u}) > t\right).$$

Taking $R \to \infty$ and using Eqs. (67) and (68), we conclude

$$\mathbb{P}\left(\max_{\boldsymbol{t}\in D} \mathcal{V}(\boldsymbol{t}) > t\right) \le 2\mathbb{P}\left(\max_{\boldsymbol{t}\in D} \mathcal{T}(\boldsymbol{t}) > t\right).$$

The strict inequalities become weak by considering $t' > t$ in place of t and taking $t' \to t$. We conclude Eq. (62). Eq. (63) follows similarly. $\square$

PROOF OF LEMMA B.5. Define

$$f(\boldsymbol{v}) := \sqrt{\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2}\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} - \frac{\boldsymbol{g}^\top\boldsymbol{v}}{n}.$$

The gradient and Hessian of $f(\boldsymbol{v})$ are

$$\nabla f(\boldsymbol{v}) = \frac{1}{n}\left(\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2\right)^{-1/2} \frac{\|\boldsymbol{h}\|_2}{\sqrt{n}}\boldsymbol{v} - \frac{\boldsymbol{g}}{n},$$

$$\nabla^2 f(\boldsymbol{v}) = \frac{1}{n}\left(\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2\right)^{-1/2} \left(\mathbf{I}_p - \left(\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2\right)^{-1} \frac{\boldsymbol{v}\boldsymbol{v}^\top}{n}\right) \frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} \preceq \frac{1}{n}\left(\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2\right)^{-1/2} \frac{\|\boldsymbol{h}\|_2}{\sqrt{n}}\mathbf{I}_p.$$

We bound

$$\begin{aligned}\|\nabla f(\boldsymbol{v})\|_2^2 &\le \frac{2\|\boldsymbol{h}\|_2^2}{n^2} + \frac{2\|\boldsymbol{g}\|_2^2}{n^2}, \\ |f(\boldsymbol{v})|\,\|\nabla^2 f(\boldsymbol{v})\|_{\mathrm{op}} &\le \frac{\|\boldsymbol{h}\|_2}{n^{3/2}}\left(\frac{\|\boldsymbol{h}\|_2}{\sqrt{n}} + \frac{\|\boldsymbol{g}\|_2\|\boldsymbol{v}\|_2}{n}\left(\frac{\|\boldsymbol{v}\|_2^2}{n} + \sigma^2\right)^{-1/2}\right) \\ &\le \frac{\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{h}\|_2\|\boldsymbol{g}\|_2}{n^2}.\end{aligned}$$

[7] Note that $\breve{T}$ is not convex-concave in $(\boldsymbol{v},\boldsymbol{t},\boldsymbol{u})$, so that the exchange of the minimization and maximization requires a different justification to that in Eq. (67).

The Hessian of $\frac{1}{2}(f(\boldsymbol{x}))_+^2$ is $[\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\top + f(\boldsymbol{x})\nabla^2 f(\boldsymbol{x})]\,\mathbf{1}_{f(\boldsymbol{x})\geq 0}$, whence on the event appearing in the statement of the lemma,

$$\|\nabla^2 \cdot (f(\boldsymbol{v})_+^2/2)\|_{\mathrm{op}} \leq \frac{2\|\boldsymbol{h}\|_2^2}{n^2} + \frac{2\|\boldsymbol{g}\|_2^2}{n^2} + \frac{\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{h}\|_2\|\boldsymbol{g}\|_2}{n^2} \leq \frac{1}{nc_0}\,,$$

where $c_0 = (4 + 4/\delta + 2(1+\delta^{-1/2}))^{-1}$. That is, $\boldsymbol{v} \mapsto \frac{1}{2}f(\boldsymbol{v})_+^2$ is $1/(nc_0)$-strongly smooth. Note that

$$\mathcal{T}(\boldsymbol{t}) = -\tilde{f}^*(-\boldsymbol{t}) - \frac{\lambda}{n}\|\boldsymbol{\theta}^*\|_1\,,$$

where $\tilde{f}(\tilde{\boldsymbol{v}}) := f(\boldsymbol{\Sigma}^{1/2}(n\tilde{\boldsymbol{v}}/\lambda - \boldsymbol{\theta}^*))^2/2$, and $\tilde{f}^*$ is the Fenchel-Legendre conjugate of $\tilde{f}$. Because $f(\boldsymbol{v})_+^2/2$ is $1/(nc_0)$-strongly smooth, $\tilde{f}$ is $\frac{n\kappa_{\max}}{c_0\lambda_{\min}^2}$-strongly smooth. By the duality of strong smoothness and strong convexity [29, Theorem 6], we conclude that $\mathcal{T}(\boldsymbol{v})$ is $\frac{c_0\lambda_{\min}^2}{n\kappa_{\max}}$-strongly concave. □

B.4. *Control of the Lasso sparsity: proof of Theorem 9.* For notational convenience, let us write

$$\bar{\boldsymbol{\Sigma}} := \frac{1}{\kappa_{\min}}\boldsymbol{\Sigma}\,, \qquad \bar{\tau}^* := \sqrt{\frac{1}{\kappa_{\min}}}\tau^*\,, \qquad \bar{\lambda} := \frac{1}{\kappa_{\min}}\lambda\,,$$

so that by Eqs. (4) and (9)

$$\widehat{\boldsymbol{\theta}}^f = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta^*}{2}\|\tau^*\boldsymbol{g} + \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1\right\} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta^*}{2}\|\bar{\tau}^*\boldsymbol{g} + \bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta})\|_2^2 + \bar{\lambda}\|\boldsymbol{\theta}\|_1\right\}. \tag{69}$$

The KKT conditions of this optimization problem are

$$\bar{\boldsymbol{\Sigma}}^{1/2}\left(\bar{\tau}^*\boldsymbol{g} + \bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}^f)\right) \in \frac{\bar{\lambda}}{\zeta^*}\partial\|\widehat{\boldsymbol{\theta}}^f\|_1\,,$$

whence

$$\widehat{\boldsymbol{\theta}}^f = \eta_{\mathrm{soft}}\left(\widehat{\boldsymbol{\theta}}^f + \bar{\boldsymbol{\Sigma}}^{1/2}\left(\bar{\tau}^*\boldsymbol{g} + \bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}^f)\right); \frac{\bar{\lambda}}{\zeta^*}\right) =: \eta_{\mathrm{soft}}\left(\breve{\boldsymbol{y}}^f; \frac{\bar{\lambda}}{\zeta^*}\right)\,, \tag{70}$$

and by Eq. (15)

$$\widehat{\boldsymbol{t}}^f = \frac{\zeta^*}{\bar{\lambda}}\left(\breve{\boldsymbol{y}}^f - \eta_{\mathrm{soft}}\left(\breve{\boldsymbol{y}}^f; \frac{\bar{\lambda}}{\zeta^*}\right)\right)\,, \tag{71}$$

where $\eta_{\mathrm{soft}}(\cdot, \alpha)$ applies $x \mapsto \mathrm{sign}(x)(|x| - \alpha)_+$ coordinates-wise. This representation is useful because the marginals of $\breve{\boldsymbol{y}}^f$ have bounded density, which will allow us to control the expected number of coordinates of $\widehat{\boldsymbol{t}}^f$ which are close to 1.

LEMMA B.6 (Anti-concentration of $\breve{\boldsymbol{y}}^f$). *For each j, the coordinate $\breve{y}_j^f$ has marginal density with respect to Lebesgue measure bounded above by $\frac{\kappa_{\min}^{1/2}\kappa_{\mathrm{cond}}}{\sqrt{2\pi}\tau_{\min}}$.*

PROOF OF LEMMA B.6. We compute

$$\breve{\boldsymbol{y}}^f = \boldsymbol{\theta}^* + \bar{\boldsymbol{\Sigma}}^{1/2}\left(\bar{\tau}^*\boldsymbol{g} + (\mathbf{I}_p - \bar{\boldsymbol{\Sigma}}^{-1})\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}^f)\right) =: \boldsymbol{\theta}^* + \bar{\boldsymbol{\Sigma}}^{1/2}(\bar{\tau}^*\boldsymbol{g} + f(\bar{\tau}^*\boldsymbol{g}))\,. \tag{72}$$

By definition, all eigenvalues of $\bar{\boldsymbol{\Sigma}}$ are bounded below and above by 1 and κ_{cond}, respectively, so that all eigenvalues of $(\mathbf{I}_p - \bar{\boldsymbol{\Sigma}}^{-1})$ are between 0 and $1 - \kappa_{\mathrm{cond}}^{-1}$. Because $\bar{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\theta}}^f$ is 1-Lipschitz in $\bar{\tau}^*\boldsymbol{g}$ (by Eq. (69), using [40, pg. 131]), the function f is $(1 - \kappa_{\mathrm{cond}}^{-1})$-Lipschitz.

Let $\bar{\boldsymbol{\sigma}}_i$ be the i^{th} row of $\bar{\boldsymbol{\Sigma}}^{1/2}$. Let $\mathsf{P}_i^\perp$ be the projection operator onto the orthogonal complement of the span of $\bar{\boldsymbol{\sigma}}_i$. Then

$$\breve{y}_i^f = \theta_i^* + \bar{\tau}^* \bar{\boldsymbol{\sigma}}_i^\top \boldsymbol{g} + \bar{\boldsymbol{\sigma}}_i^\top f\left(\bar{\tau}^*(\bar{\boldsymbol{\sigma}}_i^\top \boldsymbol{g})\bar{\boldsymbol{\sigma}}_i/\|\bar{\boldsymbol{\sigma}}_i\|_2^2 + \bar{\tau}^* \mathsf{P}_i^\perp \boldsymbol{g}\right).$$

Consider the function

$$h(x) := \bar{\tau}^* x + \bar{\boldsymbol{\sigma}}_i^\top f\left(\bar{\tau}^* x \bar{\boldsymbol{\sigma}}_i/\|\bar{\boldsymbol{\sigma}}_i\|_2^2 + \bar{\tau}^* \mathsf{P}_i^\perp \boldsymbol{g}\right).$$

Since f is $(1-\kappa_{\text{cond}}^{-1})$-Lipschitz, for any $x_1 < x_2$, $x_1, x_2 \in \mathbb{R}$, we have

$$h(x_2) - h(x_1) \geq \bar{\tau}^* \kappa_{\text{cond}}^{-1}(x_2 - x_1). \tag{73}$$

Because $\bar{\boldsymbol{\sigma}}_i^\top \boldsymbol{g} \sim \mathsf{N}(0, \|\bar{\boldsymbol{\sigma}}_i\|_2^2)$, its density is upper bounded by $(2\pi\|\bar{\boldsymbol{\sigma}}_i\|_2^2)^{-1/2}$. Further, it is independent of $\mathsf{P}_i^\perp \boldsymbol{g}$. Thus, the lower bound (73) implies that $\breve{y}_i^f$ has density $q(y)$ upper bounded by

$$\sup_y q(y) \leq \frac{1}{\sqrt{2\pi}\|\bar{\boldsymbol{\sigma}}_i\|_2} \cdot \frac{1}{\inf_y h'(y)} \leq \frac{\kappa_{\min}^{1/2}\kappa_{\text{cond}}}{\sqrt{2\pi}\tau_{\min}},$$

where we have used that $\|\bar{\boldsymbol{\sigma}}_i\|_2$ is no smaller than the minimal singular value of $\bar{\boldsymbol{\Sigma}}^{1/2}$ which is no smaller than 1 by construction, and that $\bar{\tau}^* = \tau^*/\kappa_{\min}^{1/2} \geq \tau_{\min}/\kappa_{\min}^{1/2}$. $\square$

We are now ready to complete the proof of Theorem 9. We prove high-probability upper and lower bounds on the sparsity separately. The arguments are almost identical, but establishing the upper bound involves analyzing the subgradient $\widehat{\boldsymbol{t}}$ and establishing the lower bound involves analyzing $\widehat{\boldsymbol{\theta}}$.

Upper bound on sparsity via the subgradient: The lasso sparsity is upper bounded in terms of the lasso subgradient:

$$\frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n} \leq \frac{|\{j \in [p] : |\widehat{t}_j| = 1\}|}{n}. \tag{74}$$

We prove a high-probability upper bound on the right-hand side. Given any $\Delta \geq -1$, define $T(\breve{\boldsymbol{y}}, \Delta) := \{j \subset [p] \mid |\breve{y}_j| \geq \bar{\lambda}(1+\Delta)/\zeta^*\}$. We will control quantity $T(\breve{\boldsymbol{y}}, -\Delta)$ for $\Delta \geq 1$.. Consider the function

$$\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}, \Delta) := \frac{1}{n}\sum_{j=1}^p \phi_1^{\mathsf{ub}}(\breve{y}_j, \Delta) \quad \text{where} \quad \phi_1^{\mathsf{ub}}(\breve{y}, \Delta) := \min(1, \zeta^*|\breve{y}|/(\bar{\lambda}\Delta) - 1/\Delta + 2)_+ .$$

The function ϕ_1^{ub} is 0 on $[-\bar{\lambda}(1-2\Delta)/\zeta^*, \bar{\lambda}(1-2\Delta)/\zeta^*]$, 1 on $[-\bar{\lambda}(1-\Delta)/\zeta^*, \bar{\lambda}(1-\Delta)/\zeta^*]^c$, and linearly interpolates between the function values on these sets everywhere else. Unlike $\breve{\boldsymbol{y}} \mapsto T(\breve{\boldsymbol{y}}, -\Delta)$, the function $\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}, \Delta)$ is $\frac{\zeta^*}{\bar{\lambda}\Delta\delta^{1/2}\sqrt{n}}$-Lipschitz in $\breve{\boldsymbol{y}}$. For all $\breve{\boldsymbol{y}}$, by definition we have

$$\frac{|T(\breve{\boldsymbol{y}}, -\Delta)|}{n} \leq \phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}, \Delta).$$

(The preceding display justifies the superscript ub, which stands for "upper bound".) Moreover, by Eq. (70)

$$\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}^f, \Delta) \leq \frac{\|\widehat{\boldsymbol{\theta}}^f\|_0}{n} + \frac{|\{j \in [p] \mid 1 - \zeta^*|\breve{y}_j^f|/\bar{\lambda} \in [0, 2\Delta]\}|}{n},$$

whence

$$\mathbb{E}[\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}^f,\Delta)] \le 1-\zeta^* + \frac{4\bar{\lambda}\Delta}{\delta\zeta^*}\frac{\kappa_{\min}^{1/2}\kappa_{\mathrm{cond}}}{\sqrt{2\pi}\tau_{\min}} \le 1-\zeta^* + \frac{4\lambda_{\max}\kappa_{\mathrm{cond}}\Delta}{\delta\tau_{\min}\zeta_{\min}\sqrt{2\pi\kappa_{\min}}}\,,$$

where we have applied Lemma B.6. By the definition of $\breve{\boldsymbol{y}}^f$ in Eq. (72) and recalling that $\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f$ is τ^*-Lipschitz in $\boldsymbol{g}$, we have that $\boldsymbol{g} \mapsto \breve{\boldsymbol{y}}$ is $\kappa_{\mathrm{cond}}^{1/2}\bar{\tau}^* + \kappa_{\mathrm{cond}}^{1/2}\tau^*/\kappa_{\min}^{1/2} = 2\kappa_{\mathrm{cond}}^{1/2}\tau_{\max}/\kappa_{\min}^{1/2}$-Lipschitz in $\boldsymbol{g}$. By Gaussian concentration of Lipschitz functions,

$$\begin{aligned}\mathbb{P}\left(\frac{|T(\breve{\boldsymbol{y}}^f,-\Delta)|}{n} \ge 1-\zeta^* + \frac{4\lambda_{\max}\kappa_{\mathrm{cond}}\Delta}{\delta\tau_{\min}\zeta_{\min}\sqrt{2\pi\kappa_{\min}}} + \epsilon\right) &\le \mathbb{P}\Big(\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}^f,\Delta) \ge \mathbb{E}[\phi^{\mathsf{ub}}(\breve{\boldsymbol{y}}^f,\Delta)] + \epsilon\Big)\\ &\le \exp\left(-\frac{n\delta\bar{\lambda}^2}{2\zeta^{*2}}\Delta^2\cdot\frac{\kappa_{\min}}{4\kappa_{\mathrm{cond}}\tau_{\max}^2}\epsilon^2\right)\\ &\le \exp\left(-\frac{n\delta\lambda_{\min}^2}{8\kappa_{\min}\zeta_{\max}^2\kappa_{\mathrm{cond}}\tau_{\max}^2}\Delta^2\epsilon^2\right)\,.\end{aligned}$$

Plugging in $\epsilon = \Delta$ and absorbing constants appropriately, we conclude there exists $c_1, c_2 > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\Delta \ge 0$

$$\mathbb{P}\left(\frac{|T(\breve{\boldsymbol{y}}^f,-\Delta)|}{n} \ge 1-\zeta^* + c_1\Delta\right) \le \exp\left(-c_2 n\Delta^4\right)\,.$$

By Eq. (71), if $\frac{|T(\breve{\boldsymbol{y}}^f,-\Delta)|}{n} < 1-\zeta^* + c_1\Delta$, then for all $\boldsymbol{t} \in \mathbb{R}^p$ with $|\{j \in [p] \mid |t_j| \ge 1\}|/n \ge 1-\zeta^* + 2c_1\Delta$,

$$\frac{\|\widehat{\boldsymbol{t}}^f - \boldsymbol{t}\|_2^2}{n} \ge c_1\Delta^3\,,$$

because there are at least $c_1\Delta n$ coordinates where $\widehat{\boldsymbol{t}}^f$ and $\boldsymbol{t}$ differ by at least Δ. Absorbing constants and taking $D = \{\boldsymbol{t} \in \mathbb{R}^p \mid |\{j \in [p] \mid 1 - |t_j| \le \Delta\}|/n \le 1-\zeta^* + c_1\Delta\}$ in Lemma 10, there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\Delta < c'$

$$\mathbb{P}\left(\frac{|\{j \in [p] \mid |\widehat{t}_j| \ge 1\}|}{n} \ge 1-\zeta^* + 2\Delta\right) \le 2\exp\Big(-cn\Delta^4\Big) + \frac{C}{\Delta^3}\exp\left(-cn\Delta^6\right)\,.$$

We may absorb the first term into the second at the cost of changing the constants C, c, c' because the bound applies only to $\Delta < c'$. By Eq. (74), $\mathbb{P}(\|\widehat{\boldsymbol{\theta}}\|_0/n > 1-\zeta^* + \Delta) \le \frac{C}{\Delta^3}e^{-cn\Delta^6}$. A high probability upper bound on the sparsity of the lasso solution has been established.

Lower bound on sparsity via the lasso estimate: Define

$$\phi^{\mathsf{lb}}(\breve{\boldsymbol{y}},\Delta) := \frac{1}{n}\sum_{j=1}^p \phi_1^{\mathsf{lb}}(\breve{y}_j,\Delta) \quad\text{where}\quad \phi_1^{\mathsf{lb}}(\breve{y},\Delta) := \min(1, \zeta^*|\breve{y}|/(\bar{\lambda}\Delta) - 1/\Delta - 1)_+\,.$$

The function ϕ_1^{lb} is 0 on $[-\bar{\lambda}(1+\Delta)/\zeta^*, \bar{\lambda}(1+\Delta)/\zeta^*]$, 1 on $[-\bar{\lambda}(1+2\Delta)/\zeta^*, \bar{\lambda}(1+2\Delta)/\zeta^*]^c$, and linearly interpolates between the function values on these sets everywhere else. The function ϕ^{lb} is a $\frac{\zeta^*}{\bar{\lambda}\Delta\delta^{1/2}\sqrt{n}}$-Lipschitz lower bound for $|T(\breve{\boldsymbol{y}},\Delta)|/n$:

$$\frac{|T(\breve{\boldsymbol{y}},\Delta)|}{n} \ge \phi^{\mathsf{lb}}(\breve{\boldsymbol{y}},\Delta)\,.$$

Moreover, by Eq. (70)

$$\phi^{\mathsf{lb}}(\breve{\boldsymbol{y}}^f,\Delta) \ge \frac{\|\widehat{\boldsymbol{\theta}}^f\|_0}{n} - \frac{|\{j \in [p] \mid \zeta^*|\breve{y}_j^f|/\bar{\lambda} - 1 \in [0, 2\Delta]\}|}{n}\,,$$

whence

$$\mathbb{E}[\phi^{\mathsf{lb}}(\check{\boldsymbol{y}}^f,\Delta)] \ge 1-\zeta^* - \frac{4\bar{\lambda}\Delta}{\delta\zeta^*}\frac{\kappa_{\min}^{1/2}\kappa_{\mathrm{cond}}}{\sqrt{2\pi}\tau_{\min}} \le 1-\zeta^* - \frac{4\lambda_{\max}\kappa_{\mathrm{cond}}\Delta}{\delta\tau_{\min}\zeta_{\min}\sqrt{2\pi\kappa_{\min}}}\,,$$

where we have applied Lemma B.6. Following the same argument used to establish the upper bound, we conclude there exists $c_1, c_2 > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\Delta \ge 0$

$$\mathbb{P}\left(\frac{|T(\check{\boldsymbol{y}}^f,\Delta)|}{n} \le 1-\zeta^* - c_1\Delta\right) \le \exp\left(-c_2 n\Delta^4\right)\,. \tag{75}$$

By Eq. (70), if $\frac{|T(\check{\boldsymbol{y}}^f,\Delta)|}{n} > 1-\zeta^*-c_1\Delta$, then $\frac{|\{j\in[p] \mid |\widehat{\theta}_j^f|\ge\Delta\}|}{n} > 1-\zeta^*-c_1\Delta$. Then for all $\boldsymbol{\theta}\in\mathbb{R}^p$ with $\|\boldsymbol{\theta}\|_0/n \le 1-\zeta^*-2c_1\Delta$,

$$\frac{\|\widehat{\boldsymbol{\theta}}^f-\boldsymbol{\theta}\|_2^2}{n} \ge c_1\Delta^3\,,$$

because there are at least $c_1\Delta n$ coordinates where $\widehat{\boldsymbol{\theta}}^f$ and $\boldsymbol{\theta}$ differ by at least Δ. In particular, taking $D := \{\boldsymbol{\theta}\in\mathbb{R}^p \mid \frac{|\{j\in[p] \mid |\widehat{\theta}_j^f|\ge\Delta\}|}{n} > 1-\zeta^*-c_1\Delta\}$ and $D_\epsilon := \{\boldsymbol{x}\in\mathbb{R}^p \mid \inf_{\boldsymbol{x}'\in D}\|\boldsymbol{x}-\boldsymbol{x}'\|_2/\sqrt{p}\ge\epsilon\}$, we have that $\{\boldsymbol{\theta}\in\mathbb{R}^p \mid \|\boldsymbol{\theta}\|_0/n \le 1-\zeta^*-2c_1\Delta\}\subset D_{\epsilon/2}$ for $\epsilon/2=\sqrt{c_1\delta\Delta^3}$. Eq. (75) says $\mathbb{P}(\widehat{\boldsymbol{\theta}}^f\notin D)\le e^{-c_2n\Delta^4}$. Thus, by the proof of Theorem B.1 in Appendix B.1 — in particular, Eq. (55)— we conclude there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\Delta < c'$

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n} \le 1-\zeta^*-\Delta\right) \le \exp\Big(-cn\Delta^4\Big) + \frac{C}{\Delta^3}\exp\left(-cn\Delta^6\right)\,.$$

We may absorb the first term into the second at the cost of changing the constants C, c, c' because the bound applies only to $\Delta < c'$. (In applying Eq. (55), recall $\widehat{\boldsymbol{v}}^f = \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}^f-\theta)$, with the definition of D modified according to this change of variables). A high probability lower bound on the sparsity of the lasso solution has been established.

Theorem 9 follows by putting together the upper and lower bounds.

B.5. *Control of the debiased Lasso: proofs of Theorem 11 and Corollary 12.* We control the debiased Lasso by approximating it with the debiased α-smoothed Lasso, which turns out to be easier to study due to the Lipschitz differentiability of the M_α (cf. (35)). Define the debiased α-smoothed Lasso

$$\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}} := \widehat{\boldsymbol{\theta}}_\alpha + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}_\alpha)}{\zeta_\alpha^*}\,.$$

This definition is analogous to (6) except that $1-\|\widehat{\boldsymbol{\theta}}\|_0/n$ is replaced by the constant ζ_α^*. Because it is not feasible to calculate ζ_α^* exactly without knowing $\boldsymbol{\theta}^*$, therefore $\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}$ cannot be computed either. Rather, $\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}$ is a theoretical tool.

To establish Theorem 11, we first characterize the behavior of $\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}$ and second show that $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ is close to $\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}$ with high-probability. The next lemma characterizes $\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}$.

LEMMA B.7 (Characterization of the debiased α-smoothed Lasso). *Let $\alpha > 0$. Under assumptions A1 and A2$_\alpha$, there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for any 1-Lipschitz $\phi:\mathbb{R}^p\to\mathbb{R}$, we have for all $\epsilon < c'$*

$$\mathbb{P}\left(\left|\phi\Big(\frac{\widehat{\boldsymbol{\theta}}_\alpha^{\mathrm{d}}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^*+\tau_\alpha^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}\Big)\Big]\right| > \Big(1+\frac{\lambda_{\max}}{\kappa_{\min}^{1/2}\zeta_{\min}\alpha}\Big)\epsilon\right) \le \frac{C}{\epsilon^2}e^{-cn\epsilon^4}\,,$$

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

We leave the proof of Lemma B.7 at the end of this section.

The following lemma will allow us to show that $\widehat{\boldsymbol{\theta}}^{\rm d}$ and $\widehat{\boldsymbol{\theta}}^{\rm d}_\alpha$ are close with high probability.

LEMMA B.8 (Closeness of the Lasso and α-smoothed Lasso). *There exists* $C_1, C, c, \alpha_{\max} > 0$ *depending only on* $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, *and* δ *such that*

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\theta}}_\alpha - \widehat{\boldsymbol{\theta}}\|_2}{\sqrt{p}} \le C_1\sqrt{\alpha}, \ \textit{for all } \alpha \le \alpha_{\max}\right) \ge 1 - Ce^{-cn}.$$

We prove Lemma B.8 at the end of this section. Equipped with these two lemma, we are now ready to prove Theorem 11.

B.5.1. *Proof of Theorem 11: characterization of the debiased Lasso.* For any $\alpha > 0$, direct calculations give (setting $\zeta^* = \zeta^*_0$)

(76)
$$\begin{aligned}
\frac{\|\widehat{\boldsymbol{\theta}}^{\rm d} - \widehat{\boldsymbol{\theta}}^{\rm d}_\alpha\|_2}{\sqrt{p}} &\le \frac{1}{\sqrt{p}}\Big\|\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})\Big(\frac{1}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n} - \frac{1}{\zeta^*_\alpha}\Big)\Big\|_2 + \frac{1}{\sqrt{p}}\Big\|(\mathbf{I}_p - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top\boldsymbol{X}/\zeta^*_\alpha)(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_\alpha)\Big\|_2 \\
&\le \kappa_{\min}^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\|_{\rm op}\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{p}}\left(\left|\frac{1}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n} - \frac{1}{\zeta^*}\right| + \left|\frac{1}{\zeta^*} - \frac{1}{\zeta^*_\alpha}\right|\right) \\
&\quad + \left(1 + \frac{\kappa_{\min}^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\|_{\rm op}\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\rm op}\|\boldsymbol{\Sigma}^{1/2}\|_{\rm op}}{\zeta^*_\alpha}\right)\frac{\|\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_\alpha\|_2}{\sqrt{p}} \\
&=: T_1 + T_2.
\end{aligned}$$

Bounding T_1. We start with bounding the T_1 term in Eq. (76). By [53, Corollary 5.35] and Theorem 8, there exist $C_1, C, c > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that with probability at least $1 - Ce^{-cn}$

$$\kappa_{\min}^{1/2}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\|_{\rm op}\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{p}} \le C_1. \tag{77}$$

Let L_τ and L_ζ be as in Lemma A.5, and let $\alpha_{\max}$ be the minimum of the corresponding quantities in Lemma A.5 and Lemma B.8. Let $\alpha'_{\max} = \min\{\alpha_{\max}, \tau_{\min}^2/(4L_\tau^2), \zeta_{\min}^2/(4L_\zeta^2)\}$. By Lemma A.5, for all $\alpha < \alpha'_{\max}$, one has

$$\tau_{\min}/2 \le \tau^*_\alpha \le \tau_{\max} + \tau_{\min}/2, \qquad \zeta_{\min}/2 \le \zeta^*_\alpha \le \zeta_{\max} + \zeta_{\min}/2.$$

For $\alpha \le \alpha'_{\max}$, by Lemma A.5

$$|1/\zeta^* - 1/\zeta^*_\alpha| \le (4/\zeta_{\min}^2)L_\zeta\sqrt{\alpha}. \tag{78}$$

By Theorem 9, there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that for $\alpha < c'$, with probability $1 - \frac{C}{\alpha^{3/2}}e^{-cn\alpha^3}$

$$|1/(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n) - 1/\zeta^*| \le (4/\zeta_{\min}^2)\sqrt{\alpha}. \tag{79}$$

Combining the Eqs. (77), (78), and (79), we conclude there exists $C_1, C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, and δ such that for $\alpha < c'$, with probability $1 - \frac{C}{\alpha^{3/2}}e^{-cn\alpha^3}$ the first term on the right-hand side of Eq. (76) is bounded by $C_1\sqrt{\alpha}$.

Bounding T_2. We now bound the T_2 term of Eq. (76). Assumption $\mathsf{A2}_\alpha$ is satisfied with $\mathcal{P}'_{\mathrm{fixPt}} = (\tau_{\min}/2, \tau_{\max} + \tau_{\min}/2, \zeta_{\min}/2, \zeta_{\max} + \zeta_{\min}/2)$ in place of $\mathcal{P}_{\mathrm{fixPt}}$ for all $\alpha < \alpha'_{\max}$. Because $\zeta^*_\alpha \geq \zeta_{\min}/2$, by [53, Corollary 5.35] there exist $C_2, C, c > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that with probability at least $1 - Ce^{-cn}$

$$\left(1 + \frac{\kappa_{\min}^{-1/2} \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\|_{\mathrm{op}} \|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}} \|\boldsymbol{\Sigma}^{1/2}\|_{\mathrm{op}}}{\zeta^*_\alpha}\right) \leq C_2 \,.$$

Combining this bound with Lemma B.8, absorbing parameters into constants, and absorbing smaller terms into larger ones, we conclude there exists $C_1, C, c > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\alpha < \alpha'_{\max}$, with probability $1 - Ce^{-cn}$ the second term on the right-hand side of Eq. (76) is bounded by $C_1\sqrt{\alpha}$.

Combining the high-probability upper bounds on the terms on the right-hand side of Eq. (76), we conclude there exists $C_1, C, c, \alpha_{\max} > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\alpha < \alpha_{\max}$,

$$\mathbb{P}\left(\frac{|\phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}}) - \phi(\widehat{\boldsymbol{\theta}}^{\mathrm{d}}_\alpha)|}{\sqrt{p}} > C_1\sqrt{\alpha}\right) \leq \mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\theta}}^{\mathrm{d}} - \widehat{\boldsymbol{\theta}}^{\mathrm{d}}_\alpha\|_2}{\sqrt{p}} > C_1\sqrt{\alpha}\right) \leq \frac{C}{\alpha^{3/2}} e^{-cn\alpha^3} \,. \tag{80}$$

Further, for $\alpha < \alpha_{\max}$, by Lemma A.5,

$$\left|\mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}\Big)\Big] - \mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^* + \tau^*_\alpha\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}\Big)\Big]\right| \leq C_1\sqrt{\alpha} \,. \tag{81}$$

Taking $\epsilon = \alpha^3$ in Lemma B.7 (and assuming $\alpha^3 < c'$ for c' in that lemma),

$$\mathbb{P}\left(\left|\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}_\alpha}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^* + \tau^*_\alpha\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}\Big)\Big]\right| > C_1\alpha^2\right) \leq \frac{C}{\alpha^6} e^{-cn\alpha^{12}} \,. \tag{82}$$

Combining Eqs. (80), (81), and (82) and appropriately adjusting constants, we conclude there exists $C, C', c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\epsilon < c'$

$$\mathbb{P}\left(\left|\phi\Big(\frac{\widehat{\boldsymbol{\theta}}^{\mathrm{d}}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}}{\sqrt{p}}\Big)\Big]\right| > C_1\epsilon\right) \leq \frac{C}{\epsilon^3} e^{-cn\epsilon^6} \,.$$

We complete the proof of Theorem 11.

B.5.2. *Proof of Lemma B.7: characterization of the debiased α-smoothed Lasso.* By the KKT conditions for the optimization defining the α-smoothed Lasso (cf. (37)), $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}_\alpha = \widehat{\boldsymbol{\theta}}_\alpha + \frac{\lambda\boldsymbol{\Sigma}^{-1/2}\nabla\mathsf{M}_\alpha(\widehat{\boldsymbol{\theta}}_\alpha)}{\zeta^*_\alpha}$. Since $\boldsymbol{\theta} \mapsto \nabla\mathsf{M}_\alpha(\boldsymbol{\theta})$ is $1/\alpha$-Lipschitz, $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}_\alpha$ is a $\left(1 + \frac{\lambda_{\max}}{\kappa_{\min}^{1/2}\zeta_{\min}\alpha}\right)$-Lipschitz function of $\widehat{\boldsymbol{\theta}}_\alpha$. The function

$$\tilde{\phi}\left(\frac{\boldsymbol{\theta}}{\sqrt{p}}\right) := \left(1 + \frac{\lambda_{\max}}{\kappa_{\min}^{1/2}\zeta_{\min}\alpha}\right)^{-1} \phi\left(\frac{\boldsymbol{\theta} + \lambda\boldsymbol{\Sigma}^{-1}\nabla\mathsf{M}_\alpha(\boldsymbol{\theta})/\zeta^*_\alpha}{\sqrt{p}}\right) ,$$

is 1-Lipschitz. Moreover, by the KKT conditions for the optimization defining the α-smoothed Lasso in the fixed design model (Eq. (38)),

$$\widehat{\boldsymbol{\theta}}^f_\alpha + \frac{\lambda\boldsymbol{\Sigma}^{-1}\nabla\mathsf{M}_\alpha(\widehat{\boldsymbol{\theta}}^f_\alpha)}{\zeta^*_\alpha} = \widehat{\boldsymbol{\theta}}^f_\alpha + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{1/2}(\tau^*_\alpha\boldsymbol{g} - \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}^f_\alpha - \boldsymbol{\theta}^*)) = \boldsymbol{\theta}^* + \tau^*_\alpha\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g} \,.$$

Because $\mathcal{R}_\alpha(\widehat{\boldsymbol{\theta}}_\alpha) \leq \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{R}_\alpha(\boldsymbol{\theta}) + \gamma\epsilon^2$ for any $\gamma, \epsilon > 0$, Theorem B.1 and the previous two displays imply the result.

B.5.3. *Proof of Lemma B.8: closeness of the Lasso and the α-smoothed Lasso.* The proof of Lemma B.8 relies on showing that with high-probability, the Lasso objective is strongly convex locally around its minimizer. We then show that because the value of the α-smoothed Lasso objective is close to that of the Lasso objective pointwise, the minimizers of the two objectives must also be close.

LEMMA B.9 (Local strong convexity of Lasso objective). *Assume $n\zeta^*/8 \geq 1$. Then there exists $C, c, c', c_1 > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that with probability at least $1 - Ce^{-cn}$ the following occurs: for all $\boldsymbol{\theta} \in \mathbb{R}^p$ with $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2/\sqrt{p} \leq c'$,*

$$\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \frac{c_1}{p}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 .$$

Let us take Lemma B.9 momentarily and provide its proof below. By Eq. (36), $\mathcal{R}(\boldsymbol{\theta}) \geq \mathcal{R}_\alpha(\boldsymbol{\theta}) \geq \mathcal{R}(\boldsymbol{\theta}) - \frac{\lambda\alpha}{2\delta}$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$. On the event of Lemma B.9, for $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2/\sqrt{p} \leq c'$

$$\mathcal{R}_\alpha(\boldsymbol{\theta}) \geq \mathcal{R}(\boldsymbol{\theta}) - \frac{\lambda\alpha}{2\delta} \geq \mathcal{R}(\widehat{\boldsymbol{\theta}}) + \frac{c_1}{p}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 - \frac{\lambda\alpha}{2\delta} \geq \mathcal{R}_\alpha(\widehat{\boldsymbol{\theta}}) + \frac{c_1}{p}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 - \frac{\lambda\alpha}{2\delta} .$$

Since $\widehat{\boldsymbol{\theta}}_\alpha$ minimizes $\mathcal{R}_\alpha(\boldsymbol{\theta})$, we conclude that for $\alpha \leq \alpha_{\max} := \frac{2c_1 c'^2\delta}{\lambda_{\max}}$

$$\frac{\|\widehat{\boldsymbol{\theta}}_\alpha - \widehat{\boldsymbol{\theta}}\|_2}{\sqrt{p}} \leq \sqrt{\frac{\lambda_{\max}\alpha}{2c_1\delta}} .$$

The proof of Lemma B.8 is complete.

B.5.4. *Proof of Lemma B.9: local strong convexity of the Lasso objective.* We make the observation that with high probability, the Lasso subgradient $\widehat{\boldsymbol{t}} = \frac{1}{\lambda}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})$ (cf. (14)), cannot have too many coordinates with magnitude close to 1, even off of the Lasso support. The next lemma makes this precise.

LEMMA B.1. *There exists $C, c, \Delta > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ, such that*

$$\mathbb{P}\left(\frac{|\{j \in [p] \mid |\widehat{t}_j| \geq 1 - \Delta/2\}|}{n} \geq 1 - \zeta^*/2\right) \leq Ce^{-cn} . \tag{83}$$

PROOF OF LEMMA B.1. The proof is as for Theorem 9 with the following minor changes. We apply Eq. (75) with $\Delta = \zeta^*/(4c_1)$ with c_1 as in in that equation. By Eq. (71) and with this choice of Δ, if $\frac{|T(\breve{\boldsymbol{y}}^f, \Delta)|}{n} < 1 - 3\zeta^*/4$, then for all $\boldsymbol{t} \in \mathbb{R}^p$ with $|\{j \in [p] \mid |t_j| \geq 1 - \Delta/2\}|/n \geq 1 - \zeta^*/2$,

$$\frac{\|\widehat{\boldsymbol{t}}^f - \boldsymbol{t}\|_2^2}{n} \geq \frac{\Delta^2\zeta^*}{16} = \frac{\zeta^{*4}}{256c_1^2} ,$$

because there are at least $\zeta^* n/4$ coordinates where $\widehat{\boldsymbol{t}}^f$ and $\boldsymbol{t}$ differ by at least $\Delta/2$. Absorbing constants and taking $D = \{\boldsymbol{t} \in \mathbb{R}^p \mid |\{j \in [p] \mid 1 - |t_j| \leq \Delta\}|/n \leq 1 - 3\zeta^*/4\}$ in Lemma 10, there exists $C, c > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that Eq. (83) holds. □

We are now ready to prove Lemma B.9. Define the minimum singular value of $\boldsymbol{X}$ over a set $S \subset [p]$ by

$$\kappa_-(\boldsymbol{X}, S) = \inf\left\{\|\boldsymbol{X}\boldsymbol{w}\|_2 \mid \text{supp}(\boldsymbol{w}) \subset S,\, \|\boldsymbol{w}\|_2 = 1\right\} ,$$

and the s sparse singular value by

$$\kappa_-(\boldsymbol{X}, s) = \min_{|S| \le s} \kappa_-(\boldsymbol{X}, S).$$

Consider the event

$$\mathcal{A} := \Big\{\kappa_-\big(\boldsymbol{X}, n(1-\zeta^*/4)\big) \ge \kappa'_{\min}\Big\} \cap \Big\{\|\boldsymbol{X}\|_{\mathrm{op}} \le C\Big\} \cap \Big\{\frac{|\{j \in [p] \mid |\widehat{t}_j| \ge 1 - \Delta/2\}|}{n} \le 1 - \zeta^*/2\Big\}.$$

(Note that we need not assume that $(1-\zeta^*/2)n \le p$ or $(1-\zeta^*/4)n \le p$ for these definitions or events to make sense).

We aim to show that there exist $\kappa'_{\min}, \Delta, C, c > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that

$$\mathbb{P}(\mathcal{A}) \ge 1 - Ce^{-cn}. \tag{84}$$

The second event in the definition of $\mathcal{A}$ is controlled by [53, Corollary 5.35] and the third event by Lemma B.1. Now it is sufficient to consider the first event in the definition of $\mathcal{A}$.

Case $\delta > 1$. For $\delta > 1$, we have $\mathbb{P}(\kappa_-\big(\boldsymbol{X}, n(1-\zeta^*/4)\big) \ge \kappa'_{\min}) \ge 1 - Ce^{-cn}$ because $\kappa_-(\boldsymbol{X}, n(1-\zeta^*/4)) \ge \kappa_-(\boldsymbol{X}, p)$ is the minimum singular value of $\boldsymbol{X}$, whence we invoke [53, Corollary 5.35].

Case $\delta \le 1$. Consider now $\delta \le 1$. Let $k = \lfloor n(1-\zeta^*/4) \rfloor$ and note that $k < p$ because $n \le p$. Because $\kappa_-(\boldsymbol{X}, S') \ge \kappa_-(\boldsymbol{X}, S)$ when $S' \subset S$, we have that $\kappa_-(\boldsymbol{X}, n(1-\zeta^*/4)) = \min_{|S|=k} \kappa_-(\boldsymbol{x}, S)$. By a union bound, for any $t > 0$

$$\mathbb{P}\left(\kappa_-(\boldsymbol{X}, n(1-\zeta^*/4)) \le t\right) \le \sum_{|S|=k} \mathbb{P}(\kappa_-(\boldsymbol{X}_S) \le t). \tag{85}$$

The matrix $\boldsymbol{X}_S = \widetilde{\boldsymbol{X}}_S \boldsymbol{\Sigma}_{S,S}^{1/2}$ where $\widetilde{\boldsymbol{X}}_S$ has entries distributed i.i.d. $\mathsf{N}(0, 1/n)$. Thus, one has

$$\kappa_-(\boldsymbol{X}_S) \ge \kappa_-(\widetilde{\boldsymbol{X}}_S)\kappa_-(\boldsymbol{\Sigma}_{S,S}^{1/2}) \ge \kappa_-(\widetilde{\boldsymbol{X}}_S)\kappa_{\min}^{1/2}.$$

Invoking the fact that $\widetilde{\boldsymbol{X}}_S$ has the same distribution for all $|S| = k$, expression (85) implies

$$\mathbb{P}\left(\kappa_-(\boldsymbol{X}, n(1-\zeta^*/4)) \le t\right) \le \binom{p}{k} \mathbb{P}(\kappa_-(\widetilde{\boldsymbol{X}}_S) \le t/\kappa_{\min}^{1/2}),$$

where the S appearing on the right-hand side can be any S with cardinality k. By Lemma 2.9 of [12],

$$\mathbb{P}(\kappa_-(\widetilde{\boldsymbol{X}}_S) \le t/\kappa_{\min}^{1/2}) \le C(n, t/\kappa_{\min}^{1/2}) \exp\left(n\psi(k/n, t/\kappa_{\min}^{1/2})\right),$$

where $C(a,b)$ is a universal polynomial in a, b and $\psi(a,b) := \frac{1}{2}[(1-a)\log b + 1 - a + a\log a - b]$. (Lemma 2.9 of [12] states a bound on the density of $\kappa_-(\widetilde{\boldsymbol{X}}_S)$, but a deviation bound incurs only a factor $t/\kappa_{\min}^{1/2}$ which we may absorb into the polynomial term). Note also that $\binom{p}{k} \le C'(p)\exp(nH(k/p)/\delta)$, where C' is a universal polynomial. We conclude that

$$\mathbb{P}\left(\kappa_-(\boldsymbol{X}, n(1-\zeta^*/4)) \le t\right) \le C(n, p, t/\kappa_{\min}^{1/2}) \exp\left(n(H(k/p)/\delta + \psi(k/n, t/\kappa_{\min}^{1/2}))\right).$$

Note that $\psi(a,b) \le \frac{\zeta^*}{8}\log b$ for all $a = k/n \le 1 - \zeta^*/4$ and $b \in (0,1)$. Thus, there exists $c, \kappa'_{\min} > 0$, depending only on $\delta, \kappa_{\min}, \zeta_{\min}$, such that $H(k/p)/\delta + \psi(k/n, \kappa'_{\min}/\kappa_{\min}^{1/2}) <$

$-2c$. Because $C(n,p,\kappa'_{\min}/\kappa_{\min}^{1/2})e^{-cn}$ is upper bounded by a constant C depending only on $\delta,\kappa_{\min},c$, we conclude there exists $C,c,\kappa'_{\min}>0$ depending only on $\delta,\kappa_{\min}^{1/2},\zeta_{\min}$ such that

$$\mathbb{P}\left(\kappa_-(\boldsymbol{X},n(1-\zeta^*/4))\le\kappa'_{\min}\right)\le Ce^{-cn}\,.$$

This conclude the proof of the high-probability bound Eq. (84).

The remainder of the argument takes place on the high-probability event $\mathcal{A}$. Consider any $\boldsymbol{\theta}\in\mathbb{R}^p$. We first construct $S_+\supset S(\Delta/2)$ such that

$$\text{(86)}\qquad \textit{(i)}\ |S_+|\le n(1-\zeta^*/4)\quad\text{and}\quad \textit{(ii)}\ \frac{1}{\sqrt{p}}\|\boldsymbol{\theta}_{S_+^c}\|_2\le\frac{2\sqrt{2}}{p\sqrt{\zeta^*\delta}}\|\boldsymbol{\theta}_{S(\Delta/2)^c}\|_1\,,$$

where we adopt the convention that $\|\boldsymbol{\theta}_\emptyset\|_1=\|\boldsymbol{\theta}_\emptyset\|_2=0$. We establish this by considering two cases.

Case 1: $p\le n(1-\zeta^*/4)$. In this case, let $S_+=[p]$. Then Eq. (86) holds trivially.

Case 2: $p>n(1-\zeta^*/4)$. Let $S_1,\dots,S_k$ be a partition of $[p]\setminus S(\Delta/2)$ satisfying the following properties: first, $|S_i|\ge n\zeta^*/8$ for $i=1,\dots,k-1$; second, $|S_1|\ge\cdots\ge|S_k|$; third, $|S(\Delta/2)\cup S_1|\le n(1-\zeta^*/4)$; and fourth, $|\theta_j|\ge|\theta_{j'}|$ if $j\in S_i$ and $j'\in S_{i'}$ for $i\le i'$. This is possible because $|S(\Delta/2)|\le n(1-\zeta^*/4)$ and, because $n\zeta^*/8\ge 1$, there exists an integer between $n(1-\zeta^*/4)$ an $n(1-\zeta^*/8)$. In this case, let $S_+=S(\Delta/2)\cup S_1$. Condition *(i)* holds by construction. To verify condition *(ii)*, observe

$$\begin{aligned}\frac{1}{p}\|\boldsymbol{\theta}_{S_+^c}\|_2^2=\frac{1}{p}\sum_{i=2}^k\|\boldsymbol{\theta}_{S_i}\|_2^2\le\frac{1}{p}\sum_{i=2}^k|S_i|\left(\frac{\|\boldsymbol{\theta}_{S_{i-1}}\|_1}{|S_{i-1}|}\right)^2&\le\frac{1}{p\min_{i=1,\dots,k-1}\{|S_{i-1}|\}}\sum_{i=1}^{k-1}\|\boldsymbol{\theta}_{S_i}\|_1^2\\&\le\frac{8}{p^2\zeta^*\delta}\|\boldsymbol{\theta}_{S(\Delta/2)^c}\|_1^2\,,\end{aligned}$$

where the first inequality holds because $|\theta_j|\le\|\boldsymbol{\theta}_{S_{i-1}}\|_1/|S_{i-1}|$ for $j\in S_i$, the second inequality holds because $|S_i|\le|S_{i-1}|$, and the third inequality holds because $|S_i|\ge n\zeta^*/8$ for $i\le k-1$. Thus, Eq. (86) holds in this case as well.

We lower bound the growth of the Lasso objective by

$$\begin{aligned}\mathcal{R}(\boldsymbol{\theta})-\mathcal{R}(\widehat{\boldsymbol{\theta}})&=\frac{1}{2n}\|\boldsymbol{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\|_2^2+\frac{1}{n}\langle\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\rangle+\frac{\lambda}{n}\left(\|\boldsymbol{\theta}\|_1-\|\widehat{\boldsymbol{\theta}}\|_1\right)\\&=\frac{1}{2n}\|\boldsymbol{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\|_2^2+\frac{\lambda}{n}\left(\langle\widehat{\boldsymbol{t}},\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\rangle+\|\boldsymbol{\theta}\|_1-\|\widehat{\boldsymbol{\theta}}\|_1\right)\,.\end{aligned}$$

We first make the observation that

$$\langle\widehat{\boldsymbol{t}},\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\rangle+\|\boldsymbol{\theta}\|_1-\|\widehat{\boldsymbol{\theta}}\|_1\ge\frac{\Delta}{2}\|\boldsymbol{\theta}_{S(\Delta/2)^c}\|_1\,.$$

Because $\widehat{\boldsymbol{t}}\in\partial\|\widehat{\boldsymbol{\theta}}\|_1$ and $|t_j|\le 1-\Delta/2$ on $S(\Delta/2)^c$ so that $t_j(\widehat{\theta}_j-\theta_j)+|\theta_j|-|\widehat{\theta}_j|\ge 0$ for all j, and is no smaller than $\Delta|\theta_j|/2$ for $j\in S(\Delta/2)^c$. Thus, it is guaranteed that

$$\mathcal{R}(\boldsymbol{\theta})-\mathcal{R}(\widehat{\boldsymbol{\theta}})\ge\frac{\lambda\Delta}{2n}\|\boldsymbol{\theta}_{S(\Delta/2)^c}\|_1+\frac{1}{2n}\|\boldsymbol{X}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\|_2^2\,.$$

Now choose $S_+\subset[p]$ satisfying Eq. (86). Condition *(ii)* of Eq. (86) implies

$$\text{(87)}\qquad \mathcal{R}(\boldsymbol{\theta})-\mathcal{R}(\widehat{\boldsymbol{\theta}})\ge\frac{c_1}{\sqrt{p}}\|\boldsymbol{\theta}_{S_+^c}\|_2\,,$$

where $c_1 > 0$ depends on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ. Next we prove that there exists $c' > 0$ such that for $\|\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+}\|_2/\sqrt{p} < c'$,

$$\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \frac{c'}{\sqrt{p}}\|\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+}\|_2^2 \text{ holds true on event } \mathcal{A}. \tag{88}$$

In order to see this, if $\|\boldsymbol{X}_{S_+}(\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+})\|_2/2 \geq \|\boldsymbol{X}_{S_+^c}\boldsymbol{\theta}_{S_+^c}\|_2$, then

$$\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \frac{1}{8n}\|\boldsymbol{X}_{S_+}(\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+})\|_2^2 \geq \frac{\kappa_{\min}'^2}{8n}\|\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+}\|_2^2,$$

as a consequence of Eq. (84). Otherwise, if $\|\boldsymbol{X}_{S_+}(\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+})\|_2/2 < \|\boldsymbol{X}_{S_+^c}\boldsymbol{\theta}_{S_+^c}\|_2$, then $\|\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+}\|_2 \leq \kappa_{\min}'^{-1/2}\|\boldsymbol{X}_{S_+}(\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+})\|_2 \leq 2\kappa_{\min}'^{-1/2}\|\boldsymbol{X}_{S_+^c}\boldsymbol{\theta}_{S_+^c}\|_2 \leq C\|\boldsymbol{\theta}_{S_+^c}\|_2$. Thus

$$\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \frac{c_1}{\sqrt{p}}\|\boldsymbol{\theta}_{S_+^c}\|_2 \geq \frac{c_1}{\sqrt{p}}\|\boldsymbol{\theta}_{S_+} - \widehat{\boldsymbol{\theta}}_{S_+}\|_2,$$

where the value of c_1 changes between the last inequalities. Combining the previous two displays, we have established inequality (88), where again the value of c' has changed from the previous displays.

Combined with Eq. (87), we conclude there exists $c_1, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that for $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2/\sqrt{p} \leq c'$,

$$\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\widehat{\boldsymbol{\theta}}) \geq \frac{c'}{p}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2.$$

The proof is completed.

B.5.5. *Proof of Corollary 12.* To start, let us define

$$\phi_1(x, \Delta) := \min(1, x/\Delta - z_{1-q/2}/\Delta + 1)_+.$$

The function $\phi_1(x)$ equals to 0 for $x \leq z_{1-q/2} - \Delta$ and 1 for $x \geq z_{1-q/2}$, and linearly interpolates between these two regions elsewhere. Therefore, the false-coverage proportion $\mathsf{FCP} := \frac{1}{p}\sum_{j=1}^p \mathbf{1}_{\theta_j^* \notin \mathsf{CI}_j}$ can be controlled as

$$\begin{aligned}
\mathsf{FCP} &= \frac{1}{p}\sum_{j=1}^p \mathbf{1}\left\{|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*| > \frac{\Sigma_{j|-j}^{-1/2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)} z_{1-q/2}\right\} \\
&\leq \frac{1}{p}\sum_{j=1}^p \phi_1\left(\frac{\Sigma_{j|-j}^{1/2}(1 - \|\widehat{\boldsymbol{\theta}}\|_0/n)|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}}, \Delta\right) \\
&\leq \frac{1}{p}\sum_{j=1}^p \phi_1\left(\Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|/\tau^*, \Delta\right) + \frac{1}{\Delta}\left|\frac{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}} - \frac{1}{\tau^*}\right|\left(\frac{1}{p}\sum_{j=1}^p \Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|\right).
\end{aligned}$$

We bound the terms on the right-hand side respectively.

- By Theorems 8 and Theorem 9, there exist $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that for $\epsilon < c'$, we have $\left|\frac{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2/\sqrt{n}} - \frac{1}{\tau^*}\right| < \epsilon$ with probability at least $1 - \frac{C}{\epsilon^3}e^{-cn\epsilon^6}$.

- Because $\Sigma_{j|-j}^{1/2} \le \kappa_{\max}^{1/2}$ for all j, the quantity $\frac{1}{p}\sum_{j=1}^p \Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|$ is $\sqrt{\kappa_{\max}/p}$-Lipschitz in $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$. Moreover, when $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ is replaced by $\boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}$, this equantity has expectation bounded by a constant depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ. By Theorem 11, there exist $C, C', c > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that $\frac{1}{p}\sum_{j=1}^p \Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*| < C'$ with probability at least $1 - Ce^{-cn}$.
- The quantity $\frac{1}{p}\sum_{j=1}^p \phi_1\left(\Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|/\tau^*, \Delta\right)$ is $\frac{L}{\Delta\sqrt{p}}$-Lipschitz in $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$, where L is a constant depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ. By Theorem 11, we conclude there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for $\epsilon < c'$, we have

$$\frac{1}{p}\sum_{j=1}^p \phi_1\left(\Sigma_{j|-j}^{1/2}|\widehat{\theta}_j^{\mathrm{d}} - \theta_j^*|/\tau^*, \Delta\right) < \frac{1}{p}\sum_{j=1}^p \mathbb{E}\left[\phi_1\left(\Sigma_{j|-j}^{1/2}|(\tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g})_j|/\tau^*, \Delta\right)\right] + \epsilon/\Delta,$$

  with probability at least $1 - \frac{C}{\epsilon^3}e^{-cn\epsilon^6}$.
- Using the fact that the standard Gaussian density is upper bounded by $(2\pi)^{-1/2}$, we obtain the bound $\mathbb{E}\left[\phi_1\left(\Sigma_{j|-j}^{1/2}|(\tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g})_j|/\tau^*, \Delta\right)\right] \le q + \frac{2\Delta}{\sqrt{2\pi}}$.

Combining the previous bounds, we conclude there exist $C, C', c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for all $\epsilon < c'$, we have $\mathsf{FCP} \le q + C'(\Delta + \epsilon/\Delta)$ with probability at least $1 - \frac{C}{\epsilon^3}e^{-cn\epsilon^6}$. Optimizing over Δ, we conclude there exists $C, c, c' > 0$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, and δ such that for all $\epsilon < c'$, we have $\mathsf{FCP} \le q + \epsilon$ with probability at least $1 - \frac{C}{\epsilon^6}e^{-cn\epsilon^{12}}$.

The lower bound holds similarly.

B.6. *More details on confidence interval for a single coordinate.* Because they may be of independent interest, we first describe in detail the construction of the exact tests outlined in the discussion in Section 3.5 and state some results about the quantities involved in the construction (Lemma B.10 and Theorem B.11 below). The proof of Theorem 13 uses a special case of Lemma B.10, whereas Theorem B.11 is independent of any future development, and is stated only for general interest.

B.6.1. *Description of exact test.* For any $\omega \in \mathbb{R}$, the statistician may construct the "pseudo-outcome" $\boldsymbol{y}^\omega := \boldsymbol{y} - \omega\breve{\boldsymbol{x}}_j^\perp$. We have

$$\boldsymbol{y}^\omega = \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\mathrm{loo}}^* + \breve{\boldsymbol{x}}_j^\perp(\theta_j^* - \omega) + \sigma\boldsymbol{z} \tag{89}$$

where $\boldsymbol{\theta}_{\mathrm{loo}}^*$ is defined in (21). As with the leave-one-out model of Section 3.5, expression (89) can be viewed as defining a linear-model with $p-1$ covariates, true parameter $\boldsymbol{\theta}_{\mathrm{loo}}^*$, and noise variance $\sigma_{\mathrm{loo}}^2(\omega) = \sigma^2 + \frac{\Sigma_{j|-j}(\theta_j^* - \omega)^2}{n}$. Generalizing the leave-one-out lasso estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$, the variable importance statistic ξ_j, and the estimated effective noise level in the leave-one-out model $\widehat{\tau}^j$, we define

$$\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{p-1}}\left\{\frac{1}{2n}\|\boldsymbol{y}^\omega - \boldsymbol{X}_{-j}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{\theta}\|_1\right\},$$

and similarly

$$\xi_j^\omega := \frac{(\breve{\boldsymbol{x}}_j^\perp)^\top(\breve{\boldsymbol{y}}^\omega - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega)}{\Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega\|_0/n)},$$

and

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

Whereas the statistic ξ_j should be large when $\theta_j^* \neq 0$, the statistic ξ_j^{ω} should be large when $\theta_j^* \neq \omega$. It is normally distributed conditional on $(\boldsymbol{y}, \boldsymbol{X}_{-j})$ when $\theta_j^* = \omega$, and is always approximately normally distributed unconditionally.

LEMMA B.10. *We have the following.*

(a) (Exact conditional normality of ξ_j^{ω} when $\theta_j^* = \omega$). *If $\theta_j^* = \omega$, then*

$$\xi_j^{\omega}/\widehat{\tau}_{\text{loo}}^{\omega,j} \sim \mathsf{N}(0, \Sigma_{j|-j}^{-1})\,. \tag{90}$$

(b) (Approximate normality of ξ_j^{ω} in general). *Assume $p \geq 2$. Let $\delta_{\text{loo}} = n/(p-1)$. Assume λ, $\boldsymbol{\Sigma}$, and σ satisfy assumption A1, and that $\boldsymbol{\theta}_{-j}^*$ is $(s, \sqrt{\delta_{\text{loo}}}(1-\Delta_{\min}), M)$-approximately sparse with respect to covariance $\boldsymbol{\Sigma}_{-j,-j}$ for some $s/(p-1) \geq \nu_{\min} > 0$ and $1 \geq \Delta_{\min} > 0$. Let $M' > 0$ be such that $|\theta_j^* - \omega| \leq M'(p-1)^{1/4}$ and $|\theta_j^*| \leq M'(p-1)^{1/4}$.*

Then there exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, M, M', and δ_{loo} such that the following occurs. There exist random variables r_j, R_j, Z_j such that

$$(\xi_j^{\omega} - (\theta_j^* - \omega))/\widehat{\tau}_{\text{loo}}^{\omega,j} = r_j Z_j + R_j\,,$$

and for all $\epsilon < c'$

$$Z_j \sim \mathsf{N}(0, \Sigma_{j|-j}^{-1})\,, \qquad \mathbb{P}\left(|r_j - 1| > \epsilon\right) \leq \frac{C}{\epsilon^2} e^{-cn\epsilon^4}\,, \qquad \mathbb{P}\left(|R_j| > \epsilon\right) \leq \frac{C}{n\epsilon^2}.$$

Lemma B.10(a) implies that the test which rejects when $|\xi_j^{\omega}| \geq \Sigma_{j|-j}^{-1/2}\widehat{\tau}_{\text{loo}}^{\omega,j} z_{1-\alpha/2}$ is an exact level-α test of the null $\theta_j^* = \omega$. Lemma B.10(b) states that under the alternative ξ_j^{ω} is approximately normal with mean $\theta_j^* - \omega$ and standard deviation $\widehat{\tau}_{\text{loo}}^{\omega,j}$. (The latter quantity is random but concentrates). Thus, Lemma B.10(b) permits a power analysis of the exact test.

The next theorem is included because it may be of independent interest. No future development depend upon this theorem, and it can safely be skipped.

THEOREM B.11. *Let $\tau_{\text{loo}}^{\omega,*}, \zeta_{\text{loo}}^{\omega,*}$ be the solution to the fixed point equations (8a) and (8b) in the model (89) for the Lasso at regularization λ.*

(a) (Power of exact test). *There exist constants $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, M, M', and δ_{loo} such that for all $\epsilon < c'$,*

$$\left|\mathbb{P}_{\theta_j^*}\left(|\xi_j^{\omega}| \geq \Sigma_{j|-j}^{-1/2}\widehat{\tau}_{\text{loo}}^{\omega,j} z_{1-\alpha/2}\right) - \mathbb{P}\left(|\theta_j^* + \tau_{\text{loo}}^{\omega,*} G - \omega| > \tau_{\text{loo}}^{\omega,*} z_{1-\alpha/2}\right)\right| \leq C\left((1 + |\theta_j^* - \omega|)\epsilon + \frac{1}{\epsilon^2}e^{-cn\epsilon^6} + \frac{1}{n\epsilon^2}\right), \tag{91}$$

where $G \sim \mathsf{N}(0,1)$.

(b) (Insensitivity of fixed point parameter to ω). *There exists $L, M_1' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, M, M', and δ_{loo} such that for $|\theta_j^* - \omega|/\sqrt{n} \leq M_1'$, we have $|\tau_{\text{loo}}^{\omega,*} - \tau_{\text{loo}}^{\theta_j^*,*}| \leq L|\theta_j^* - \omega|/\sqrt{n}$.*

Theorem B.11 says that the exact test of the null $\theta_j^* = \omega$ described above has power approximately equal to that achieved by a standard two-sided confidence interval for Gaussian observations at noise-variance $(\tau_{\text{loo}}^{\omega,*})^2$. Part (b) states that this noise variance is effectively constant for all $\omega = o(\sqrt{n})$.

Because we have an exact test for $\theta_j^* = \omega$ for all $\omega \in \mathbb{R}$, we may in principle construct exact confidence intervals by inverting this collection of tests. As described in Section 3.5, constructing such confidence intervals would require computing a Lasso estimate at each value of ω.

B.6.2. *Proof of Lemma B.10, Theorem 13, and Theorem B.11.*

Proof of Lemma B.10(a). Wehn $\theta_j^* = \omega$, the data $(\boldsymbol{y}^\omega, \boldsymbol{X}_{-j})$ is independent of $\breve{\boldsymbol{x}}_j^\perp$. Because $\widehat{\boldsymbol{\theta}}_{\text{loo}}^\omega$ is $\sigma(\boldsymbol{y}^\omega, \boldsymbol{X}_{-j})$-measurable,

$$\xi_j^\omega / \widehat{\tau}_{\text{loo}}^{\omega,j} = \frac{\sqrt{n}(\breve{\boldsymbol{x}}_j^\perp)^\top (\boldsymbol{y}^\omega - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}^\omega)}{\Sigma_{j|-j}\|\boldsymbol{y}^\omega - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}^\omega\|_2}\,.$$

Because $\breve{\boldsymbol{x}}_j^\perp \sim \mathsf{N}(0, \Sigma_{j|-j}\mathbf{I}_p/n)$ and is independent of $\boldsymbol{y}^\omega - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}^\omega$, conditionally on $\boldsymbol{y}^\omega, \boldsymbol{X}_{-j}$ the quantity is distributed $\mathsf{N}(0, \Sigma_{j|-j}^{-1})$. Thus, it is distributed $\mathsf{N}(0, \Sigma_{j|-j}^{-1})$ unconditionally as well. We have established (90).

Proof of Lemma B.10(b). We may without loss of generality consider the case $\omega = 0$. Indeed, the joint distribution of $(\boldsymbol{y}^\omega, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp, \widehat{\boldsymbol{\theta}}_{\text{loo}}^\omega)$ under θ_j^* is equal to the joint distribution of $(\boldsymbol{y}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp, \widehat{\boldsymbol{\theta}}_{\text{loo}}^0)$ if the j^{th} coordinate were instead $\theta_j^* - \omega$. Under this transformation, the conditions of the Theorem are still met, possibly with M' replaced by $2M'$.

Thus, consider the case $\omega = 0$. Note $\boldsymbol{y}^0 = \boldsymbol{y}$, $\widehat{\boldsymbol{\theta}}_{\text{loo}}^0 = \widehat{\boldsymbol{\theta}}_{\text{loo}}$, and $\xi_j^0 = \xi_j$. To simplify notation, we thus remove the superscript 0 in the remainder of the argument. Define the quantity

$$\tilde{\xi}_j := (\breve{\boldsymbol{x}}_j^\perp)^\top (\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}) - \theta_j^* \Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n)\,.$$

Direct calculations give

$$\begin{aligned}\tilde{\xi}_j &= (\breve{\boldsymbol{x}}_j^\perp)^\top (\sigma\boldsymbol{z} + \breve{\boldsymbol{x}}_j^\perp\theta_j^* + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}) - \theta_j^*\Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n) \\ &= \underbrace{(\breve{\boldsymbol{x}}_j^\perp)^\top (\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\text{loo}})}_{\Delta_1} + \underbrace{(\breve{\boldsymbol{x}}_j^\perp)^\top (\breve{\boldsymbol{x}}_j^\perp\theta_j^* + \boldsymbol{X}_{-j}(\widehat{\boldsymbol{\theta}}'_{\text{loo}} - \widehat{\boldsymbol{\theta}}_{\text{loo}})) - \theta_j^*\Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n)}_{\Delta_2}\,,\end{aligned}$$

where

$$\widehat{\boldsymbol{\theta}}'_{\text{loo}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{\theta}\|_1 \right\}.$$

In particular, $\widehat{\boldsymbol{\theta}}'_{\text{loo}}$ is $\sigma(\boldsymbol{z}, \boldsymbol{X}_{-j})$-measurable, so is independent of $\breve{\boldsymbol{x}}_j^\perp$, whence

$$\Delta_1 \,|\, \boldsymbol{z}, \boldsymbol{X}_{-j} \sim \mathsf{N}\Big(0, \Sigma_{j|-j}\|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\text{loo}}\|_2^2/n\Big)\,.$$

The estimate $\widehat{\boldsymbol{\theta}}_{\text{loo}}$ is a function of $\boldsymbol{z}$, $\boldsymbol{X}_{-j}$, and $\breve{\boldsymbol{x}}_j^\perp$. We make this explicit by writing $\widehat{\boldsymbol{\theta}}_{\text{loo}}(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)$. Following this notation, $\widehat{\boldsymbol{\theta}}'_{\text{loo}}$ defined above is equal to $\widehat{\boldsymbol{\theta}}_{\text{loo}}(\boldsymbol{z}, \boldsymbol{X}_{-j}, \mathbf{0})$.

Next consider the term Δ_2. First define

$$F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp) := \breve{\boldsymbol{x}}_j^\perp\theta_j^* + \boldsymbol{X}_{-j}(\widehat{\boldsymbol{\theta}}_{\text{loo}}(\boldsymbol{z}, \boldsymbol{X}_{-j}, \mathbf{0}) - \widehat{\boldsymbol{\theta}}_{\text{loo}}(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp))\,.$$

Use $\nabla_{\breve{\boldsymbol{x}}_j^\perp}$ to denote the Jacobian with respect to $\breve{\boldsymbol{x}}_j^\perp$. Almost surely, $\nabla_{\breve{\boldsymbol{x}}_j^\perp} F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp) = \theta_j^*(\mathbf{I}_n - \mathsf{P}_{\boldsymbol{X}_{-j\widehat{S}}})$, where $\mathsf{P}_{\boldsymbol{X}_{-j\widehat{S}}}$ is the projector onto the span of $\{\breve{\boldsymbol{x}}_k \mid k \in \widehat{S}\}$ and $\widehat{S}$ is the support of $\widehat{\boldsymbol{\theta}}_{\text{loo}}(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)$ [55]. The function F is θ_j^*-Lipschitz in $\breve{\boldsymbol{x}}_j^\perp$ for fixed $\boldsymbol{z}, \boldsymbol{X}_{-j}$. Therefore we conclude that $\Delta_2 = (\breve{\boldsymbol{x}}_j^\perp)^\top F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp) - \Sigma_{j|-j} \operatorname{div}_{\breve{\boldsymbol{x}}_j^\perp} F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)/n$. Applying Stein's formula and the second-order Stein's formula [8, Eq. (2.1) and Theorem 2.1], we get

$$\mathbb{E}[\Delta_2 | \boldsymbol{z}, \boldsymbol{X}_{-j}] = 0 \text{ and } \operatorname{Var}(\Delta_2 | \boldsymbol{z}, \boldsymbol{X}_{-j}) = \frac{\Sigma_{j|-j}}{n}\Big(\mathbb{E}\Big[\|F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)\|_2^2 + \frac{\Sigma_{j|-j}}{n}\|\nabla_{\breve{\boldsymbol{x}}_j^\perp} F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)\|_{\mathrm{F}}^2 \;\Big|\; \boldsymbol{z}, \boldsymbol{X}_{-j}$$

Note that, almost surely $\|\nabla_{\breve{\boldsymbol{x}}_j^\perp} F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)\|_{\mathrm{F}}^2 = \theta_j^{*2}(n - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0)$. Further, $\|F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp)\|_2^2 \le \theta_j^{*2}\|\breve{\boldsymbol{x}}_j^\perp\|_2^2$ because F is θ_j^*-Lipschitz and $F(\boldsymbol{z}, \boldsymbol{X}_{-j}, \mathbf{0}) = \mathbf{0}$. Thus,

$$\operatorname{Var}(\Delta_2 | \boldsymbol{z}, \boldsymbol{X}_{-j}) \le \frac{\Sigma_{j|-j}}{n}\mathbb{E}\Big[\theta_j^{*2}\Big(\|\breve{\boldsymbol{x}}_j^\perp\|_2^2 + \Sigma_{j|-j}\Big(1 - \frac{\|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0}{n}\Big)\Big)\Big|\boldsymbol{z}, \boldsymbol{X}_{-j}\Big] \le \frac{2\Sigma_{j|-j}^2\theta_j^{*2}}{n} \text{ almost surely.}$$

Because the $\mathbb{E}[\Delta_2|\boldsymbol{z}, \boldsymbol{X}_{-j}] = 0$ almost surely, we have $\operatorname{Var}(\Delta_2) \le 2\Sigma_{j|-j}^2\theta_j^{*2}/n$ as well. Next observe that $\frac{\tilde{\xi}_j}{\Sigma_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n)} = \xi_j - \theta_j^*$. Thus,

$$\begin{aligned}\frac{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_0/n)(\xi_j - \theta_j^*)}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_2} &= \frac{\|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}'\|_2}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_2}\frac{\Delta_1}{\Sigma_{j|-j}\|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}'\|_2/\sqrt{n}} \\ &\qquad + \frac{\Delta_2}{\Sigma_{j|-j}\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_2/\sqrt{n}} \\ &=: r_j Z_j + R_j\,,\end{aligned}$$

where $Z_j := \frac{\Delta_1}{\Sigma_{j|-j}\|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}'\|_2/\sqrt{n}} \sim \mathsf{N}(0, \Sigma_{j|-j}^{-1})$ (and normality follows by the proof of Eq. (90)).

The singular values of $\boldsymbol{\Sigma}_{-j,-j}$ are bounded between the minimal and maximal singular values of $\boldsymbol{\Sigma}$. Thus, the matrix $\boldsymbol{\Sigma}_{-j,-j}$ satisfies assumption A1(b) because $\boldsymbol{\Sigma}$ does. In particular, the triple λ, $\boldsymbol{\Sigma}_{-j,-j}$, and σ satisfy assumption A1. Because $\boldsymbol{\theta}_{-j}^*$ is $(s, \sqrt{\delta_{\text{loo}}}(1 - \Delta_{\min}), M)$-approximately sparse, we may choose $\boldsymbol{x}_{-j} \in \{-1, 0, 1\}^{p-1}$ and $\bar{\boldsymbol{\theta}}_{-j}^* \in \mathbb{R}^{p-1}$ such that

$$\begin{gathered}\|\boldsymbol{\theta}_{-j}^* - \bar{\boldsymbol{\theta}}_{-j}^*\|_1/(p-1) \le M, \quad \boldsymbol{x}_{-j} \in \partial\|\bar{\boldsymbol{\theta}}_{-j}^*\|_1, \\ \text{and } \|\boldsymbol{x}_{-j}\|_0/(p-1) \ge \nu_{\min}, \quad \mathcal{G}(\boldsymbol{x}_{-j}, \boldsymbol{\Sigma}_{-j,-j}) \le \sqrt{n/(p-1)}(1 - \Delta_{\min}).\end{gathered}$$

Also, $\|\theta_j^*\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}\|_1/(p-1) \le M'\|\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}\|_2/\sqrt{p-1} \le M'\kappa_{\min}^{-1/2}\kappa_{\max}/\sqrt{p-1}$. Using the same sparse approximation $\bar{\boldsymbol{\theta}}_{-j}^*$ and subgradient $\boldsymbol{x}_{-j}$, we conclude that

$$\boldsymbol{\theta}_{\text{loo}}^* \text{ is } (s, \sqrt{\delta_{\text{loo}}}(1 - \Delta_{\min}), M + M'\kappa_{\min}^{-1/2}\kappa_{\max}/\sqrt{p-1}) - \text{approximately sparse.}$$

By Theorem 2, there exists $\mathcal{P}_{\text{fixPt}}$ depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, and δ such that assumption A2 is satisfied by the observations $\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^*$ and design matrix $\boldsymbol{X}_{-j}$ which are used to fit $\widehat{\boldsymbol{\theta}}_{\text{loo}}'$.

Because F is θ_j^*-Lipschitz in $\breve{\boldsymbol{x}}_j^\perp$, we have $|\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}\|_2 - \|\sigma\boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}_{\text{loo}}^* - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}'\|_2| < \theta_j^*\|\breve{\boldsymbol{x}}_j^\perp\|_2$. By Theorem 8 and since $\theta_j^{*2}\|\breve{\boldsymbol{x}}_j^\perp\|_2^2/n \sim \theta_j^{*2}\chi_n^2/n^2$, there exist

$C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, δ, and M' such that for $\epsilon < c'$, it is guaranteed that

$$\begin{aligned}\mathbb{P}\left(|r_j - 1| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{\|\sigma \boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}^*_{\rm loo} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\rm loo}\|_2/\sqrt{n}}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\rm loo}\|_2/\sqrt{n}} - 1\right| > \epsilon\right)\\ &= \mathbb{P}\left(\left|\frac{\|\sigma \boldsymbol{z} + \boldsymbol{X}_{-j}\boldsymbol{\theta}^*_{\rm loo} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\rm loo}\|_2/\sqrt{n} - \|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\rm loo}\|_2/\sqrt{n}}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\rm loo}\|_2/\sqrt{n}}\right| > \epsilon\right)\\ &\le \mathbb{P}\left(\frac{|\theta^*_j|\|\breve{\boldsymbol{x}}^\perp_j\|_2/\sqrt{n}}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\rm loo}\|_2/\sqrt{n} - |\theta^*_j|\|\breve{\boldsymbol{x}}^\perp_j\|_2/\sqrt{n}} > \epsilon\right) < \frac{C}{\epsilon^2}e^{-cn\epsilon^4},\end{aligned}$$

Where the final inequality has the following justification: First, we have used that $\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\rm loo}\|_2/\sqrt{n}$ concentrates on a quantity for which we have a lower bound by Theorem 8. Second, we have used that $\mathbb{P}(|\theta^*_j|\|\breve{\boldsymbol{x}}^\perp_j\|_2/\sqrt{n} > M'(p-1)^{1/4}/\sqrt{n}\Sigma^{1/2}_{j|-j} + t) \le C\exp(-cn^2t^2)$. The right-hand side of the preceding display is larger than 1 from $\epsilon = O(n^{-1/4})$ and $M'(p-1)^{1/4}/\sqrt{n}\Sigma^{1/2}_{j|-j} = O(n^{-1/4})$, whence we get $\mathbb{P}(|\theta^*_j|\|\breve{\boldsymbol{x}}^\perp_j\|_2/\sqrt{n} > \epsilon) \le \frac{C}{\epsilon^2}\exp(-cn\epsilon^4)$, as desired. The O's hide constants depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, δ, and M'.

Similarly, combining the concentration of $\Sigma_{j|-j}\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}'_{\rm loo}\|_2/\sqrt{n}$ on a quantity for which we have a lower bound, the high probability upper bound on $|\theta^*_j|\|\breve{\boldsymbol{x}}^\perp_j\|_2/\sqrt{n}$, and Chebyshev's inequality applied to Δ_2, there exists $C, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, δ, and M' such that for $\epsilon < c'$ such that

$$\mathbb{P}\left(|R_j| > \epsilon\right) < \frac{C}{n\epsilon^2}.$$

The proof of the lemma is complete.

Proof of Theorem 13(a). The event $\theta \notin \mathsf{CI}^{\rm loo}_j$ is equivalent to

$$\frac{\Sigma^{1/2}_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\rm loo}\|_0/n)|\xi_j - \theta|}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\rm loo}\|_2/\sqrt{n}} \ge z_{1-\alpha/2}.$$

With r_j, R_j defined as in Theorem B.10 for $\omega = 0$, this is equivalent to

$$A := \Sigma^{1/2}_{j|-j}(r_jZ_j + R_j) + \frac{\Sigma^{1/2}_{j|-j}(1 - \|\widehat{\boldsymbol{\theta}}_{\rm loo}\|_0/n)(\theta^*_j - \theta)}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\rm loo}\|_2/\sqrt{n}} \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}].$$

By Theorems 8 and 9 on concentration of the Lasso residual and sparsity and Theorem B.10 on the concentration of r_j and R_j, there exist $C, c, c' > 0$ depending only on $\mathcal{P}_{\rm model}$, $\mathcal{P}_{\rm fixPt}$, $\delta_{\rm loo}$, and M' such that for all $\epsilon < c'$,

$$\mathbb{P}\left(\left|A - \Sigma^{1/2}_{j|-j}Z_j - \frac{\Sigma^{1/2}_{j|-j}(\theta^*_j - \theta)}{\tau^{*,0}_{\rm loo}}\right| > (1 + |\theta^*_j - \theta|)\epsilon\right) \le \frac{C}{\epsilon^3}e^{-cn\epsilon^6} + \frac{C}{n\epsilon^2}.$$

Thus, by direct calculation (where C may take different values between lines)

$$
\begin{aligned}
\mathbb{P}\left(A \notin [-z_{1-\alpha/2}, z_{1+\alpha/2}]\right) &\geq \mathbb{P}\left(\left|\Sigma_{j|-j}^{1/2} Z_j + \frac{\Sigma_{j|-j}^{1/2}(\theta_j^* - \theta)}{\tau_{\mathrm{loo}}^{*,0}}\right| > z_{1-\alpha/2} + (1 + |\theta_j^* - \theta|)\epsilon\right) \\
&\qquad - \mathbb{P}\left(\left|A - \Sigma_{j|-j}^{1/2} Z_j - \frac{\Sigma_{j|-j}^{1/2}(\theta_j^* - \theta)}{\tau_{\mathrm{loo}}^{*,0}}\right| > (1 + |\theta_j^* - \theta|)\epsilon\right) \\
&\geq \mathbb{P}\left(\left|\Sigma_{j|-j}^{1/2} Z_j + \frac{\Sigma_{j|-j}^{1/2}(\theta_j^* - \theta)}{\tau_{\mathrm{loo}}^{*,0}}\right| > z_{1-\alpha/2} + (1 + |\theta_j^* - \theta|)\epsilon\right) - C\left(\frac{1}{\epsilon^3} e^{-cn\epsilon^6} + \frac{1}{n\epsilon^2}\right) \\
&\geq \mathbb{P}\left(|\theta_j^* + \Sigma_{j|-j}^{-1/2} \tau_{\mathrm{loo}}^{*,0} G - \theta| \geq \Sigma_{j|-j}^{-1/2} \tau_{\mathrm{loo}}^{*,0} z_{1-\alpha/2}\right) - C\left((1 + |\theta_j^* - \theta|)\epsilon + \frac{1}{\epsilon^3} e^{-cn\epsilon^6} + \frac{1}{n\epsilon^2}\right).
\end{aligned}
$$

The reverse inequality is obtained similarly.

Proof of Theorem 13(b). By definition, one has

$$
\frac{\widehat{\tau}_{\mathrm{loo}}^{j}}{\tau_{\mathrm{loo}}^{*}} = \frac{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}\|_2/\sqrt{n}}{(1 - \|\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}\|_0/n)\tau_{\mathrm{loo}}^{*}}.
$$

The singular values of $\boldsymbol{\Sigma}_{-j,-j}$ are bounded between the minimal and maximal singular values of $\boldsymbol{\Sigma}$. Thus, the matrix $\boldsymbol{\Sigma}_{-j,-j}$ satisfies assumption A1(b) because $\boldsymbol{\Sigma}$ does. Further, $\sigma_{\min}^2 \leq \sigma^2 \leq \sigma_{\mathrm{loo}}^2 \leq \sigma_{\max}^2 + \frac{\kappa_{\max} M'^2 (p-1)^{1/2}}{n} \leq \sigma_{\max}^2 + \kappa_{\max} M'^2$. In particular, the triple λ, $\boldsymbol{\Sigma}_{-j,-j}$, and σ_{loo} satisfy assumption A1 with $\sigma_{\max}^2$ replaced by $\sigma_{\max}^2 + \kappa_{\max} M'^2$. (In fact, that as $n, p \to \infty$, we have an upper bound on σ_{loo} which converges to $\sigma_{\max}$).

Further, as argued in the proof of Theorem B.10(b), $\boldsymbol{\theta}_{\mathrm{loo}}^*$ is $(s, \sqrt{\delta_{\mathrm{loo}}}(1 - \Delta_{\min}), M + M'\kappa_{\min}^{-1/2}\kappa_{\max}/\sqrt{p-1})$-approximately sparse, so is in fact $(s, \sqrt{\delta_{\mathrm{loo}}}(1 - \Delta_{\min}), M + M'\kappa_{\min}^{-1/2}\kappa_{\max})$-approximately sparse. (In fact, we have as $n, p \to \infty$ that the ℓ_1-approximation constant $M + M'\kappa_{\min}^{-1/2}\kappa_{\max}/\sqrt{p-1}$ converges to M). By Theorem 2, there exists $\mathcal{P}_{\mathrm{fixPt}}$ such that assumption A2 holds on the model $(\boldsymbol{y}, \boldsymbol{X}_{-j})$. Equation (25) follows from Theorems 8 and 9 on the concentration results for the Lasso residual and the sparsity.

Proof of Lemma B.11(a). The joint distribution of $(\boldsymbol{y}^\omega, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp, \widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega)$ under θ_j^* is equal to the joint distribution of $(\boldsymbol{y}, \boldsymbol{X}_{-j}, \breve{\boldsymbol{x}}_j^\perp, \widehat{\boldsymbol{\theta}}_{\mathrm{loo}})$ if the j^{th} coordinate were instead $\theta_j^* - \omega$. (We also used this fact in the proof of Theorem B.10(b)). Thus,

$$
\begin{aligned}
\mathbb{P}_{\theta_j^*}(|\xi_j^\omega| \geq \Sigma_{j|-j}^{-1/2}\widehat{\tau}_{\mathrm{loo}}^{\omega,j} z_{1-\alpha/2}) &= \mathbb{P}_{\theta_j^*}\left(\frac{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega\|_0/n)|\xi_j^\omega|}{\|\boldsymbol{y}^\omega - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}^\omega\|_2} > \Sigma_{j|-j}^{-1/2} z_{1-\alpha/2}\right) \\
&= \mathbb{P}_{\theta_j^* - \theta}\left(\frac{\sqrt{n}(1 - \|\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}\|_0/n)|\xi_j|}{\|\boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}\|_2} > \Sigma_{j|-j}^{-1/2} z_{1-\alpha/2}\right) \\
&= \mathbb{P}_{\theta_j^* - \omega}\left(0 \notin \mathsf{CI}_j^{\mathrm{loo}}\right).
\end{aligned}
$$

Then Eq. (91) follows from Eq. (24).

Proof of Theorem B.11(b). As argued in the proof of Theorem B.10(b), the leave-one-out parameter vector $\boldsymbol{\theta}_{\mathrm{loo}}^*$ is $(s, \sqrt{\delta_{\mathrm{loo}}}(1 - \Delta_{\min}), M + M_1'\kappa_{\min}^{-1/2}\kappa_{\max}/\sqrt{p-1})$-approximately sparse. We view (89) as defining linear model with aspect ratio $\delta_{\mathrm{loo}} = n/(p-1)$, design covariance $\boldsymbol{\Sigma}_{-j,-j}$, noise variance $\sigma_{\mathrm{loo}}^2(\omega) = \sigma^2 + \frac{\Sigma_{j|-j}(\theta_j^* - \omega)^2}{n}$, and true parameter $\boldsymbol{\theta}_{\mathrm{loo}}^*$. Note $\sigma_{\mathrm{loo}}^2(\theta_j^*) = \sigma^2$ is bounded by $\sigma_{\min}^2 \le \sigma^2 \le \sigma_{\max}^2$ by assumption. Thus, by Theorem 2, there exist parameters $\mathcal{P}_{\mathrm{fixPt}} = (\tau_{\min}, \tau_{\max}, \zeta_{\min}, \zeta_{\max})$ depending only on $\mathcal{P}_{\mathrm{model}}$, $\nu_{\min}$, $\Delta_{\min}$, M, M_1' such that $\tau_{\min} \le \tau_{\mathrm{loo}}^{\theta_j^*,*} \le \tau_{\max}$ and $\zeta_{\min} \le \zeta_{\mathrm{loo}}^{\theta_j^*,*} \le \zeta_{\max}$. The notation $\mathcal{P}_{\mathrm{fixPt}}$ will refer to the parameters in this lower and upper bound for the remainder of the proof.

To control the fixed point parameter $\tau_{\mathrm{loo}}^{\omega,*}$, we will study the functional objective $\mathscr{E}_0$ of Eq. (41) for the linear model (89) as we vary ω. For simplicity of notation, we will drop the subscript on $\mathscr{E}_0$. As we vary ω, the only parameter defining the leave-one-out model which changes is the noise variance $\sigma_{\mathrm{loo}}^2(\omega)$. Thus, we will make the dependence of the functional objective on σ_{loo} explicit but leave its dependency on all other parameters implicit. In particular, we write

$$\mathscr{E}^{\sigma_{\mathrm{loo}}}(\boldsymbol{v}) := \frac{1}{2}\left(\sqrt{\frac{\|\boldsymbol{v}(\boldsymbol{g})\|_{L^2}^2}{n} + \sigma_{\mathrm{loo}}^2} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}}{n}\right)_+^2 + \frac{\lambda}{n}\mathbb{E}\left\{\|\boldsymbol{\theta}_{\mathrm{loo}}^* + (\boldsymbol{\Sigma}_{-j,-j})^{-1/2}\boldsymbol{v}(\boldsymbol{g})\| - \|\boldsymbol{\theta}_{\mathrm{loo}}^*\|_1\right)\right\},$$

where we emphasize that $\mathscr{E}^{\sigma_{\mathrm{loo}}}$ is a function $L^2(\mathbb{R}^{p-1};\mathbb{R}^{p-1}) \to \mathbb{R}$ and where we take $\boldsymbol{g}$ to be the identity function in $L^2(\mathbb{R}^p;\mathbb{R}^p)$. The model (89) for $\omega = \theta_j^*$ corresponds to the choice $\sigma_{\mathrm{loo}}^2 = \sigma^2$. Denote the unique minimizer of $\mathscr{E}^{\sigma_{\mathrm{loo}}}$ by $\boldsymbol{v}^{\sigma_{\mathrm{loo}},*}$. Existence and uniqueness is guaranteed by the proof of Lemma A.2. Also by the proof of Lemma A.2,

$$\tau_{\mathrm{loo}}^{\omega,*} = \sqrt{\sigma_{\mathrm{loo}}^2(\omega) + \|\boldsymbol{v}^{\sigma_{\mathrm{loo}},*}\|_{L^2}^2/n}. \tag{92}$$

The objective $\mathscr{E}^{\sigma_{\mathrm{loo}}}$ is L-Lipschitz in σ_{loo}^2 on $\sigma_{\mathrm{loo}}^2 > \sigma_{\min}^2$ for some L depending only on $\mathcal{P}_{\mathrm{fixPt}}$. By the proof of Lemma A.5, there exists $r, a > 0$ depending only on $\mathcal{P}_{\mathrm{fixPt}}$ and δ_{loo} such that $\mathscr{E}^{\sigma_{\mathrm{loo}}}$ is a/n-strongly convex in $\boldsymbol{v}$ on $\|\boldsymbol{v} - \boldsymbol{v}^{\sigma_{\mathrm{loo}},*}\|_2/\sqrt{n} \le r$. Thus, for $\|\boldsymbol{v} - \boldsymbol{v}^{\sigma_{\mathrm{loo}},*}\|_{L^2}/\sqrt{n} \le r$,

$$\begin{aligned}\mathscr{E}^{\sigma_{\mathrm{loo}}}(\boldsymbol{v}) &\ge \mathscr{E}^{\sigma}(\boldsymbol{v}) - L|\sigma_{\mathrm{loo}}^2 - \sigma^2| \ge \mathscr{E}^{\sigma}(\boldsymbol{v}^{\sigma,*}) + \frac{a}{n}\|\boldsymbol{v} - \boldsymbol{v}^{\sigma,*}\|_{L^2}^2 - L|\sigma_{\mathrm{loo}}^2 - \sigma^2| \\ &\ge \mathscr{E}^{\sigma_{\mathrm{loo}}}(\boldsymbol{v}^{\sigma,*}) + \frac{a}{n}\|\boldsymbol{v} - \boldsymbol{v}^{\sigma,*}\|_{L^2}^2 - 2L|\sigma_{\mathrm{loo}}^2 - \sigma^2|\,.\end{aligned}$$

We conclude that if $\sqrt{2L|\sigma_{\mathrm{loo}}^2 - \sigma^2|/a} \le r$, then $\|\boldsymbol{v}^{\sigma_{\mathrm{loo}},*} - \boldsymbol{v}^{\sigma,*}\|_{L^2}/\sqrt{n} \le \sqrt{2L|\sigma_{\mathrm{loo}}^2 - \sigma^2|^{1/2}/a}$. Recalling that $\sigma_{\mathrm{loo}}^2(\omega) - \sigma^2 = \Sigma_{j|-j}(\theta_j^* - \omega)^2/n$, we see that for $|\theta_j^* - \omega|/\sqrt{n} \le M_3' := r\sqrt{a/(2L\kappa_{\max})}$ we have $\|\boldsymbol{v}^{\sigma_{\mathrm{loo}}(\omega),*} - \boldsymbol{v}^{\sigma,*}\|_{L^2}/\sqrt{n} \le \sqrt{2L\Sigma_{j|-j}/a}|\theta_j^* - \theta|/\sqrt{n}$. By Eq. (92),

$$\begin{aligned}|\tau_{\mathrm{loo}}^{\omega,*} - \tau_{\mathrm{loo}}^{\theta_j^*,*}| &= \left|\sqrt{\sigma_{\mathrm{loo}}^2(\omega) + \|\boldsymbol{v}^{\sigma_{\mathrm{loo}}(\omega),*}\|_{L^2}^2/n} - \sqrt{\sigma^2 + \|\boldsymbol{v}^{\sigma,*}\|_{L^2}^2/n}\right| \le \sqrt{\frac{2L\Sigma_{j|-j}}{a}}\frac{|\theta_j^* - \omega|}{\sqrt{n}} + \frac{1}{2\sigma}\frac{\Sigma_{j|-j}(\theta_j^* - \omega)^2}{n} \\ &\le L\frac{|\theta_j^* - \omega|}{\sqrt{n}}\,,\end{aligned}$$

where the L in the final line differs from the one in the preceding line and depends only on $\mathcal{P}_{\mathrm{model}}$, $\mathcal{P}_{\mathrm{fixPt}}$, δ_{loo}, and M_1'.

The proof is complete.

B.7. *Uniform control over λ: proof of Theorem 6.* To make the dependence of the Lasso objective on λ explicit, we write $\mathcal{R}^\lambda(\boldsymbol{\theta})$ for Eq. (1). As before, $\mathcal{C}^\lambda(\boldsymbol{v})$ is a re-parametrization of $\mathcal{R}^\lambda(\boldsymbol{\theta})$, namely $\mathcal{C}^\lambda(\boldsymbol{v}) := \mathcal{R}^\lambda(\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v})$. We also write $\widehat{\boldsymbol{\theta}}^\lambda$ for the minimizer of $\mathcal{R}^\lambda(\boldsymbol{\theta})$

and $\widehat{\boldsymbol{v}}^{\lambda}$ for the minimizer of $\mathcal{C}^{\lambda}(\boldsymbol{v})$ (in particular $\widehat{\boldsymbol{\theta}}^{\lambda} = \boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{\lambda}$). Finally, in order to expose the full dependency of η on the regularization parameter λ, we redefine

$$\eta(\boldsymbol{y}^f, \zeta/\lambda) := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{\zeta}{2\lambda}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \|\boldsymbol{\theta}\|_1\right\}.$$

Throughout this section we will use this definition instead of Eq. (4). We denote the Lasso error vector in the fixed-design model at regularization λ by

$$\widehat{\boldsymbol{v}}^{f,\lambda} := \boldsymbol{\Sigma}^{1/2}(\eta(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}, \zeta^*/\lambda) - \boldsymbol{\theta}^*),$$

where implicitly τ^*, ζ^* depend on λ via the fixed-point Eqs. (8a) and (8b). For simplicity, we write $\tilde{\phi}\Big(\frac{\boldsymbol{v}}{\sqrt{p}}\Big) = \phi\Big(\frac{\boldsymbol{\theta}^*+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)$. For $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, let

$$D_\epsilon^\lambda := \left\{\boldsymbol{v}\in\mathbb{R}^p \;\Big|\; \Big|\tilde{\phi}\Big(\frac{\boldsymbol{v}}{\sqrt{p}}\Big) - \mathbb{E}\Big[\tilde{\phi}\Big(\frac{\widehat{\boldsymbol{v}}^{f,\lambda}}{\sqrt{p}}\Big)\Big]\Big| > \epsilon\right\}.$$

Define $\mathscr{E}^\lambda : L^2(\mathbb{R}^p;\mathbb{R}^p) \to \mathbb{R}$ as $\mathscr{E}(\boldsymbol{v}) = \mathscr{E}_0(\boldsymbol{v})$, where $\mathscr{E}_0$ is as in the proof of Lemma A.3, and we make dependence on λ explicit in the notation. In particular,

$$\mathscr{E}^\lambda(\boldsymbol{v}) := \frac{1}{2}\Big(\sqrt{\frac{\|\boldsymbol{v}(\boldsymbol{g})\|_{L^2}^2}{n} + \sigma^2} - \frac{\langle \boldsymbol{g}, \boldsymbol{v}\rangle_{L^2}}{n}\Big)_+^2 + \frac{\lambda}{n}\mathbb{E}\Big\{\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{v}(\boldsymbol{g})\|_1 - \|\boldsymbol{\theta}^*\|_1\Big\}.$$

We emphasize that the argument $\boldsymbol{v}$ is not a vector but a function $\boldsymbol{v} : \mathbb{R}^p \to \mathbb{R}^p$. Recall, by the proof of Lemma A.2, that $\widehat{\boldsymbol{v}}^{f,\lambda}$, viewed as a function of $\boldsymbol{g}$ and thus a member of $L^2(\mathbb{R}^p;\mathbb{R}^p)$, is the unique minimizer of $\mathscr{E}^\lambda$.

The proof of Theorem 6 relies on two lemmas. The first quantifies the sensitivity of the Lasso problem (1) to the regularization parameter λ. The second quantifies the continuity of the minimizer of the objective function $\mathscr{E}^\lambda$ in the regularization parameter λ.

LEMMA B.12. *Assume $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse for the matrix $\boldsymbol{\Sigma}$. Then, under assumption A1, there exist constants $K, C_0, c_0 > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\Delta_{\min}$, M, and δ such that*

$$\mathbb{P}\left(\forall\lambda,\lambda' \in [\lambda_{\min},\lambda_{\max}],\; \mathcal{C}^{\lambda'}(\widehat{\boldsymbol{v}}^{\lambda}) \le \min_{\boldsymbol{v}\in\mathbb{R}^p}\mathcal{C}^{\lambda'}(\boldsymbol{v}) + K|\lambda-\lambda'|\right) \ge 1 - C_0 e^{-c_0 n}.$$

LEMMA B.13. *Assume $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse for the matrix $\boldsymbol{\Sigma}$. Then, under assumption A1, there exists constants $K, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\Delta_{\min}$, M, and δ such that for all $\lambda, \lambda' \in [\lambda_{\min}, \lambda_{\max}]$ with $|\lambda - \lambda'| < c'$ we have*

$$\text{for all } \lambda,\lambda' \in [\lambda_{\min},\lambda_{\max}], \qquad \frac{1}{\sqrt{p}}\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2} \le K|\lambda'-\lambda|^{1/2},$$

where in the previous display we view $\widehat{\boldsymbol{v}}^{f,\lambda}$, $\widehat{\boldsymbol{v}}^{f,\lambda'}$ as functions of the same random vector $\boldsymbol{g}$ and thus as members of $L^2(\mathbb{R}^p;\mathbb{R}^p)$.

The characterization of the Lasso solution involves only the distribution of $\widehat{\boldsymbol{v}}^{f,\lambda}$. The preceding lemma implicitly constructs a coupling between these distributions defined for different values of λ by using the same source of randomness $\boldsymbol{g}$ in defining $\widehat{\boldsymbol{v}}^{f,\lambda}$ and $\widehat{\boldsymbol{v}}^{f,\lambda'}$. We prove Lemma B.12 and B.13 in Sections B.7.1 and B.7.2 respectively.

To achieve a uniform control over $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, we invoke an ϵ-net argument. Because $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse and assumption A1 is satisfied, assumption A2 is satisfied for some $\mathcal{P}_{\text{fixPt}}$ depending only on $\mathcal{P}_{\text{model}}$, δ, $\nu_{\min}$, and $\Delta_{\min}$. Consider $\epsilon < c'$, where c' is as in Theorem 4. Let C_0, c_0 be as in Lemma B.12 and Lemma and let K_1, K_2 be the K's which appear in Lemma B.12 and Lemma B.13, respectively. Set $\epsilon' = \min\left\{\gamma\epsilon^2/K_1, \epsilon\kappa_{\min}^{1/2}/K_2\right\}$. Define $\lambda_i = \lambda_{\min} + i\epsilon'$ for $i = 1, \ldots, k$, $k := \lfloor \frac{\lambda_{\min}-\lambda_{\max}}{\epsilon'} \rfloor$ and $\lambda_{k+1} = \lambda_{\max}$.

By a union bound over λ_i, Theorem 4 implies that, for $C, c, c', \gamma > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ, with probability at least $1 - \frac{C(k+1)}{\epsilon^2}\exp(-cn\epsilon^4)$,

$$\forall \boldsymbol{v} \in \mathbb{R}^p,\ \forall \lambda_i, \quad \mathcal{C}_{\lambda_i}(\boldsymbol{v}) \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_{\lambda_i}(\boldsymbol{v}) + \gamma\epsilon^2 \ \Rightarrow\ \boldsymbol{v} \in (D_\epsilon^{\lambda_i})^c\,. \tag{93}$$

Further, Lemma B.12 implies that with probability at least $1 - C_0 e^{-c_0 n}$, the following occurs: for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$

$$\mathcal{C}_{\lambda_i}(\widehat{\boldsymbol{v}}^\lambda) \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_{\lambda_i}(\boldsymbol{v}) + K|\lambda - \lambda_i| \le \min_{\boldsymbol{v}\in\mathbb{R}^p} \mathcal{C}_{\lambda_i}(\boldsymbol{v}) + \gamma\epsilon^2,$$

where $i = i(\lambda)$ is chosen such that $\lambda \in [\lambda_i, \lambda_{i+1}]$ and the inequality holds by the choice of ϵ'. Combining with inequality (93), we conclude that

$$\text{for all } \lambda,\ \widehat{\boldsymbol{v}}^\lambda \in (D_\epsilon^{\lambda_i})^c \ \text{ where } i = i(\lambda) \text{ is such that } \lambda \in [\lambda_i, \lambda_{i+1}]\,.$$

Because ϕ is 1-Lipschitz,

$$\begin{aligned}\left|\mathbb{E}\tilde{\phi}\Big(\frac{\widehat{\boldsymbol{v}}^{f,\lambda}}{\sqrt{p}}\Big) - \mathbb{E}\tilde{\phi}\Big(\frac{\widehat{\boldsymbol{v}}^{f,\lambda_i}}{\sqrt{p}}\Big)\right| &\le \frac{1}{\sqrt{p}}\mathbb{E}\left[\|\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{v}}^{f,\lambda} - \widehat{\boldsymbol{v}}^f(\lambda_i)\|_2\right] \le \frac{\kappa_{\min}^{-1/2}}{\sqrt{p}}\|\widehat{\boldsymbol{v}}^{f,\lambda} - \widehat{\boldsymbol{v}}^{f,\lambda_i}\|_{L^2} \\ &\le K\kappa_{\min}^{-1/2}|\lambda_i - \lambda|^{1/2} \le \epsilon\end{aligned}$$

where the third-to-last inequality holds by Jensen's inequality, and the second-to-last inequality holds by Lemma B.13, and the last inequality holds by the choice of ϵ'. Note we have compared the two expectations on the left-hand side by constructing a coupling between the distribution of $\widehat{\boldsymbol{v}}^{f,\lambda}$ defined for different values of λ; see comment following Lemma B.13. By the triangle inequality, if $\widehat{\boldsymbol{v}}^\lambda \in (D_\epsilon^{\lambda_i})^c$, then $\widehat{\boldsymbol{v}}^\lambda \in (D_{2\epsilon}^{\lambda})^c$. Thus, we conclude that with $C, c, c' > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, and δ,

$$\mathbb{P}\left(\exists \lambda \in [\lambda_{\min}, \lambda_{\max}],\ \ \widehat{\boldsymbol{v}}^\lambda \in D_{2\epsilon}^\lambda\right) \ge 1 - \frac{C(k+1)}{\epsilon^2}e^{-cn\epsilon^4}.$$

For $\epsilon < c'$, we have $(k+1) \le C/\epsilon^2$ for some C depending only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, and δ. Absorbing constants appropriately, the proof of Theorem 6 is complete.

B.7.1. *Proof of Lemma B.12.*

LEMMA B.14. *Assume $\boldsymbol{x} \in \{-1, 0, 1\}^p$ is such that $\boldsymbol{x} \in \partial\|\bar{\boldsymbol{\theta}}^*\|_1$ with $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \le (1 - \Delta_{\min})\sqrt{\delta}$. Then there exist finite constants $a, c_0, C_0 > 0$ depending only on $\Delta_{\min}, \kappa_{\min}, \kappa_{\max}$, and δ such that if $n \ge \sqrt{2}/\Delta_{\min}$ the following happens with probability at least $1 - C_0 e^{-c_0 n}$. For any $\boldsymbol{w} \in \mathbb{R}^p$:*

$$\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{w}\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 \le 0 \ \Rightarrow\ \|\boldsymbol{X}\boldsymbol{w}\|_2 \ge a\|\boldsymbol{w}\|_2\,. \tag{94}$$

PROOF OF LEMMA B.14. The Gaussian width $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma})$ is an upper bound on the standard notion of Gaussian width $\mathcal{G}_{\text{std}}(\boldsymbol{x}, \boldsymbol{\Sigma})$ defined in Eq. (12). Thus,

$$(1 - \Delta_{\min})\sqrt{\delta} \ge \mathcal{G}_{\text{std}}(\boldsymbol{x}, \boldsymbol{\Sigma}) = \frac{1}{p} = \frac{1}{p}\mathbb{E}\Big[\max_{\substack{\boldsymbol{v}\in\mathcal{K}(\boldsymbol{x},\boldsymbol{\Sigma}) \\ \|\boldsymbol{v}\|_2^2/p\le 1}} \langle \boldsymbol{v}, \boldsymbol{g}\rangle\Big]\,. \tag{95}$$

The result then follows from standard results; see, for example, Corollary 3.3 of [16] and its proof. We repeat the proof here for convenience.

For simplicity of notation, we will denote $\mathcal{K} = \mathcal{K}(\boldsymbol{x}, \boldsymbol{\Sigma})$. Note that

$$\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{w}\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 = \|\boldsymbol{w}_{S^c}\|_1 + \|(\bar{\boldsymbol{\theta}}^* + \boldsymbol{w})_S\|_1 - \|\bar{\boldsymbol{\theta}}^*_S\|_1 \geq \|\boldsymbol{w}_{S^c}\|_1 + \sum_{j \in \mathrm{supp}(\boldsymbol{x})} x_j \boldsymbol{w}_j \,,$$

whence $\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{w}\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 \leq 0$ implies $\boldsymbol{\Sigma}\boldsymbol{w} \in \mathcal{K}$. Thus, it suffices to show that with probability at least $1 - C_0 e^{-c_0 n}$, one has

$$\boldsymbol{\Sigma}\boldsymbol{w} \in \mathcal{K} \;\Rightarrow\; \|\boldsymbol{X}\boldsymbol{w}\|_2 \geq a\|\boldsymbol{w}\|_2\,.$$

Define the minimum singular value over $\mathcal{K}$ as

$$\kappa_-(\boldsymbol{X}, \mathcal{K}) := \inf\left\{\|\boldsymbol{X}\boldsymbol{w}\|_2 \mid \boldsymbol{w} \in \mathcal{K}, \|\boldsymbol{w}\|_2 = 1\right\},$$

and define $\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K}) := \inf\left\{\|\boldsymbol{X}\boldsymbol{w}\|_2 \mid \boldsymbol{w} \in \mathcal{K}, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 = 1\right\}$. Then, because $\mathcal{K}$ is a cone (and so is scale invariant),

$$\kappa_-(\boldsymbol{X}, \mathcal{K}) \geq \tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K}) \cdot \min_{\|\boldsymbol{w}\|_2 = 1} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 \geq \tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})\kappa_{\min}^{1/2}\,.$$

Thus, it suffices to show there exists $a > 0$ depending on $\Delta_{\min}, \kappa_{\min}, \kappa_{\max}$, and δ such with high-probability $\kappa_-(\boldsymbol{X}, \mathcal{K}) \geq a$.

By definition,

$$-\mathbb{E}[\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})] = \mathbb{E}[\max_{\substack{\boldsymbol{w} \in \mathcal{K} \\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 = 1}} -\|\boldsymbol{X}\boldsymbol{w}\|_2] = \mathbb{E}[\max_{\substack{\boldsymbol{w} \in \mathcal{K} \\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 = 1}} \min_{\|\boldsymbol{u}\|_2 = 1} \boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{w}]\,.$$

Recall that the rows of $\boldsymbol{X}$ are distributed iid from $\mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}/n)$. By Gordon's lemma (Corollary G.1 of [35])

$$\begin{aligned}
-\mathbb{E}[\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})] &\leq \mathbb{E}\Big[\max_{\substack{\boldsymbol{w} \in \mathcal{K} \\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 = 1}} \min_{\|\boldsymbol{u}\|_2 = 1} \frac{1}{\sqrt{n}}\|\boldsymbol{\Sigma}\boldsymbol{w}\|_2\langle \boldsymbol{h}, \boldsymbol{u}\rangle + \frac{1}{\sqrt{n}}\|\boldsymbol{u}\|_2\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}, \boldsymbol{g}\rangle\Big] \\
&= \frac{1}{\sqrt{n}}\mathbb{E}\Big[\max_{\substack{\boldsymbol{w} \in \mathcal{K} \\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{w}\|_2 = 1}} -\|\boldsymbol{h}\|_2 + \langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}, \boldsymbol{g}\rangle\Big] \leq -\sqrt{\frac{n}{n+1}} + \sqrt{\frac{p}{n}}\widetilde{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{\Sigma}) \\
&\leq -\sqrt{\frac{n}{n+1}} + 1 - \Delta_{\min}\,,
\end{aligned}$$

where the second-to-last equality uses $\mathbb{E}[\|\boldsymbol{h}\|_2] \geq \frac{n}{\sqrt{n+1}}$ and the definition of $\widetilde{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{\Sigma})$, and the last inequality uses the upper bound on the Gaussian width (95). For all $n \geq 2$ we have $\sqrt{n/(n+1)} \geq \sqrt{(n-1)/n} \geq 1 - \frac{1}{\sqrt{2}n}$. Thus, for $n \geq \sqrt{2}/\Delta_{\min}$, $\mathbb{E}[\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})] \geq \Delta_{\min}/2$.

The quantity $\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})$ as a function of $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ is $\frac{1}{\sqrt{n}}$-Lipschitz with respect to the Frobenius norm. Thus

$$\mathbb{P}(\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K}) \leq \mathbb{E}[\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})] - t) \leq e^{-nt^2/2}.$$

Taking $t = \Delta_{\min}/4$ and using $\mathbb{E}[\tilde{\kappa}_-(\boldsymbol{X}, \mathcal{K})] \geq \Delta_{\min}/2$ gives $\mathbb{P}(\kappa_-(\boldsymbol{X}, \mathcal{K}) \leq \kappa_{\min}^{1/2}\Delta_{\min}/4) \leq e^{-n\Delta_{\min}^2/32}$. The proof of inequality (94) is complete. $\square$

LEMMA B.15. *Assume $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1 - \Delta_{\min}), M)$-approximately sparse for the matrix $\boldsymbol{\Sigma}$. If $n \geq \sqrt{2}/\Delta_{\min}$, then there exist constants $C, C_0, c_0 > 0$ depending only on $\sigma, \Delta_{\min}, \kappa_{\min}, \kappa_{\max}, \lambda_{\min}, \lambda_{\max}$, and δ such that*

$$\mathbb{P}\left(\forall \lambda \in [\lambda_{\min}, \lambda_{\max}] : \frac{1}{n}\left|\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\boldsymbol{\theta}^*\|_1\right| \leq C\right) \geq 1 - C_0 e^{-c_0 n}\,. \tag{96}$$

PROOF OF LEMMA B.15. The proof follows almost exactly that for [35, Proposition C.4], using Lemma B.14. The primary difference is the approximation of $\boldsymbol{\theta}^*$ by $\bar{\boldsymbol{\theta}}^*$.

Because $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse, there exists $\bar{\boldsymbol{\theta}}^* \in \mathbb{R}^p$ and $\boldsymbol{x} \in \{-1,0,1\}^p$ such that $\frac{1}{p}\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1 \le M$, $\boldsymbol{x} \in \partial\|\bar{\boldsymbol{\theta}}^*\|_1$, and $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \le \sqrt{\delta}(1-\Delta_{\min})$. Note that

$$\begin{aligned}\frac{\lambda}{n}\Big(\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\boldsymbol{\theta}^*\|_1\Big) &\ge \frac{\lambda}{n}\Big(\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 - 2\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1\Big)\\ &\ge \frac{\lambda}{n}\Big(\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1\Big) - 2\lambda M\,.\end{aligned}$$

We show that the high probability event (96) is implied by the event

$$\mathcal{A} := \big\{\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{w}\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 \le 0 \;\Rightarrow\; \|\boldsymbol{X}\boldsymbol{w}\|_2 \ge a\|\boldsymbol{w}\|_2\big\} \bigcap \big\{\|\boldsymbol{z}\|_2 \le 2\sqrt{n}\big\}\,,$$

where a is as in Lemma B.14. On this event, $\mathcal{C}(\widehat{\boldsymbol{v}}^\lambda) \le \mathcal{C}_\lambda(\mathbf{0}) = \sigma^2\|\boldsymbol{z}\|_2^2/(2n) \le 2\sigma^2$, whence

$$\frac{1}{n}\Big(\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1\Big) \le 2\sigma^2/\lambda_{\min} + 2M\,,$$

which further implies

$$\frac{1}{n}\Big(\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\boldsymbol{\theta}^*\|_1\Big) \le 2\sigma^2/\lambda_{\min} + 4M\,. \tag{97}$$

Let $\widehat{\boldsymbol{w}}^\lambda = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda$. On the event $\mathcal{A}$, we also have

$$\begin{aligned}2\sigma^2 \ge \mathcal{C}(\widehat{\boldsymbol{v}}^\lambda) &\ge \frac{1}{2n}\|\sigma\boldsymbol{z} - \boldsymbol{X}\widehat{\boldsymbol{w}}^\lambda\|_2^2 - \frac{\lambda}{n}\|\widehat{\boldsymbol{w}}^\lambda\|_1 - 2\lambda M\\ &\ge \frac{1}{4n}\|\boldsymbol{X}\widehat{\boldsymbol{w}}^\lambda\|_2^2 - \frac{\sigma^2}{2n}\|\boldsymbol{z}\|_2^2 - \frac{\lambda}{\sqrt{\delta n}}\|\widehat{\boldsymbol{w}}^\lambda\|_2 - 2\lambda M\\ &\ge \frac{a}{4n}\|\widehat{\boldsymbol{w}}^\lambda\|_2^2 - 4\sigma^2 - \frac{\lambda\sqrt{p}}{n}\|\widehat{\boldsymbol{w}}^\lambda\|_2 - 2\lambda M\,.\end{aligned}$$

We conclude that

$$\frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{w}}^\lambda\|_2 \le 2\sqrt{2\sigma^2 + 2\lambda M + \frac{\lambda^2}{a\delta}} \le C(1+\lambda+M),$$

for C depending only on σ, a, δ, so in fact only on $\sigma, \Delta_{\min}, \delta, \kappa_{\min}, \kappa_{\max}$. Then

$$-C\delta^{-1/2}(1+\lambda_{\max}+M) \le -\frac{1}{n}\|\widehat{\boldsymbol{w}}^\lambda\|_1 \le \frac{1}{n}(\|\boldsymbol{\theta}^* + \widehat{\boldsymbol{w}}^\lambda\|_1 - \|\boldsymbol{\theta}^*\|_1). \tag{98}$$

The event $\mathcal{A}$ has probability at least $1 - C_0 e^{-c_0 n}$ by Lemma B.14 and concentration of Lipschitz functions of Gaussian vectors, where C_0, c_0 depend only δ, $\Delta_{\min}$, $\kappa_{\min}$, and $\kappa_{\max}$. Lemma B.15 now follows by combining (97) and (98). ☐

Lemma B.12 follows from Lemma B.15 by exactly the same argument in the proof of [35, Lemma C.5].

B.7.2. *Proof of Lemma B.13.* Recall from the proof of Lemma A.2 (in particular, Eq. (42)) that $\widehat{\boldsymbol{v}}^{f,\lambda}$, where the latter is viewed as a function of $\boldsymbol{g}$ and hence an element of $L^2(\mathbb{R}^p;\mathbb{R}^p)$, is the unique minimizer of $\mathscr{E}^\lambda$. By optimality,

$$\frac{\sigma^2}{2} = \mathscr{E}^\lambda(\mathbf{0}) \ge \mathscr{E}^\lambda(\widehat{\boldsymbol{v}}^{f,\lambda}) \ge \frac{\lambda}{n}\mathbb{E}\Big\{\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{f,\lambda}(\boldsymbol{g})\|_1 - \|\boldsymbol{\theta}^*\|_1\,,\Big\}\,. \tag{99}$$

We now find also a lower bound on the right-hand side. Because $\boldsymbol{\theta}^*$ is $(s, \sqrt{\delta}(1-\Delta_{\min}), M)$-approximately sparse, there exists $\bar{\boldsymbol{\theta}}^* \in \mathbb{R}^p$ and $\boldsymbol{x} \in \{-1,0,1\}^p$ such that $\frac{1}{p}\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1 \le M$, $\boldsymbol{x} \in \partial\|\bar{\boldsymbol{\theta}}^*\|_1$, and $\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma}) \le \sqrt{\delta}(1-\Delta_{\min})$. Note that

$$\begin{aligned}\frac{\lambda}{n}\mathbb{E}\left\{\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\boldsymbol{\theta}^*\|_1\right\} &\ge \frac{\lambda}{n}\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1 - 2\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1\right\} \\ &\ge \frac{\lambda}{n}\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1\right\} - 2\lambda M\,.\end{aligned}$$

By the definition of Gaussian width (see Eq. (10)), either

$$\frac{\lambda}{n}\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^\lambda\|_1 - \|\bar{\boldsymbol{\theta}}^*\|_1\right\} \ge 0\,,$$

or

$$\begin{aligned}\frac{\sigma^2}{2} = \mathscr{E}^\lambda(\mathbf{0}) \ge \mathscr{E}^\lambda(\widehat{\boldsymbol{v}}^{f,\lambda}) &\ge \frac{1}{2}\left(\frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}}{\sqrt{n}} - \frac{\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma})\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}}{\sqrt{n\delta}}\right)_+^2 - \frac{\lambda}{n}\mathbb{E}\left\{\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{f,\lambda}(\boldsymbol{g})\|_1\right\} - 2\lambda M \\ &\ge \frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}^2}{2n}\Delta_{\min}^2 - \frac{\lambda}{\sqrt{n\delta}}\mathbb{E}\left\{\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{f,\lambda}(\boldsymbol{g})\|_2\right\} - 2\lambda M \\ &\ge \frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}^2}{2n}\Delta_{\min}^2 - \frac{\lambda\kappa_{\max}}{\sqrt{\delta}}\frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}}{\sqrt{n}} - 2\lambda M\,.\end{aligned}$$

In the latter case, we conclude

$$\frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}}{\sqrt{n}} \le 2\sqrt{2\lambda M + \frac{\sigma^2}{2} + \frac{\lambda^2\kappa_{\max}^2}{2\delta\Delta_{\min}^2}}\,.$$

Thus, in this case

(100)

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left\{\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{f,\lambda}\|_1 - \|\boldsymbol{\theta}^*\|_1\right\} &\ge -\frac{\lambda\kappa_{\max}}{\sqrt{\delta}}\frac{\|\widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}}{\sqrt{n}} - 2\lambda M \\ &\ge -\frac{\kappa_{\max}}{\sqrt{\delta}}2\sqrt{2\lambda_{\max}M + \frac{\sigma_{\max}^2}{2} + \frac{\lambda_{\max}^2\kappa_{\max}^2}{2\delta\Delta_{\min}^2}} - 2M\,.\end{aligned}$$

Combining Eqs. (99) and (100), there exists C depending only on $\mathcal{P}_{\text{model}}$, $\Delta_{\min}$, M, and δ such that

$$\left|\frac{1}{n}\mathbb{E}\left\{\|\boldsymbol{\theta}^* + \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{v}}^{f,\lambda}\|_1 - \|\boldsymbol{\theta}^*\|_1\right\}\right| \le C\,.$$

By Theorem 2, the solutions to the fixed point equations (8a) and (8b) are bounded by some parameters $\mathcal{P}_{\text{fixPt}} = (\zeta_{\min}, \zeta_{\max}, \tau_{\min}, \tau_{\max})$ which depend only on $\mathcal{P}_{\text{model}}$, $\nu_{\min}$, $\Delta_{\min}$, M, and δ. By the proof of Lemma A.5, there exists $r, a > 0$ depending only on $\mathcal{P}_{\text{model}}$, $\mathcal{P}_{\text{fixPt}}$, and δ such that $\mathscr{E}^\lambda$ is a/n strongly-convex in the neighborhood $\|\boldsymbol{v} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n} \le r$ around its minimizer. Thus, we conclude for any $\boldsymbol{v} \in L^2$ we have

$$\mathscr{E}^\lambda(\boldsymbol{v}) \ge \mathscr{E}^\lambda(\widehat{\boldsymbol{v}}^{f,\lambda}) + h(\|\boldsymbol{v} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n})$$

where $h(x) := \min\{ax^2/2, ar|x|/2\}$. It worth emphasizing that this bound holds for with the same a, r for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Then direct calculations further give

$$\begin{aligned}\mathscr{E}^\lambda(\widehat{\boldsymbol{v}}^{f,\lambda'}) &\ge \mathscr{E}^\lambda(\widehat{\boldsymbol{v}}^{f,\lambda}) + h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) \ge \mathscr{E}^{\lambda'}(\widehat{\boldsymbol{v}}^{f,\lambda}) + h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) - C|\lambda' - \lambda| \\ &\ge \mathscr{E}^{\lambda'}(\widehat{\boldsymbol{v}}^{f,\lambda'}) + 2h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) - C|\lambda' - \lambda| \\ &\ge \mathscr{E}^{\lambda}(\widehat{\boldsymbol{v}}^{f,\lambda}) + 4h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) - 2C|\lambda' - \lambda|\,,\end{aligned}$$

where the last inequality holds by the same string of manipulations justifying the first three. Take $c' = ar^2/C$. If $|\lambda - \lambda'| < c'$, we have $h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) \leq ar^2/2$, whence in fact $h(\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{n}) = \frac{a\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}^2}{2n}$. We conclude $\|\widehat{\boldsymbol{v}}^{f,\lambda'} - \widehat{\boldsymbol{v}}^{f,\lambda}\|_{L^2}/\sqrt{p} \leq \sqrt{\frac{C\delta}{a}}|\lambda - \lambda'|^{1/2}$. Therefore the claimed result holds with $K = \sqrt{\frac{C\delta}{a}}$.

B.8. *Control of the emprical distribution: proof of Corollary 7.* In the fixed design model, let I be uniformly distributed on $[p]$ independently of $\boldsymbol{g}$. Let μ_* be the distribution of $(\theta_I^*, \widehat{\theta}_I^f)$. For any k, we have $\frac{1}{p}\sum_{i=1}^p \phi_k(\theta_i^*, \widehat{\theta}_i^f)$ is $\tau\kappa_{\min}^{-1/2}/\sqrt{p}$-Lipschitz in $\boldsymbol{g}$, so that by Gaussian concentration of Lipschitz functions,

$$\mathbb{P}\left(\Big|\frac{1}{p}\sum_{i=1}^p \phi_k(\theta_i^*, \widehat{\theta}_i^f) - \mathbb{E}\Big[\frac{1}{p}\sum_{i=1}^p \phi_k(\theta_i^*, \widehat{\theta}_i^f)\Big]\Big| > t\right) \leq 2e^{-2p\kappa_{\min}t^2/\tau_{\max}^2},$$

whence $\mathbb{E}\left[\Big|\frac{1}{p}\sum_{i=1}^p \phi_k(\theta_i^*, \widehat{\theta}_i^f) - \mathbb{E}\Big[\frac{1}{p}\sum_{i=1}^p \phi_k(\theta_i^*, \widehat{\theta}_i^f)\Big]\Big|\right] \leq C/\sqrt{p}$, for some C depending on $\mathcal{P}_{\text{model}}, \mathcal{P}_{\text{fixPt}}$. Summing the above inequality over $k = 1, \ldots, \infty$, we obtain that $\mathbb{E}\left[d_{w^*}\left(\frac{1}{p}\sum_{i=1}^p \delta_{\theta_i^*, \widehat{\theta}_i^f}, \mu_*\right)\right] \leq C/\sqrt{p}$. Note further that $d_{w^*}\left(\frac{1}{p}\sum_{i=1}^p \delta_{\theta_i^*, \widehat{\theta}_i^f}, \mu_*\right)$ is $\tau_{\max}\kappa_{\min}^{-1/2}/\sqrt{p}$-Lipschitz in $\boldsymbol{g}$. Applying Gaussian Lipschitz concentration in the fixed design model, we conclude the second inequality in Corollary 7. In addition, applying Theorem 6, we conclude the first inequality in Corollary 7.

C. Auxiliary results and proofs.

C.1. *Gaussian width under correlated designs: proof of Proposition 3.*

PROOF OF PROPOSITION 3. Let us first establish the following relation

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\Sigma}) \leq \kappa_{\text{cond}}^{1/2}\omega^*(\|\boldsymbol{x}\|_0/p)\,. \tag{101}$$

To start with, notice that

$$\sup_{\substack{\boldsymbol{v}\in\mathcal{D}(\boldsymbol{x},\boldsymbol{\Sigma})\\ \|\boldsymbol{v}\|_{L^2}^2/p\leq 1}} \frac{1}{p}\langle \boldsymbol{v}, \boldsymbol{g}\rangle_{L^2} = \sup_{\substack{\boldsymbol{w}\in\mathcal{D}(\boldsymbol{x},\mathbf{I}_p)\\ \|\boldsymbol{w}\|_{L^2}^2/p\leq 1}} \frac{\|\boldsymbol{w}\|_{L^2}\langle \boldsymbol{\Sigma}\boldsymbol{w}, \boldsymbol{g}\rangle_{L^2}}{p\|\boldsymbol{\Sigma}\boldsymbol{w}\|_{L^2}} \leq \frac{\kappa_{\max}^{1/2}}{\kappa_{\min}^{1/2}} \sup_{\substack{\boldsymbol{w}\in\mathcal{D}(\boldsymbol{x},\mathbf{I}_p)\\ \|\boldsymbol{w}\|_{L^2}^2/p\leq 1}} \frac{\langle \boldsymbol{\Sigma}\boldsymbol{w}, \boldsymbol{g}\rangle_{L^2}}{p\kappa_{\max}^{1/2}}, \tag{102}$$

where in the equality we have used that $\boldsymbol{w} \leftrightarrow \|\boldsymbol{w}\|_{L^2}\boldsymbol{\Sigma}\boldsymbol{w}/\|\boldsymbol{\Sigma}\boldsymbol{w}\|_{L^2}$ is a bijection between the sets over which the suprema are taken, and in the inequality we have used that the supremum is positive (because $\boldsymbol{w} = \mathbf{0}$ is feasible) and $\|\boldsymbol{w}\|_{L^2}/\|\boldsymbol{\Sigma}\boldsymbol{w}\|_{L^2} \geq \kappa_{\min}^{1/2}$. The Lagrangian for the maximization on the right-hand side is

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\Sigma}}(\boldsymbol{w}; \kappa, \xi) &:= \frac{1}{p\kappa_{\max}^{1/2}}\mathbb{E}\big[\boldsymbol{w}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}\big] + \frac{\kappa}{2}\left(1 - \frac{1}{p}\mathbb{E}\left[\|\boldsymbol{w}\|_2^2\right]\right) - \frac{\xi}{p}\mathbb{E}\Big[\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\Big] \\
&= \frac{\kappa}{2} + \frac{1}{p}\mathbb{E}\left[\frac{\boldsymbol{w}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}}{\kappa_{\max}^{1/2}} - \frac{\kappa}{2}\|\boldsymbol{w}\|_2^2 - \xi\left(\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\right)\right].
\end{aligned}$$

The optimal $\boldsymbol{w} \in L^2$ maximizes the integrand for almost every $\boldsymbol{g}$, whence

$$\sup_{\boldsymbol{w}\in L^2} \mathcal{L}_{\boldsymbol{\Sigma}}(\boldsymbol{w}; \kappa, \xi) = \frac{\kappa}{2} + \frac{1}{p}\mathbb{E}\left[\sup_{\boldsymbol{w}\in\mathbb{R}^p}\left\{\frac{\boldsymbol{w}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}}{\kappa_{\max}^{1/2}} - \frac{\kappa}{2}\|\boldsymbol{w}\|_2^2 - \xi\left(\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\right)\right\}\right].$$

We emphasize that the dummy variable $\boldsymbol{w}$ is in L^2 on the left-hand side and $\mathbb{R}^p$ on the right-hand side. We apply the Sudakov-Fernique inequality [1, Theorem 2.2.3] to upper bound the expectation in the preceding display. Indeed, for $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^p$, we have $\mathbb{E}[\boldsymbol{w}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}/\kappa_{\max}^{1/2}] = \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{g}] = 0$ and $\mathrm{Var}((\boldsymbol{w}-\boldsymbol{w}')^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}/\kappa_{\max}^{1/2}) \leq \mathrm{Var}((\boldsymbol{w}-\boldsymbol{w}')^\top \boldsymbol{g})$ because $\|\boldsymbol{\Sigma}^{1/2}/\kappa_{\max}^{1/2}\|_{\mathrm{op}} \leq 1$. Thus, the Sudakov-Fernique inequality implies

$$\sup_{\boldsymbol{w}\in L^2} \mathcal{L}_{\boldsymbol{\Sigma}}(\boldsymbol{w};\kappa,\xi) \leq \frac{\kappa}{2} + \frac{1}{p}\mathbb{E}\left[\sup_{\boldsymbol{w}\in\mathbb{R}^p}\left\{\boldsymbol{w}^\top\boldsymbol{g} - \frac{\kappa}{2}\|\boldsymbol{w}\|_2^2 - \xi\left(\sum_{j\in S} x_j w_j + \|\boldsymbol{w}_{S^c}\|_1\right)\right\}\right] = \sup_{\boldsymbol{w}\in L^2}\mathcal{L}_{\mathbf{I}_p}(\boldsymbol{w};\kappa,\xi).$$

For any $\kappa, \xi \geq 0$, $\sup_{\boldsymbol{w}\in L^2}\mathcal{L}_{\boldsymbol{\Sigma}}(\boldsymbol{w};\kappa,\xi) \geq \sup_{\substack{\boldsymbol{w}\in\mathcal{D}(\boldsymbol{x},\mathbf{I}_p)\\ \|\boldsymbol{w}\|_{L^2}^2/p\leq 1}} \frac{\langle \boldsymbol{\Sigma}\boldsymbol{w},\boldsymbol{g}\rangle_{L^2}}{p\kappa_{\max}^{1/2}}$, whence, by Eq. (102),

$$\mathcal{G}(\boldsymbol{x},\boldsymbol{\Sigma}) = \sup_{\substack{\boldsymbol{v}\in\mathcal{D}(\boldsymbol{x},\boldsymbol{\Sigma})\\ \|\boldsymbol{v}\|_{L^2}^2/p\leq 1}} \frac{1}{p}\langle \boldsymbol{v},\boldsymbol{g}\rangle_{L^2} \leq \kappa_{\mathrm{cond}}^{1/2}\sup_{\boldsymbol{w}\in L^2}\mathcal{L}_{\mathbf{I}_p}(\boldsymbol{w};\kappa,\xi).$$

Note that $\mathcal{L}_{\mathbf{I}_p}(\boldsymbol{w};\kappa,\xi)$ is the Lagrangian for the optimization Eq. (10) defining $\mathcal{G}(\boldsymbol{x},\mathbf{I}_p)$. Because the constraints on $\boldsymbol{w}$ in this optimization are strictly feasible, strong duality holds. Thus, Eq. (101) follows by taking the infimum over $\kappa,\xi\geq 0$ in the preceding display.

Parts (a) and (b) of Proposition 3 now follow from the following constructions.

(a) If $\|\boldsymbol{\theta}^*\|_q^q/p \leq \nu^q$ for some $\nu > 0$ and $q > 0$, take $\boldsymbol{x}$ to be supported on the largest $\lfloor p\varepsilon^*(\kappa_{\mathrm{cond}},\delta/2)\rfloor$ coordinates of $\boldsymbol{\theta}^*$ and take $\bar{\theta}_j^* = \theta_j^*$ for $j\in\mathrm{supp}(\boldsymbol{x})$ and 0 otherwise. We have an upper bound on the Gaussian width of $\sqrt{\delta/2}$ and an upper bound on $\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1/p$ of $M = \nu(1-\varepsilon^*(\kappa_{\mathrm{cond}},\delta/2))$.

(b) If $\|\boldsymbol{\theta}^*\|_0/p \leq \varepsilon^*(\kappa_{\mathrm{cond}},\alpha)$ for some $\alpha<\delta$, take $\boldsymbol{x}$ to have support size $\lfloor p\varepsilon^*(\kappa_{\mathrm{cond}},\alpha)\rfloor$ with support containing $\mathrm{supp}(\boldsymbol{\theta}^*)$[8] and take $\bar{\boldsymbol{\theta}}^* = \boldsymbol{\theta}^*$. We have an upper bound on the Gaussian width of $\sqrt{\alpha}$ and an upper bound on $\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_1/p$ of $M=0$,

The proof of Proposition 3 is complete. □

C.2. *Properties of the design matrix.* Given every integer $j\in\{1,\dots,p\}$, each row of our design matrix is sampled independently from a multivariate Gaussian distribution, namely

$$\text{for } i=1,\dots,n \quad (X_{i,j},\boldsymbol{X}_{i.-j}) \sim \mathsf{N}\left(0,\frac{1}{n}\boldsymbol{\Sigma}\right) \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\Sigma_{j,j} & \boldsymbol{\Sigma}_{j,-j}\\ \boldsymbol{\Sigma}_{-j,j} & \boldsymbol{\Sigma}_{-j,-j}\end{pmatrix},$$

where the $1/n$ factor is due to the normalization of the design matrix. Here $\boldsymbol{X}_{.,j}$ stands for the covariate corresponding to the j-th coordinate of $\boldsymbol{\theta}$ and $\boldsymbol{X}_{.,-j}\in\mathbb{R}^{(p-1)}$ stands for covariates corresponding to rest of $\boldsymbol{\theta}$.

Let us further define $X_j^\perp := X_j - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{X}_{-j}$ for every $j\in\{1,\dots,p\}$ and the sampled version $\breve{\boldsymbol{x}}^\perp := \boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}\in\mathbb{R}^n$. Then the linear model can be written as

$$\boldsymbol{y} = \breve{\boldsymbol{x}}^\perp\theta_j^* + \boldsymbol{X}_{-j}(\boldsymbol{\theta}_{-j}^* + \theta_j^*\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}) + \sigma\boldsymbol{z}.$$

In addition, we state without proof the following straightforward properties.

- $X_j|\boldsymbol{X}_{-j} \sim \mathsf{N}(\boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{X}_{-j},\ \frac{1}{n}(\Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}))$.
- $X_j^\perp|\boldsymbol{X}_{-j} \sim \mathsf{N}(0,\ \frac{1}{n}(\Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}))$.
- $X_j^\perp \sim \mathsf{N}(0,\ \frac{1}{n}(\Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}))$.
- The entries of $\breve{\boldsymbol{x}}^\perp$ are i.i.d with distribution $\mathsf{N}(0,\ \frac{1}{n}(\Sigma_{j,j} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}))$.

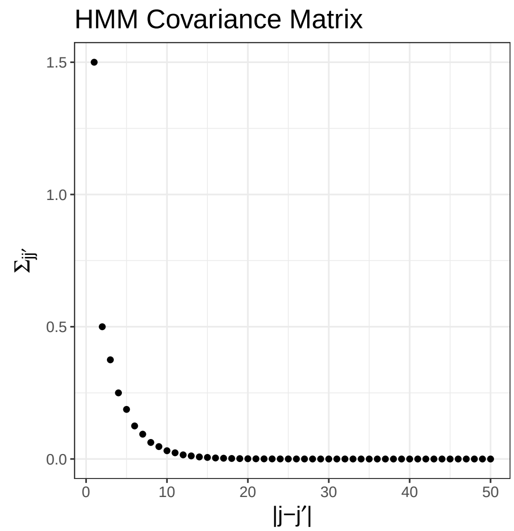
[8]Note it is important that $\boldsymbol{x}$ have support large enough, even if $\boldsymbol{\theta}^*$ is sparser than $\varepsilon^*(\kappa_{\mathrm{cond}},\alpha)$.

D. Additional Simulations.

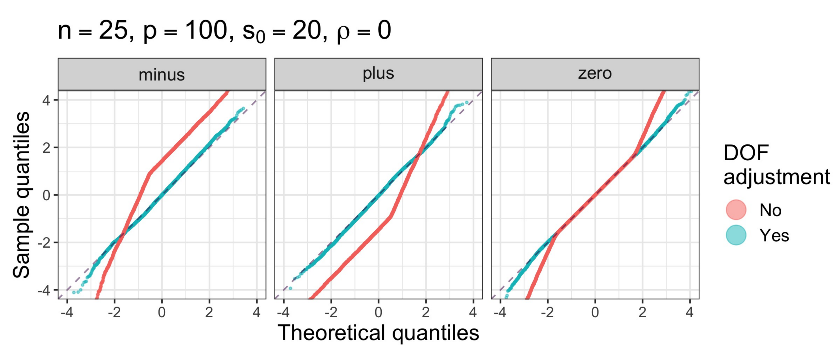
D.1. *Hidden Markov model specification.* In the hidden Markov model, the covariates x_{ij} are conditionally independent given latent states s_{ij} which are generated according to a Markov chain. In particular, the distribution satisfies $\mathbb{P}(s_{i(j+1)} \mid \{s_{i\ell}\}_{\ell \le j}) = \mathbb{P}(s_{i(j+1)} \mid s_{ij})$. The latent states (values for s_{ij}) and observed values (values of x_{ij}) in the hidden Markov model that we consider here, take values in $\{1,2,3,4,5\}$. Both the transition and emission probabilities are given by a symmetric random walk with reflection at the boundary; that is,

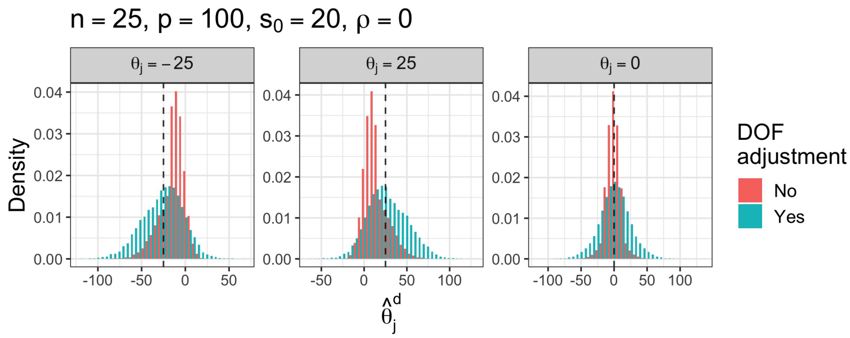
$$\mathbb{P}(s_{i(j+1)} = a | s_{ij} = b) = \mathbb{P}(x_{ij} = a | s_{ij} = b) = \begin{cases} 1/2 & b \in \{2,3,4\} \text{ and } |a-b| = 1, \\ 1 & b \in \{1,2\} \text{ and } |a-b| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

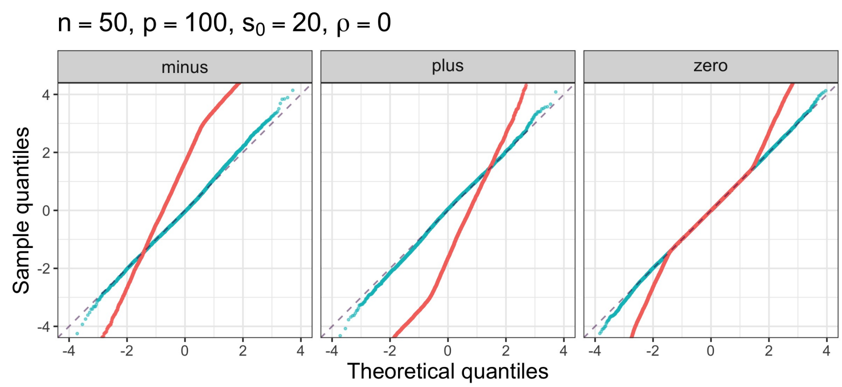
We initialize this Markov chain (i.e., s_{i1}) from its stationary distribution. In this case, the covariance of x_{ij} and $x_{ij'}$ is only a function of $|j-j'|$, as plotted below. We see that covariates which are within approximately distance 10 of each other have non-trivial correlation.

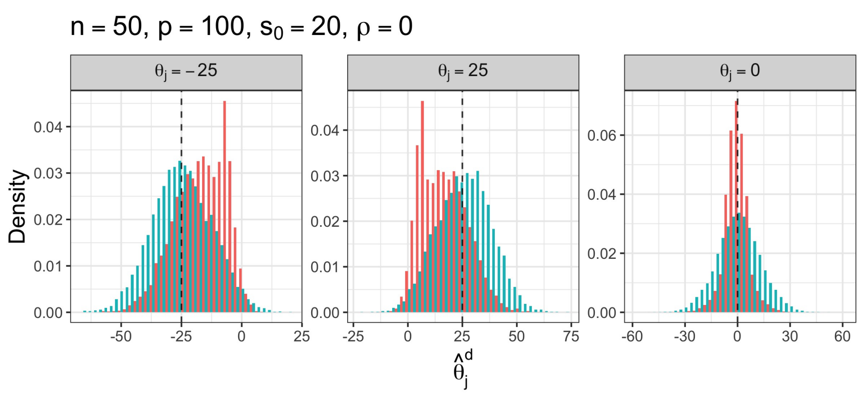


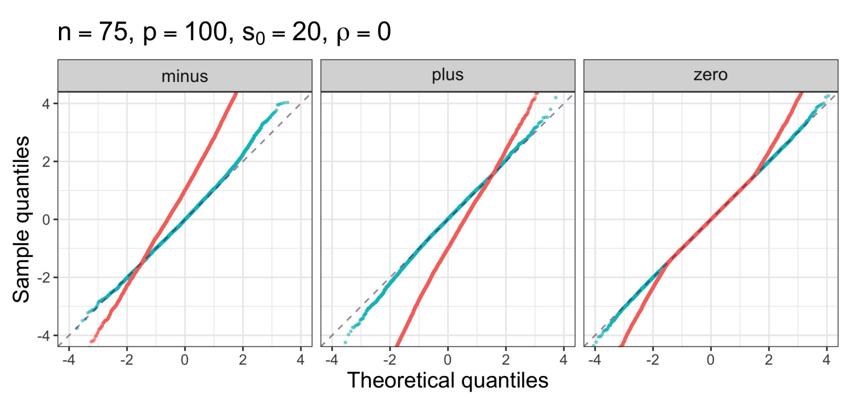
D.2. *Debiasing under Gaussian AR1 models.* Here we collect simulations which repeat those in Figure 1 at different model parameters. These simulations demonstrate the success of debiasing at many settings of the model parameters. In particular, we run the simulations varying the correlation parameter $\rho = 0, .5, .8$ and the sample size $n = 25, 50, 75, 100$. We show the legend for the first two plots. The legend for the remaining plots is the same.

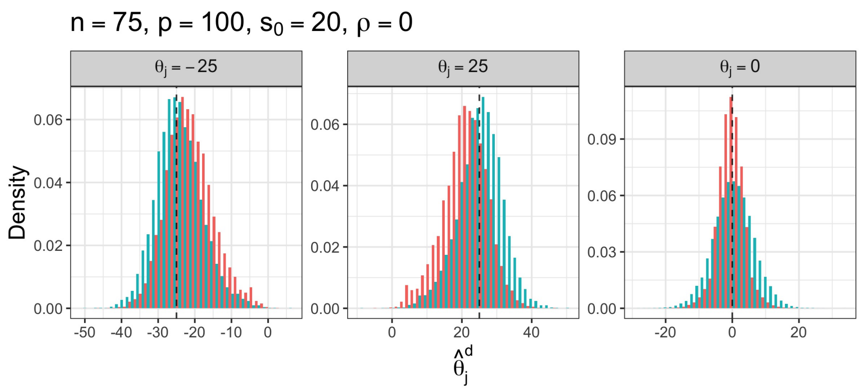


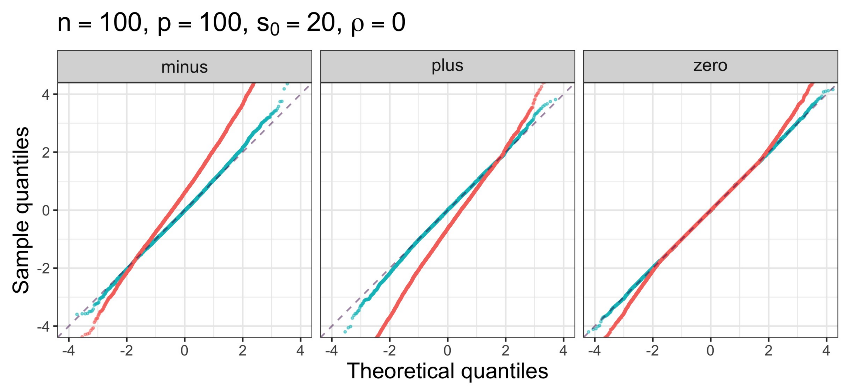

n = 25, p = 100, s0 = 20, ρ = 0
θj = −25
θj = 25
θj = 0
Density
DOF adjustment
No
Yes

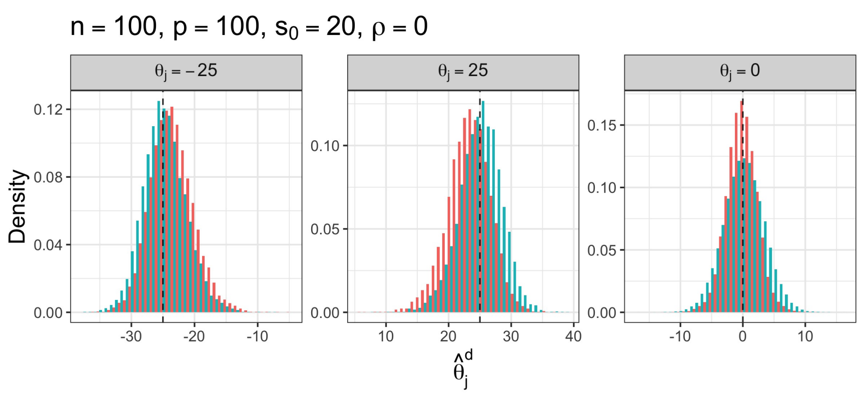

n = 50, p = 100, s0 = 20, ρ = 0
minus
plus
zero
Sample quantiles
Theoretical quantiles

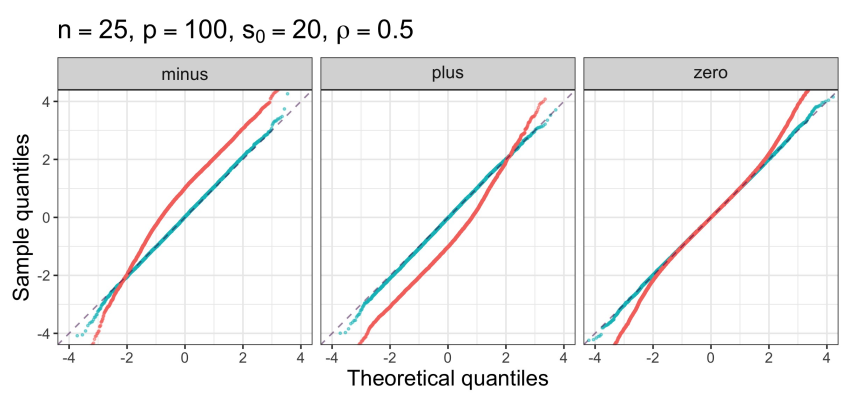

n = 50, p = 100, s0 = 20, ρ = 0
θj = −25
θj = 25
θj = 0
Density

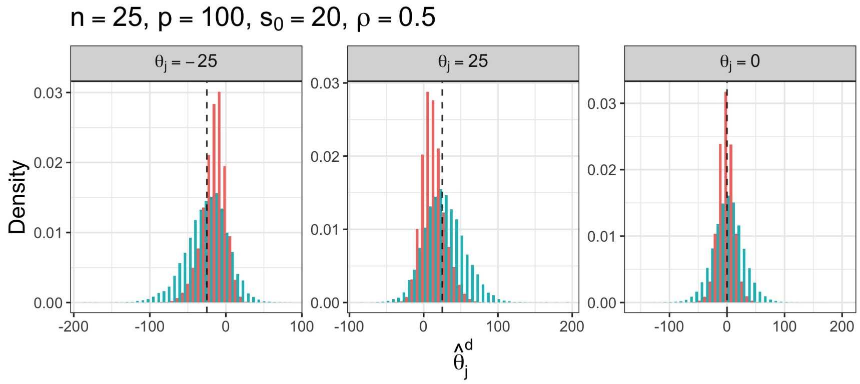

n = 75, p = 100, s0 = 20, ρ = 0
minus
plus
zero
Sample quantiles
Theoretical quantiles

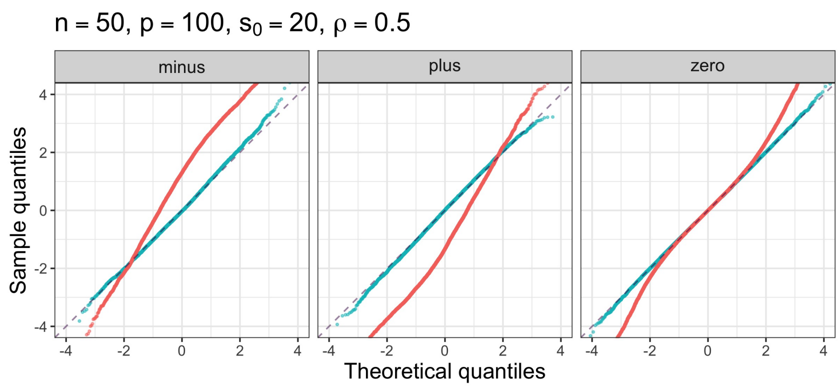

n = 75, p = 100, s0 = 20, ρ = 0
θj = −25
θj = 25
θj = 0
Density
θ̂j d

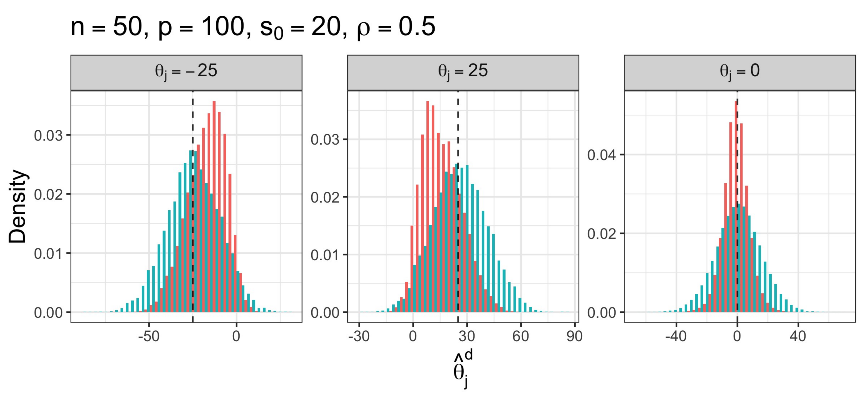

n = 100, p = 100, s0 = 20, ρ = 0
minus
plus
zero
Sample quantiles
Theoretical quantiles

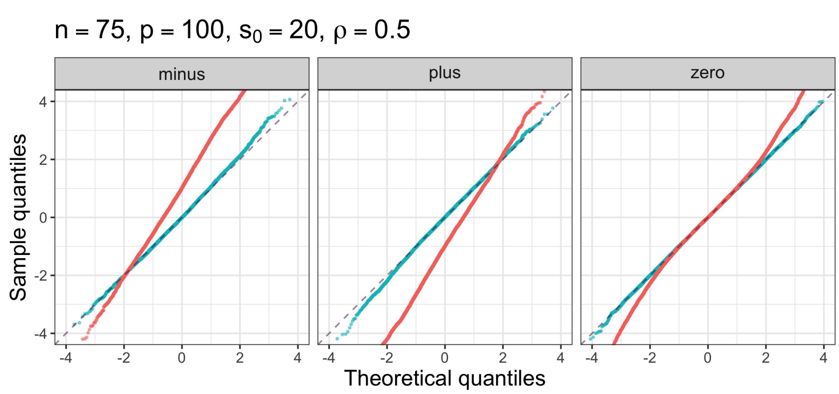

n = 100, p = 100, s0 = 20, ρ = 0
θj = −25
θj = 25
θj = 0
Density
θ̂j d

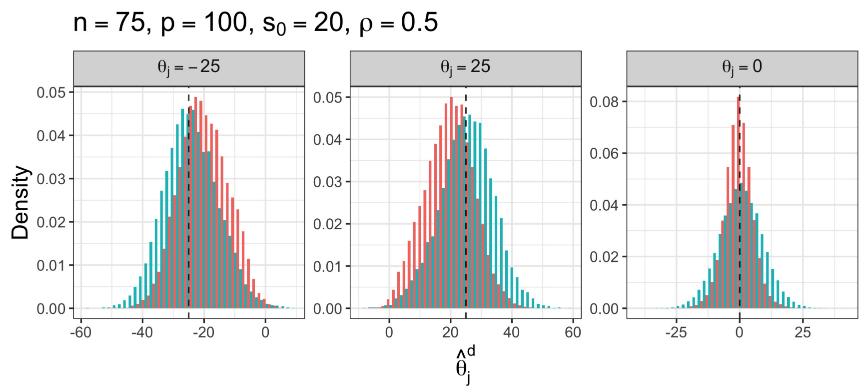

n = 25, p = 100, s0 = 20, ρ = 0.5
minus
plus
zero
Sample quantiles
Theoretical quantiles

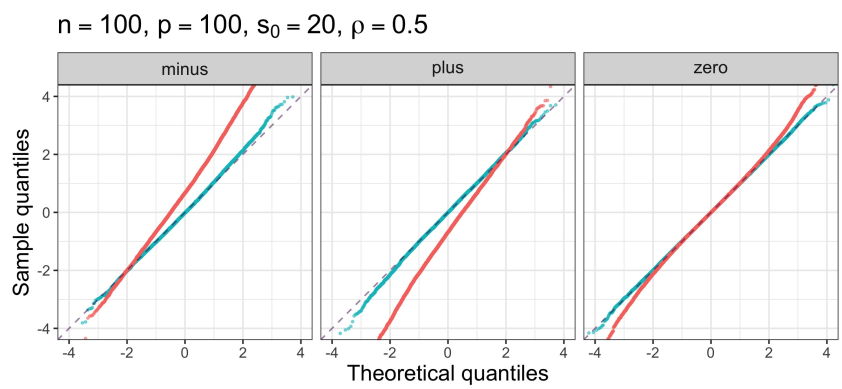

n = 25, p = 100, s0 = 20, ρ = 0.5
θj = −25
θj = 25
θj = 0
Density
θ̂j^d

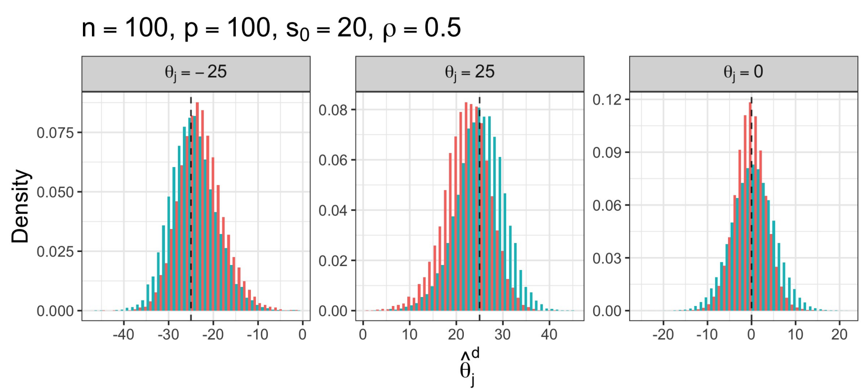

n = 50, p = 100, s0 = 20, ρ = 0.5
minus
plus
zero
Sample quantiles
Theoretical quantiles

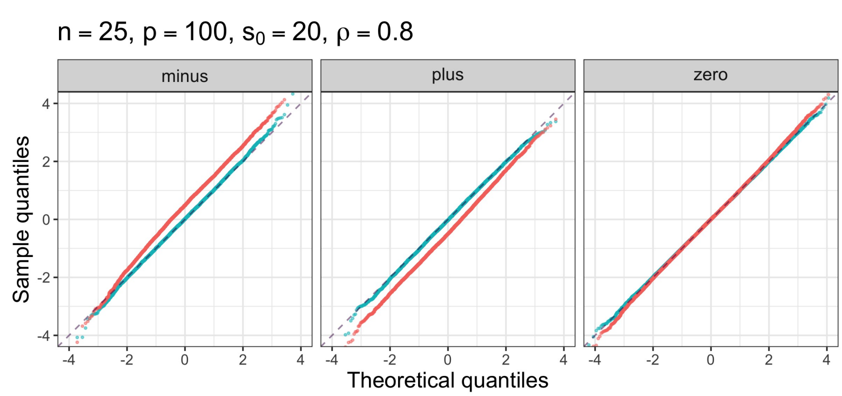

n = 50, p = 100, s0 = 20, ρ = 0.5
θj = −25
θj = 25
θj = 0
Density
θ̂j^d

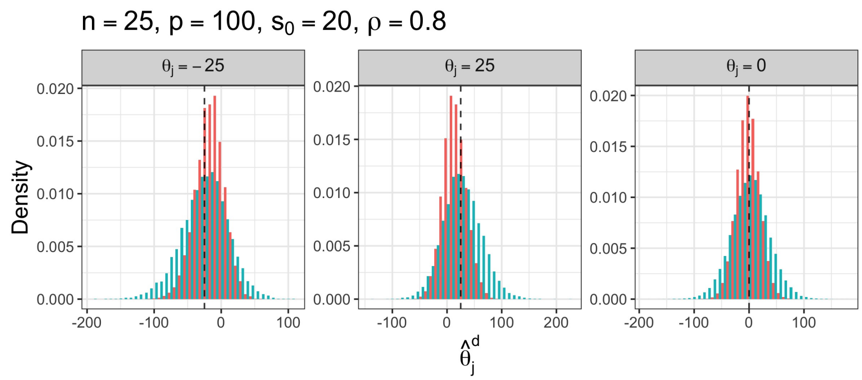

n = 75, p = 100, s0 = 20, ρ = 0.5
minus
plus
zero
Sample quantiles
Theoretical quantiles

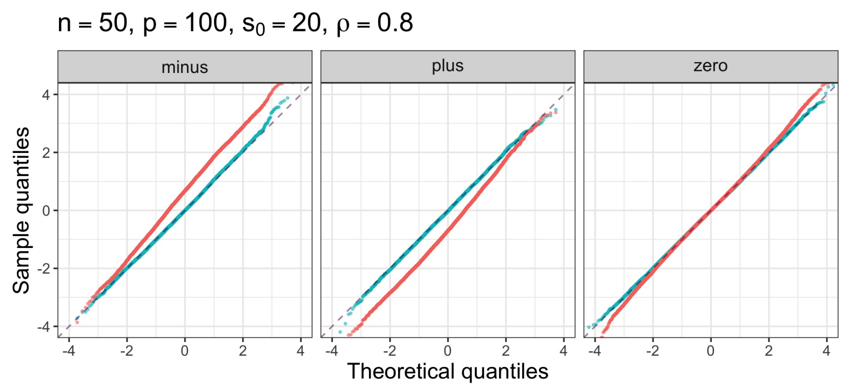

n = 75, p = 100, s0 = 20, ρ = 0.5
θj = −25
θj = 25
θj = 0
Density
θ̂j^d

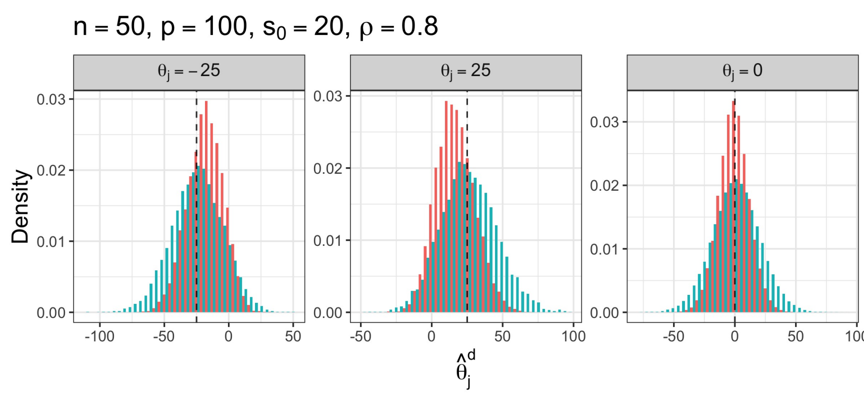

n = 100, p = 100, s0 = 20, ρ = 0.5
minus
plus
zero
Sample quantiles
Theoretical quantiles

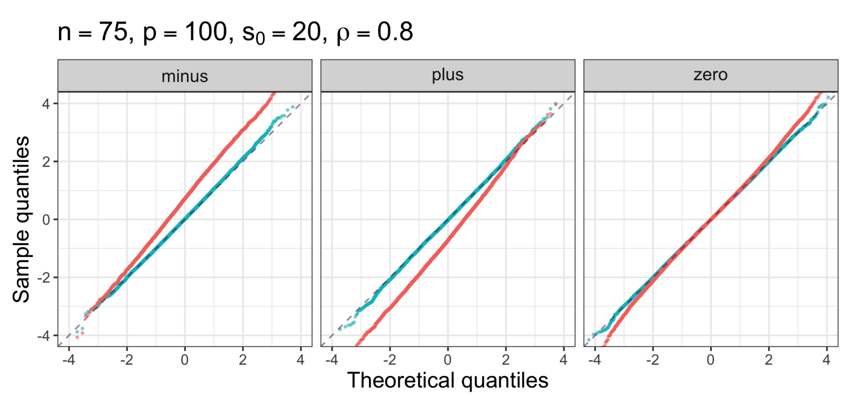

n = 100, p = 100, s0 = 20, ρ = 0.5
θj = −25
θj = 25
θj = 0
Density
θ̂j^d

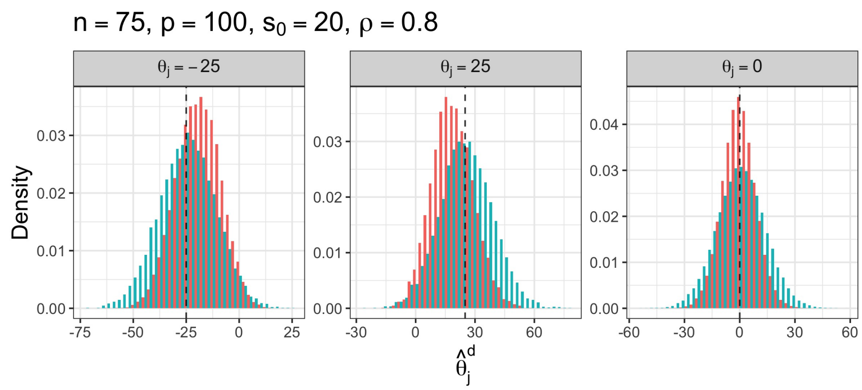

n = 25, p = 100, s0 = 20, ρ = 0.8
minus
plus
zero
Sample quantiles
Theoretical quantiles

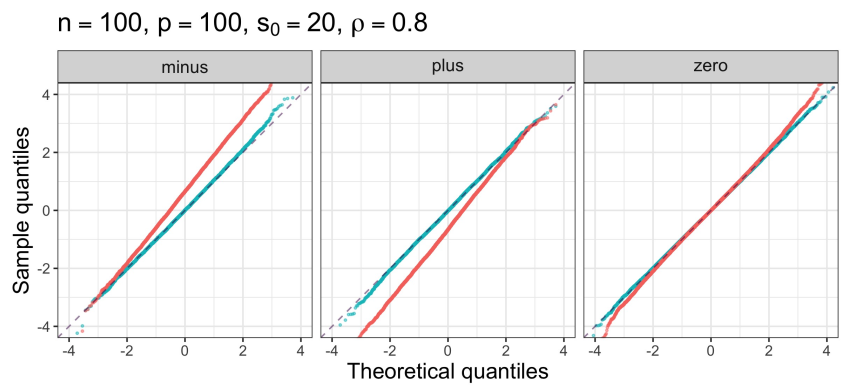

n = 25, p = 100, s_0 = 20, ρ = 0.8
$\theta_j = -25$
$\theta_j = 25$
$\theta_j = 0$
Density
$\hat{\theta}_j^d$

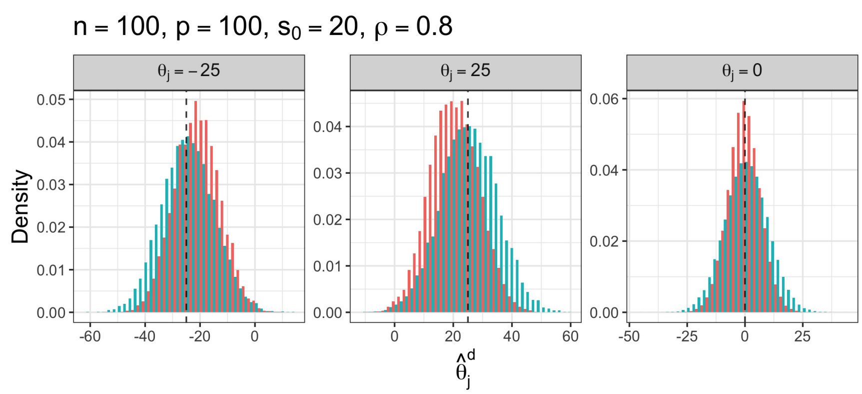

n = 50, p = 100, s_0 = 20, ρ = 0.8
minus
plus
zero
Sample quantiles
Theoretical quantiles


n = 50, p = 100, s_0 = 20, ρ = 0.8
$\theta_j = -25$
$\theta_j = 25$
$\theta_j = 0$
Density
$\hat{\theta}_j^d$


n = 75, p = 100, s_0 = 20, ρ = 0.8
minus
plus
zero
Sample quantiles
Theoretical quantiles


n = 75, p = 100, s$_0$ = 20, ρ = 0.8
θ$_j$ = −25
θ$_j$ = 25
θ$_j$ = 0
Density
$\hat{\theta}_j^d$


n = 100, p = 100, s$_0$ = 20, ρ = 0.8
minus
plus
zero
Sample quantiles
Theoretical quantiles


n = 100, p = 100, s$_0$ = 20, ρ = 0.8
θ$_j$ = −25
θ$_j$ = 25
θ$_j$ = 0
Density
$\hat{\theta}_j^d$