The Annals of Statistics
2023, Vol. 51, No. 2, 391–436
https://doi.org/10.1214/22-AOS2243

DEBIASING CONVEX REGULARIZED ESTIMATORS AND INTERVAL ESTIMATION IN LINEAR MODELS

BY PIERRE C. BELLEC[a] AND CUN-HUI ZHANG[b]

Department of Statistics, Rutgers University, [a]*pierre.bellec@rutgers.edu,* [b]*czhang@stat.rutgers.edu*

New upper bounds are developed for the L_2 distance between $\xi/\mathrm{Var}[\xi]^{1/2}$ and linear and quadratic functions of $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ for random variables of the form $\xi = z^\top f(z) - \mathrm{div}\, f(z)$. The linear approximation yields a central limit theorem when the squared norm of $f(z)$ dominates the squared Frobenius norm of $\nabla f(z)$ in expectation.

Applications of this normal approximation are given for the asymptotic normality of debiased estimators in linear regression with correlated design and convex penalty in the regime $p/n \le \gamma$ for constant $\gamma \in (0, \infty)$. For the estimation of linear functions $\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ of the unknown coefficient vector $\boldsymbol{\beta}$, this analysis leads to asymptotic normality of the debiased estimate for most normalized directions $\boldsymbol{a}_0$, where "most" is quantified in a precise sense. This asymptotic normality holds for any convex penalty if $\gamma < 1$ and for any strongly convex penalty if $\gamma \ge 1$. In particular, the penalty needs not be separable or permutation invariant. By allowing arbitrary regularizers, the results vastly broaden the scope of applicability of debiasing methodologies to obtain confidence intervals in high dimensions. In the absence of strong convexity for $p > n$, asymptotic normality of the debiased estimate is obtained for the Lasso and the group Lasso under additional conditions. For general convex penalties, our analysis also provides prediction and estimation error bounds of independent interest.

1. Introduction.

Consider the linear model

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

with an unknown coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$, a Gaussian noise vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ and a Gaussian design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$ rows independent of $\boldsymbol{\varepsilon}$. We assume throughout the sequel that $\boldsymbol{\Sigma}$ is invertible. The paper develops confidence intervals for $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ from a given regularized initial estimator $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$, using a technique referred to as *debiasing*: a correction to the initial estimate $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle$ in the direction $\boldsymbol{a}_0$ is constructed so that the *"debiased"* estimate can be used for inference about $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$.

1.1. *Regularization induces bias.*

If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, the unregulated least-squares estimate $\widehat{\boldsymbol{\beta}}^{ls} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is unbiased, that is, $\mathbb{E}[\widehat{\boldsymbol{\beta}}^{ls} - \boldsymbol{\beta} | \boldsymbol{X}] = \mathbf{0}$. On the other hand, if the square loss is regularized with an additive penalty,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b}) \tag{1.2}$$

for penalty functions commonly used in high-dimensional statistics such as $g(\boldsymbol{b}) = \lambda \|\boldsymbol{b}\|_1$ for $\lambda > 0$ (Lasso) or $g(\boldsymbol{b}) = \mu \|\boldsymbol{b}\|_2^2$ for $\mu > 0$ (ridge regression), then $\widehat{\boldsymbol{\beta}}$ is biased.

For ridge regression $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, this bias can be quantified explicitly when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ as a shrinkage to the origin. Let $\sum_{i=1}^r \boldsymbol{u}_i s_i \boldsymbol{v}_i^\top$ be the SVD of $\boldsymbol{X}$ with $s_i > 0$ and

Received July 2020; revised October 2022.
MSC2020 subject classifications. Primary 62H12, 62G15; secondary 62F35, 62J07.
Key words and phrases. Bias correction, central limit theorem, confidence intervals, convex regularization, Gaussian Poincaré inequality, high-dimensional linear models, Lasso, Stein's formula, variance estimation.

$r = \min(n, p)$. By rotational invariance, $\boldsymbol{v}_i$ is independent of s_i and uniformly distributed in the unit sphere in $\mathbb{R}^p$. Thus, with G_γ being the Marchenko–Pastur law,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}\left[\sum_{i=1}^{r} \frac{s_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{\beta}}{s_i^2 + n\mu}\right] = \mathbb{E}\left[\sum_{i=1}^{r} \frac{p^{-1} s_i^2}{s_i^2 + n\mu}\right] \boldsymbol{\beta} \approx \boldsymbol{\beta} \int \frac{(r/p)x}{x + (r/p)\mu} G_\gamma(dx) \quad \text{as } \frac{p}{n} \to \gamma.$$

The Lasso penalty $g(\boldsymbol{b}) = \lambda \|\boldsymbol{b}\|_1$ also introduces bias. For example, for deterministic orthonormal designs, the Lasso estimator of the coefficient β_j is the soft-thresholding of $N(\beta_j, \sigma^2/n)$, which is again biased toward the origin. For Gaussian designs with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and in an average sense, the Lasso is approximately the soft-thresholding of $N(\beta_j, \tau_*^2/n)$ with certain $\tau_* \geq \sigma$ under proper conditions [1]. Thus, with $s_1 = \#\{j : |\beta_j| > \lambda\}$, the squared bias of the Lasso, $\|\boldsymbol{\beta} - \mathbb{E}[\widehat{\boldsymbol{\beta}}]\|_2^2$, is expected to have no smaller order than the lower bound $s_1\lambda^2$ for its ℓ_2 risk [3], Theorem 3.1. Alternative approaches were proposed to remove or reduce the bias of the Lasso for strong signals, for example, by using concave penalty functions (e.g., SCAD [24], MCP [48]) or iterated hard thresholding algorithms [13]. These approaches yield an error term of the order $(\|\boldsymbol{\beta}\|_0 - s_1')\lambda^2 + s_1'\sigma^2/n$ where $s_1' = \{j = 1, \ldots, p : |\beta_j| > c\lambda\}$ for some constant $c > 0$ [25, 34], alleviating the bias of the Lasso for large coefficients at typical penalty levels $\lambda > \sigma/n^{1/2}$.

Debiasing the Lasso, asymptotic normality and confidence intervals. If the goal is the estimation of a single scalar parameter $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ in a predetermined direction $\boldsymbol{a}_0$ instead of the full vector $\boldsymbol{\beta} \in \mathbb{R}^p$, it is possible to correct the bias of the Lasso and to construct confidence intervals for θ: there is already a vast literature on asymptotic normality of de-biased estimates in sparse linear regression for the Lasso [8, 9, 26–28, 33, 46, 51], among others. In this literature $\boldsymbol{a}_0$ is usually the jth canonical basis vector and β_j the scalar parameter of interest. Given the Lasso $\widehat{\boldsymbol{\beta}}$ as an initial estimator of $\boldsymbol{\beta}$, the idea is to add a debiasing term to achieve asymptotic normality, which then yields confidence intervals for $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$. If $s_0 = \|\boldsymbol{\beta}\|_0$ in (1.1), several debiased estimators have been proposed and their asymptotic normality hold under certain rate conditions on s_0, n, p. The earliest works on this topic [9, 26, 46, 51] provide asymptotic normality results in the regime $s_0 \log(p)/\sqrt{n} \to 0$. When $s_0 \log(p)/\sqrt{n} \to 0$ indeed holds, the debiasing constructions in these papers are all first-order equivalent to each other, and under normalization $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|_2 = 1$ to

$$\begin{aligned} \widehat{\theta} &= \underbrace{\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle}_{\text{initial estimate}} + \underbrace{\|\boldsymbol{z}_0\|_2^{-2} \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}_{\text{debiasing correction}}, \\ \sqrt{n}(\widehat{\theta} - \theta) &= \underbrace{\sqrt{n}\|\boldsymbol{z}_0\|_2^{-2} \boldsymbol{z}_0^\top \boldsymbol{\varepsilon}}_{\text{normal part}} + \underbrace{O_{\mathbb{P}}(R_n)}_{\text{remainder}}, \end{aligned} \tag{1.3}$$

where $\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0/\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0 \rangle$ and $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{u}_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$. While these works do not assume $\boldsymbol{\Sigma}$ known and construct an estimated score vector $\widehat{\boldsymbol{z}}$ for $\boldsymbol{z}_0$, the impact of using $\widehat{\boldsymbol{z}}$ can be absorbed into the remainder in (1.3) with $R_n = \sigma s_0 \log(p)/\sqrt{n}$. The direction $\boldsymbol{u}_0$ and the debiasing correction in (1.3) have a natural semiparametric interpretation [49]. Viewing $\theta : \mathbb{R}^p \to \mathbb{R}$ as the function $\theta(\boldsymbol{\beta}) = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$, the Fischer information for the estimation of $\theta(\boldsymbol{\beta})$ in (1.1) is $F_\theta = 1/(\sigma^2 \langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0 \rangle)$, and the direction $\boldsymbol{u}_0$ above is the only $\boldsymbol{u} \in \mathbb{R}^p$ with

$$\langle \nabla\theta(\widehat{\boldsymbol{\beta}}), \boldsymbol{u} \rangle = \langle \boldsymbol{a}_0, \boldsymbol{u} \rangle = 1 \tag{1.4}$$

such that F_θ is also the Fischer information in the one-dimensional submodel $\{\widehat{\boldsymbol{\beta}} + t\boldsymbol{u}, t \in \mathbb{R}\}$. For this reason, the line $\{\widehat{\boldsymbol{\beta}} + t\boldsymbol{u}_0, t \in \mathbb{R}\}$ is referred to as the least-favorable one-dimensional submodel for the estimation of θ. The normalization (1.4) ensures that $\theta(\widehat{\boldsymbol{\beta}} + t\boldsymbol{u}) = \theta(\widehat{\boldsymbol{\beta}}) + t$ and $\widehat{\theta} = \theta(\widehat{\boldsymbol{\beta}} + \widehat{t}\boldsymbol{u}_0)$ with $\widehat{t} = \|\boldsymbol{z}_0\|_2^{-2} \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, so that (1.3) replaces the initial $\widehat{\boldsymbol{\beta}}$ with

its one-step correction $\widehat{\boldsymbol{\beta}} + \widehat{t}\boldsymbol{u}_0$, where $\widehat{t}$ maximizes the likelihood in the least-favorable submodel. We refer to [10] for a systematic study of this semiparametric perspective.

If $s_0 \log(p)/\sqrt{n} \to +\infty$ and $\boldsymbol{\Sigma}$ is unknown with bounded spectrum, the minimax estimation error of the form $\sqrt{n}(\widehat{\theta} - \theta)$ diverges for any estimator $\widehat{\theta}$ [16]. This rules out asymptotic normally results at the $\sqrt{n}$ adjusted rate if $s_0 \log(p)/\sqrt{n} \to +\infty$ and no further assumption is made on $\boldsymbol{\Sigma}$. However, if $\boldsymbol{z}_0$ is known, (1.3) holds with $R'_n = \sqrt{s_0 \log(p/s_0)/n}(1 + s_0/\sqrt{n})$, providing asymptotic normality for sparsity levels $s_0 \lesssim n^{2/3}$ up to logarithmic factors; cf. [8], Corollary 3.3. Similarly, [28], Theorem 3.8, provides (1.3) with $\boldsymbol{a}_0 = \boldsymbol{e}_j \in \mathbb{R}^p$ a canonical basis vector and $R'_n = \log(p)\sqrt{s_0/n} \max_j \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1$. Already in the regime $\sqrt{n} \lll s_0 \lll n^{2/3}$, the arguments of [8, 28] differ significantly from the ℓ_1-ℓ_∞ Hölder inequality argument of [9, 26, 46, 51]: while these earlier works prove asymptotic normality with a remainder term of order $O_{\mathbb{P}}(s_0 \log(p)/\sqrt{n})$, [8, 28] analyze explicitly the smaller order terms hidden in this $O_{\mathbb{P}}(s_0 \log(p)/\sqrt{n})$ remainder.

For $s_0 \ggg n^{2/3}$, the debiasing correction in (1.3) needs to be modified:

$$\begin{aligned} \widehat{\theta} &= \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle + \big(n - |\widehat{S}|\big)^{-1} \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \\ \sqrt{n}(\widehat{\theta} - \theta) &= \sqrt{n}\|\boldsymbol{z}_0\|_2^{-2} \boldsymbol{z}_0^\top \boldsymbol{\varepsilon} + O_{\mathbb{P}}(R'_n) \end{aligned} \tag{1.5}$$

with $\widehat{S} = \{j \in [p] : \widehat{\beta}_j \neq 0\}$ and $R'_n = \sigma(s_0 \log(p/s_0)/n)^{1/2}$; cf. [8], Theorem 3.1. For $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$, the difference from (1.3) is the replacement of $\|\boldsymbol{z}_0\|_2^{-2} \approx n^{-1}$ in the debiasing correction with $(n - |\widehat{S}|)^{-1}$ to amplify it by a factor $(1 - |\widehat{S}|/n)^{-1}$. This modification is required as soon as $s_0 \ggg n^{2/3}$ up to logarithmic factors [8], Section 3. These asymptotic results for $s_0 \ggg \sqrt{n}$ are amenable to the lack of knowledge of $\boldsymbol{\Sigma}$: in this case, estimation of $\boldsymbol{z}_0$ is possible when $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ is sufficiently sparse; see [28] if the direction of interest $\boldsymbol{a}_0$ is canonical basis vector and [8], Section 2.2, for arbitrary direction $\boldsymbol{a}_0$. These results [8, 28] for $s_0 \ggg \sqrt{n}$ and correlated $\boldsymbol{\Sigma}$ are so far restricted to random Gaussian designs.

Inflated asymptotic variance for nonvanishing prediction error. In the results discussed so far for the Lasso, $s_0 \log(p/s_0)/n \to 0$ or stronger conditions are required for asymptotic normality, and the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$ is σ^2. The condition $s_0 \log(p/s_0)/n \to 0$ implies the consistency of the Lasso in prediction and estimation thanks to error bounds of the form $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \lesssim s_0 \log(p/s_0)/n$ [5, 8, 38, 50]. It turns out that the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$ is larger than σ^2 if $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2$ does not vanish; this is the situation studied in the present work. The literature on asymptotic normality of debiased estimates in the regime

$$p/n \to \gamma \in (0, +\infty), \qquad s_0/n \to \kappa \in (0, 1) \tag{1.6}$$

for constants $\gamma, \kappa > 0$ is more scarce. In this regime where p, n and s_0 are all of the same order, [27, 33] provide asymptotic normality results for the debiased Lasso (1.5) in the estimation of β_j (canonical $\boldsymbol{a}_0 = \boldsymbol{e}_j$) in the isotropic Gaussian design. In these works, the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$ equals a constant τ_*^2 satisfying the system of two nonlinear equations in [1] and [33], Proposition 3.1, Theorem 3.1. The constant τ_*^2 is related to the residual sum of squares [33], Corollary 4.1, and out-of-sample error [33], Theorem 3.2, as in

$$\big(1 - |\widehat{S}|/n\big)^{-2} \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/n \to^{\mathbb{P}} \tau_*^2, \qquad \sigma^2 + \big\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big\|_2^2 \to^{\mathbb{P}} \tau_*^2,$$

where $\to^{\mathbb{P}}$ denotes convergence in probability. These results for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ highlight that the asymptotic variance is strictly larger than σ^2 when p, n are of the same order as in (1.6). This phenomenon in the regime (1.6) is generic: for instance, the asymptotic variance is also larger than σ^2 for all permutation-invariant penalty functions [17], Proposition 4.3.

In this regime where n and p are of the same order, [21, 23] proved asymptotic normality and characterized the variance for unregularized M-estimators. For M-estimators, a debiasing correction is unnecessary due to the absence of regularization, and a rotational invariance argument reduces the problem of correlated designs to a corresponding uncorrelated one [23], Lemma 1. However, this rotational invariance is lost in the presence of a penalty such as the ℓ_1-norm. New techniques are called for to analyze the asymptotic behavior in the regime (1.6) and under correlated designs of estimators that are not rotational invariant. More recently, the approximate message Ppassing techniques used in [21, 27] were used to obtain similar results in logistic regression [40]; but again, these techniques cannot handle the Lasso penalty for correlated design. A more detailed comparison with these works is made in Section 3.8. To our knowledge, there is no previous asymptotic normality result for debiased estimates in the regime (1.6) for correlated designs in the presence of a penalty not depending on $\boldsymbol{\Sigma}$ (i.e., in situations where rotational invariance does not hold). A main goal of the paper is to fill this gap. Available techniques that tackle the regime (1.6) assume, in addition to uncorrelated design, that the penalty is invariant under permutations of the p coefficients [1, 15, 17, 33] and that the empirical distribution of the true $\{\sqrt{n}\beta_j, j \le p\}$ converges to some prior distribution. A second goal of the present paper is to show that asymptotic normality of debiased estimates can be obtained beyond the Lasso and beyond permutation-invariant penalty functions, without imposing the convergence of the empirical distribution of the normalized coefficients $\{\sqrt{n}\beta_j, j \le p\}$.

1.2. *A general construction of debiased estimators.* This section describes a general approach to systematically construct de-biased estimates in the linear model (1.1) where $\boldsymbol{X}$ has i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows. Our goal is to construct confidence intervals for the one-dimensional parameter $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$. Consider an initial estimator $\widehat{\boldsymbol{\beta}}$, viewed as a function of $(\boldsymbol{y}, \boldsymbol{X})$, that is, $\widehat{\boldsymbol{\beta}} : \mathbb{R}^{n\times(1+p)} \to \mathbb{R}^p$ and assume that this function $\widehat{\boldsymbol{\beta}}$ is Fréchet[1] differentiable. For a given observed data $(\boldsymbol{y}, \boldsymbol{X})$ from the linear model (1.1) and a $\widehat{\boldsymbol{\beta}}$ Fréchet differentiable at $(\boldsymbol{y}, \boldsymbol{X})$, there exist uniquely matrices $\widehat{\boldsymbol{H}} \in \mathbb{R}^{n\times n}$ and $\widehat{\boldsymbol{G}} \in \mathbb{R}^{n\times p}$ such that

$$
\begin{aligned}
\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}+\boldsymbol{\eta}, \boldsymbol{X}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{H}}^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|), \\
\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}_0^\top) - \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{G}}^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|)
\end{aligned}
\tag{1.7}
$$

for all $\boldsymbol{\eta} \in \mathbb{R}^n$. With $\boldsymbol{X} = (x_{ij})_{i\in[n], j\in[p]}$, if the partial derivatives of $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ at the observed data $(\boldsymbol{y}, \boldsymbol{X})$ are $(\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ and $(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ then (1.7) implies $\widehat{\boldsymbol{H}}^\top \boldsymbol{e}_i = \boldsymbol{X}(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ and $\widehat{\boldsymbol{G}}^\top \boldsymbol{e}_i = \sum_{j=1}^p \langle \boldsymbol{a}_0, \boldsymbol{e}_j \rangle (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ for canonical basis vectors $\boldsymbol{e}_i \in \mathbb{R}^n$ and $\boldsymbol{e}_j \in \mathbb{R}^p$. The derivatives of $\widehat{\boldsymbol{\beta}}$ and the matrices $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{G}}$ can be computed by only looking at the observed data $(\boldsymbol{y}, \boldsymbol{X})$, for instance by finite difference schemes.

Next, consider the function $\boldsymbol{\phi}$ defined as

$$
\boldsymbol{\phi} : \mathbb{R}^{n\times(1+p)} \to \mathbb{R}^n, \qquad (\boldsymbol{y}, \boldsymbol{X}) \mapsto \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{y}.
$$

If $\widehat{\boldsymbol{\beta}}$ is differentiable at $(\boldsymbol{y}, \boldsymbol{X})$, then $\boldsymbol{\phi}$ is differentiable as well. By the product and chain rules,

$$
\begin{aligned}
\boldsymbol{\phi}(\boldsymbol{y}+\boldsymbol{\eta}, \boldsymbol{X}) - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X}) &= [\widehat{\boldsymbol{H}} - \boldsymbol{I}_n]^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|), \\
\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}_0^\top) - \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X}) &= \big[\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle \boldsymbol{I}_n + \widehat{\boldsymbol{G}}\boldsymbol{X}^\top\big]^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|).
\end{aligned}
\tag{1.8}
$$

[1] Although the Fréchet derivative is the usual definition of derivative in finite dimension, we write Fréchet to emphasize that the derivative is linear. Linearity may fail for weaker notions such as Gateaux differentiability.

If the partial derivatives of $\boldsymbol{\phi}$ are $(\partial/\partial y_i)\boldsymbol{\phi}$ and $(\partial/\partial x_{ij})\boldsymbol{\phi}$, the second line of the previous display is equivalently rewritten as

$$\sum_{j=1}^{p} \langle \boldsymbol{a}_0, \boldsymbol{e}_j \rangle (\partial/\partial x_{ij}) \boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X}) = \big[\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle \boldsymbol{I}_n + \widehat{\boldsymbol{G}} \boldsymbol{X}^\top \big]^\top \boldsymbol{e}_i$$

for each canonical basis vector $\boldsymbol{\eta} = \boldsymbol{e}_i \in \mathbb{R}^n$.

Observe that the arguments $(\boldsymbol{y}, \boldsymbol{X})$ of $\boldsymbol{\phi}$ are centered and jointly normal random variables and their correlations are computed explicitly, for example, $\mathbb{E}[x_{ij} y_l] = \boldsymbol{e}_j^\top \boldsymbol{\Sigma} \boldsymbol{\beta} \boldsymbol{I}_{\{i=\ell\}}$, with basis vectors $\boldsymbol{e}_j \in \mathbb{R}^p$. One version of Stein's formula, also known as Gaussian integration by parts, is $\mathbb{E}[G h(Z_1, \ldots, Z_q)] = \sum_{k=1}^{q} \mathbb{E}[\mathrm{G} Z_k] \mathbb{E}[(\partial/\partial z_k) h(Z_1, \ldots, Z_q)]$ provided that the function $h(z_1, \ldots, z_q)$ is differentiable and that $G, Z_1, \ldots, Z_q$ are centered jointly normal random variables, provided the existence of the expectations [41], Appendix A.4. We leverage this version of Stein's formula to obtain an unbiased estimating equation involving only one unknown parameter, the scalar $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ of interest. For $G_i = \boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0$, we find $\mathbb{E}[G_i y_k] = \mathbb{E}[G_i x_{kj}] = 0$ if $i \neq k$ while $\mathbb{E}[G_i x_{ij}] = \langle \boldsymbol{a}_0, \boldsymbol{e}_j \rangle$ and $\mathbb{E}[G_i y_i] = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ so that by reading the partial derivatives in (1.8),

$$\begin{aligned} \mathbb{E}\big[G_i \phi_i(\boldsymbol{y}, \boldsymbol{X})\big] &= \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle \mathbb{E}\bigg[\frac{\partial \phi_i}{\partial y_i}(\boldsymbol{y}, \boldsymbol{X}) \bigg] + \sum_{j=1}^{p} \langle \boldsymbol{a}_0, \boldsymbol{e}_j \rangle \mathbb{E}\bigg[\frac{\partial \phi_i}{\partial x_{ij}}(\boldsymbol{y}, \boldsymbol{X}) \bigg] \\ &= \mathbb{E}\big[\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle (\widehat{\boldsymbol{H}}_{ii} - 1) \big] + \mathbb{E}\big[\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle + \boldsymbol{e}_i^\top \widehat{\boldsymbol{G}} \boldsymbol{X}^\top \boldsymbol{e}_i \big]. \end{aligned} \tag{1.9}$$

Summing over $i = 1, \ldots, n$ and using $\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{X}) = \boldsymbol{X} \widehat{\boldsymbol{\beta}} - \boldsymbol{y}$, we find that

$$\mathbb{E}\big[\big\langle \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0, \boldsymbol{X} \widehat{\boldsymbol{\beta}} - \boldsymbol{y} \big\rangle \big] = \mathbb{E}\big[-\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle \operatorname{trace}[\boldsymbol{I}_n - \widehat{\boldsymbol{H}}] + \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle n + \operatorname{trace}\big[\boldsymbol{X}^\top \widehat{\boldsymbol{G}} \big] \big].$$

To transform this equation into a form representative of the results of the paper, define the scalars $\widehat{\mathsf{df}}$ and $\widehat{A}$ by

$$\widehat{\mathsf{df}} = \operatorname{trace}[\widehat{\boldsymbol{H}}], \qquad \widehat{A} = \operatorname{trace}\big[\boldsymbol{X}^\top \widehat{\boldsymbol{G}} \big] + \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle \widehat{\mathsf{df}}. \tag{1.10}$$

The notation $\widehat{\mathsf{df}}$ underlines that $\operatorname{trace}[\widehat{\boldsymbol{H}}]$ has the interpretation of degrees-of-freedom of the estimator $\widehat{\boldsymbol{\beta}}$ in Stein's Unbiased Risk Estimate (SURE) [37]: regarding $\widehat{\boldsymbol{\mu}} = \boldsymbol{X} \widehat{\boldsymbol{\beta}}$ as an estimate of $\boldsymbol{\mu} = \boldsymbol{X} \boldsymbol{\beta}$ in the Gaussian sequence model with observation $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, the quantity $\widehat{\mathrm{SURE}} = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + 2\sigma^2 \widehat{\mathsf{df}} - \sigma^2 n$ is an unbiased estimate of the in-sample error $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$. With this notation, we obtain the *unbiased estimating equation*,

$$0 = \mathbb{E}\big[\big\langle \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0, \boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}} \big\rangle + (n - \widehat{\mathsf{df}}) \big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle - \theta \big) + \widehat{A} \big], \tag{1.11}$$

where the only unobserved quantity inside the expectation is $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$, the scalar parameter we wish to estimate. In the above application of Stein's formula, $G_i = \boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0$ was chosen on purpose so that $\boldsymbol{\beta}$ appears in (1.9) only through $\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$ thanks to $\mathbb{E}[G_i y_i] = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$. Note that replacing G_i in (1.9) by $\boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{u}$ for any $\boldsymbol{u} \in \mathbb{R}^p$ not proportional to $\boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0$ brings a scalar projection of $\boldsymbol{\beta}$ different from $\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$: this shows the unique role of the random vector $\boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0$ to derive an unbiased estimating equation for $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle$. It is notable that the direction $\boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0$ coincides with the least-favorable direction described around (1.4). Equation (1.11) is obtained for an arbitrary initial estimator $\widehat{\boldsymbol{\beta}}$ provided that its derivatives with respect to $(\boldsymbol{y}, \boldsymbol{X})$ exist and the integrability conditions hold to ensure existence of the expectations involved. From (1.11), the method of moments suggests to estimate θ with $\widehat{\theta} = \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle + (n - \widehat{\mathsf{df}})^{-1} (\langle \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0, \boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}} \rangle + \widehat{A})$, which resembles (1.5) for the Lasso for $\widehat{\mathsf{df}} = |\widehat{S}|$ and $\boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0 = z_0$ under the normalization $\langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0 \rangle = 1$.

It is useful at this point to specialize the above derivation to an estimator for which all derivatives can be computed explicitly. For Ridge regression with penalty $g(\boldsymbol{b}) = \mu \|\boldsymbol{b}\|_2^2$ for

some $\mu > 0$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = (\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ and

$$\begin{aligned} \widehat{\boldsymbol{H}}^\top &= \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top, \\ \widehat{\boldsymbol{G}}^\top &= (\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} [\boldsymbol{a}_0 (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top - \boldsymbol{X}^\top \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle]. \end{aligned} \tag{1.12}$$

Indeed, the derivatives of $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ exist as it is the composition of elementary differentiable functions. Differentiation with respect to $\boldsymbol{y}$ is straightforward as $\widehat{\boldsymbol{\beta}}$ is linear in $\boldsymbol{y}$, while in order to compute $\widehat{\boldsymbol{G}}$ we proceed by setting $\boldsymbol{b}(t) = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}(t))$ with $\boldsymbol{X}(t) = \boldsymbol{X} + t\boldsymbol{\eta}\boldsymbol{a}_0^\top$. Differentiation of the KKT conditions $\boldsymbol{X}(t)^\top (\boldsymbol{y} - \boldsymbol{X}(t)\boldsymbol{b}(t)) = n\mu \boldsymbol{b}(t)$ at $t = 0$ provides the directional derivative $(d/dt)\boldsymbol{b}(t)|_{t=0} = \widehat{\boldsymbol{G}}^\top \boldsymbol{\eta}$. This gives (1.12). It follows from (1.12) that $\widehat{\mathsf{df}} = \operatorname{trace}[\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top]$ and $\widehat{A} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1} \boldsymbol{a}_0$ for the quantities in (1.10) (for $\widehat{A}$, note the fortuitous cancelation of the term $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} \rangle \widehat{\mathsf{df}}$). For the Lasso, similar differentiability formulae are derived in [8]. It is however, unclear how to obtain closed form formulae for the derivatives of $\widehat{\boldsymbol{\beta}}$ for an arbitrary convex penalty g in (1.2).

We now set up some notation that will be useful for the rest of the paper, and derive again the unbiased estimating equation (1.11) using this new notation. Define

$$\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0 / \langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0 \rangle, \qquad z_0 = \boldsymbol{X}\boldsymbol{u}_0, \qquad \boldsymbol{Q}_0 = \boldsymbol{I}_{p \times p} - \boldsymbol{u}_0 \boldsymbol{a}_0^\top. \tag{1.13}$$

The normalizing constant in $\boldsymbol{u}_0$ is such that $\langle \boldsymbol{a}_0, \boldsymbol{u}_0 \rangle = 1$ holds so that the expression (1.13) for $\boldsymbol{u}_0$ coincides with the direction of the least-favorable submodel discussed around (1.4). The vector z_0 is independent of $\boldsymbol{X}\boldsymbol{Q}_0$ by construction as $(z_0, \boldsymbol{X}\boldsymbol{Q}_0)$ are jointly normal and uncorrelated. This follows by noting that $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ has i.i.d. $N(0, 1)$ entries and

$$z_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \boldsymbol{v} / \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{a}_0\|, \qquad \boldsymbol{X}\boldsymbol{Q}_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} (\boldsymbol{I}_p - \boldsymbol{v}\boldsymbol{v}^\top) \boldsymbol{\Sigma}^{1/2}$$

for the unit vector $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{a}_0 / \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{a}_0\|$ as by construction of $\boldsymbol{Q}_0$ matrix $\boldsymbol{I}_p - \boldsymbol{v}\boldsymbol{v}^\top = \boldsymbol{\Sigma}^{1/2} \boldsymbol{Q}_0 \boldsymbol{\Sigma}^{-1/2}$ is the orthogonal projection onto $\{\boldsymbol{v}\}^\perp$. We summarize this as

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + z_0 \boldsymbol{a}_0^\top \quad \text{with } z_0 \sim N(\boldsymbol{0}, \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{a}_0\|^{-2} \boldsymbol{I}_n) \text{ independent of } \boldsymbol{X}\boldsymbol{Q}_0. \tag{1.14}$$

For brevity, we assume in the sequel and without loss of generality that the direction of interest $\boldsymbol{a}_0$ is normalized such that

$$\|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{a}_0\|^2 = \langle \boldsymbol{a}_0, \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_0 \rangle = 1. \tag{1.15}$$

By definition of $\boldsymbol{u}_0$ and z_0, the normalization (1.15) gives $z_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$.

Conditionally on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$, define the function $f_{(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$f_{(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})}(z_0) = \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}. \tag{1.16}$$

By (1.14) and the independence of $\boldsymbol{\varepsilon}$ and $\boldsymbol{X}$, the conditional expectation given $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ can be written as integrals against the Gaussian measure of z_0, for example,

$$\mathbb{E}[z_0^\top f_{(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})}(z_0) | (\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})] = \int (z^\top f_{(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})}(z)) e^{-\|z\|_2^2/2} (\sqrt{2\pi})^{-n} \, dz$$

since $z_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$. As we argue conditionally on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$, we omit the dependence on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ and write simply $f : \mathbb{R}^n \to \mathbb{R}^n$. Since $\boldsymbol{y} = \boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + z_0 \boldsymbol{a}_0^\top$,

$$f(z_0) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{\beta} + z_0 \boldsymbol{a}_0^\top \boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{Q}_0 + z_0 \boldsymbol{a}_0^\top) - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\varepsilon}.$$

The gradient ∇f with respect to z_0, holding $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ fixed, can be computed by the product rule and the chain rule via (1.8):

$$\nabla f(z_0)^\top = \boldsymbol{I}_n \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + [\langle \boldsymbol{a}_0, \boldsymbol{\beta} \rangle \widehat{\boldsymbol{H}}^\top + \boldsymbol{X}\widehat{\boldsymbol{G}}^\top]. \tag{1.17}$$

We adopt the usual convention that the gradient of a vector valued function is the transpose of its Jacobian. Computing the directional derivative of f in a direction $\boldsymbol{\eta}$ requires considering the difference of an expression at $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0, z_0 + t\boldsymbol{\eta})$ minus the same expression at $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0, z_0)$, dividing by t and taking the limit as $t \to 0$; this is equivalent to considering the difference of an expression at $(\boldsymbol{\varepsilon}, \boldsymbol{X}(t))$ with $\boldsymbol{X}(t) = \boldsymbol{X} + t\boldsymbol{\eta}\boldsymbol{a}_0^\top$ minus the same expression at $(\boldsymbol{\varepsilon}, \boldsymbol{X})$, dividing by t and taking the limit as $t \to 0$.

Taking the trace of (1.17) and by definition of $\widehat{\mathsf{df}}$ and $\widehat{A}$ in (1.10), the identity

$$\begin{aligned} -\xi_0 &\overset{\text{def}}{=} \operatorname{div} f(z_0) - z_0^\top f(z_0) \\ &= (n - \widehat{\mathsf{df}})\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big) + \langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle + \widehat{A} \end{aligned} \tag{1.18}$$

holds where $\operatorname{div} f(z_0) = \operatorname{trace}[\nabla f(z_0)]$. Since $\mathbb{E}[\operatorname{div} f(z_0) - z_0^\top f(z_0) | (\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})] = 0$ by Stein's formula [37], this provides the unbiased estimating equation (1.11). Reasoning conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$, using Stein formulae with respect to z_0 involving conditional expectations given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ and gradients of the form $\nabla f(z_0)$ holding $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ fixed will be a recurring theme throughout the paper. In this context, the function f itself depends on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ as in (1.16), although the dependence on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ is omitted for brevity.

In order to construct confidence intervals using the unbiased estimating equation (1.11), one may hope that the quantity (1.18) above is well behaved—ideally, approximately normal with mean zero and a variance that can be consistently estimated from the observed data. By the second-order Stein's formula in Proposition 2.1 below, which was already known to Stein [37], (8.6), in a different form, the conditional variance of (1.18) given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ is

$$\begin{aligned} \operatorname{Var}_0[\xi_0] &= \mathbb{E}_0\big[\|f(z_0)\|^2 + \operatorname{trace}\big[\{\nabla f(z_0)\}^2\big]\big] \\ &= \mathbb{E}_0\big[V^*(\theta)\big]\big] \quad \text{for } V^*(\theta) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \operatorname{trace}\big[\{\nabla f(z_0)\}^2\big], \end{aligned} \tag{1.19}$$

where $\mathbb{E}_0 = \mathbb{E}[\cdot | \boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0]$ denotes the conditional expectation with respect to z_0 given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ and Var_0 denotes the conditional variance given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$. The gradient $\nabla f(z_0)$ in (1.17) and the unbiased estimate $V^*(\theta)$ of $\operatorname{Var}_0[\xi_0]$ only depend on the unknown parameter of interest θ and observable quantities, and $V^*(\theta)$ is quadratic in θ.

Assume now we are in an ideal situation in the sense that both conditions below are satisfied: (i) The quantity (1.18) is approximately normally distributed conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ and (ii) $V^*(\theta)$ is a consistent estimator of (1.19), the conditional variance of the random variable (1.18). Then the set of θ for which the inequality

$$\big[(n - \widehat{\mathsf{df}})\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big) + \langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle + \widehat{A}\big]^2 - V^*(\theta) z_{\alpha/2}^2 \le 0 \tag{1.20}$$

is satisfied is an $(1-\alpha)$-confidence interval, where $\mathbb{P}(|N(0,1)| > z_{\alpha/2}) = 1 - \alpha$. Solving the corresponding quadratic equality gives up to two solutions $\Theta_1(z_{\alpha/2}) \le \Theta_2(z_{\alpha/2})$ that are such that (1.20) holds with equality. These two solutions implicitly depend on the observables

$$\langle \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}, z_0\rangle, \qquad \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2, \qquad \widehat{\mathsf{df}}, \qquad \widehat{A}, \qquad \boldsymbol{a}_0^\top \widehat{\boldsymbol{\beta}}$$

and the derivatives of $\widehat{\boldsymbol{\beta}}$. If the coefficient of θ^2 in the left-hand side of (1.20) is positive, (i.e., if the leading coefficient of (1.20), seen as a polynomial in θ with data-driven coefficients, is positive), a $(1-\alpha)$ confidence interval for $\theta = \boldsymbol{a}_0^\top \boldsymbol{\beta}$ is then given by

$$\widehat{CI} = \big[\Theta_1(z_{\alpha/2}), \Theta_2(z_{\alpha/2})\big]. \tag{1.21}$$

We will show in the discussion surrounding (3.30) below that the dominant coefficient is positive and that the confidence interval is indeed of the above form if $\widehat{\boldsymbol{\beta}}$ is a convex penalized estimator. Although a variant of the above construction was briefly presented in [7], Section 6, (there the function $z_0 \to \boldsymbol{X}\boldsymbol{Q}_0(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}$ is used), important questions remain unanswered to prove the validity of the general confidence interval in (1.21) and its applicability to commonly used regularized estimators.

1.3. *The rest of the paper is organized as follows.* Section 2 develops an L_2 bound between $\xi/\operatorname{Var}[\xi]^{1/2}$ and $N(0,1)$ for random variables of the form $\xi = z^\top f(z) - \operatorname{div} f(z)$ where $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$. Section 3 uses this normal approximation to show the asymptotic normality of (1.18) and proves the consistency of the variance estimate $V^*(\theta)$ in (1.19) in the regime where p and n are of the same order in the linear model (1.1) with correlated design. Section 4 provides closed-form formulas to apply the results in Section 3 to the Lasso, the group Lasso and twice continuously differentiable penalty functions. Section 7 contains the proofs of the results in Section 3. Appendix A provides a technical lemma on the integrability of smallest eigenvalue of Wishart matrices, Appendix B provides the proofs of the asymptotic normality results for the Lasso and group Lasso when $p > n$ and Appendix C contains the proofs of the derivative formulae for the group Lasso.

1.4. *Notation.* For two reals $\{a, b\}$, let $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$ and $a_+ = a \vee 0$. Let $\boldsymbol{I}_d$ be the identity matrix of size $d \times d$, for example, $d = n$, p. For any $p \geq 1$, let $[p]$ be the set $\{1, \ldots, p\}$. Let $\|\cdot\|$ be the Euclidean norm and $\|\cdot\|_q$ the ℓ_q norm of vectors for any $q \geq 1$, so that $\|\cdot\| = \|\cdot\|_2$. Let $\|\cdot\|_{\mathrm{op}}$ be the operator norm of matrices and $\|\cdot\|_F$ the Frobenius norm. Let $\phi_{\min}(\boldsymbol{S})$ be the smallest eigenvalue of a symmetric matrix $\boldsymbol{S}$. We use the notation $\langle\cdot,\cdot\rangle$ for the canonical scalar product of vectors in $\mathbb{R}^n$ or $\mathbb{R}^p$, that is, $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}^\top \boldsymbol{b}$ for two vectors $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension. For any event Ω, denote by I_Ω its indicator function. The unit sphere is $S^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\| = 1\}$. Convergence in distribution is denoted by $\to^d$ and convergence in probability by $\to^{\mathbb{P}}$. Throughout the paper, $C_0, C_1, \ldots$ denote positive absolute constants, $C_k(\gamma)$ positive constants depending on γ only and $C_k(\gamma, \mu)$ on $\{\gamma, \mu\}$ only.

For any vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$ and set $A \subset [p]$, the vector $\boldsymbol{v}_A \in \mathbb{R}^{|A|}$ is the restriction $(v_j)_{j\in A}$. For any $n \times p$ matrix $\boldsymbol{M}$ with columns $(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_p)$ and any subset $A \subset [p]$, let $\boldsymbol{M}_A = (\boldsymbol{M}_j, j \in A)$ be the matrix composed of columns of $\boldsymbol{M}$ indexed by A. If $\boldsymbol{M}$ is a symmetric matrix of size $p \times p$ and $A \subset [p]$, then $\boldsymbol{M}_{A,A}$ denotes the submatrix of $\boldsymbol{M}$ with rows and columns in A, and $\boldsymbol{M}_{A,A}^{-1}$ is the inverse of $\boldsymbol{M}_{A,A}$. For any square matrix $\boldsymbol{M}$, let $\boldsymbol{M}^s = (\boldsymbol{M} + \boldsymbol{M}^\top)/2$ be its symmetrization giving the same quadratic form.

For a vector valued map $h : \mathbb{R}^n \to \mathbb{R}^q$ with coordinates $h_1, \ldots, h_q : \mathbb{R}^n \to \mathbb{R}$, the gradient $\nabla h \in \mathbb{R}^{n\times q}$ is the matrix with columns $\nabla h_1, \ldots, \nabla h_q$. Thus, ∇h is the transpose of the Jacobian of h and $h(\boldsymbol{x} + \boldsymbol{\eta}) = h(\boldsymbol{x}) + \nabla h(\boldsymbol{x})^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|)$ if each coordinate h_i is Fréchet differentiable at $\boldsymbol{x}$. For deterministic matrices $\boldsymbol{A} \in \mathbb{R}^{m\times q}$, $\nabla(\boldsymbol{A}h) = (\nabla h)\boldsymbol{A}^\top \in \mathbb{R}^{n\times m}$. For f in (2.1), $\nabla f(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$ and the divergence is $\operatorname{div} f(\boldsymbol{x}) = \operatorname{trace}[\nabla f(\boldsymbol{x})]$.

2. Normal approximation in Stein's formula. We develop in this section normal approximations for random variables of the form

$$\xi = z^\top f(z) - \operatorname{div} f(z), \tag{2.1}$$

for which Stein's formula [37] states $\mathbb{E}[\xi] = 0$, where $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ is standard normal and $f : \mathbb{R}^n \to \mathbb{R}^n$. We establish L_2 bounds for the linear and quadratic approximations of ξ and construct consistent variance estimates in the related CLT.

Throughout this paper, the ith coordinate f_i of f is a function $f_i : \mathbb{R}^n \to \mathbb{R}$ and its weak gradient is denoted by ∇f_i. Similarly, the weak derivative of $g(z)$ is denoted by ∇g. We refer to [14], Section 1.5, for definitions of weak differentiability. For the application to asymptotic normality of debiased estimates in Section 3, the functions we will consider are locally Lipschitz. By Rademacher's theorem, locally Lipschitz functions are Fréchet differentiable almost everywhere, which is stronger than the existence of directional derivatives in all directions. In this case, the weak derivatives agree with the classical partial derivatives almost everywhere. As far as the application in Section 3 is concerned, the reader unfamiliar with

weak differentiability may consider the additional assumption that f is locally Lipschitz in the following results and replace weak derivatives with classical derivatives. The variance of (2.1) is given by the following proposition.

PROPOSITION 2.1 (Second-order Stein formula, [37], equation (8.6), [7]). *Let $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ be a function with each coordinate f_i being squared integrable and weakly differentiable with squared integrable gradient, that is, $\mathbb{E}[f_i(z)^2] + \mathbb{E}[\|\nabla f_i(z)\|^2] < +\infty$. Then*

$$\mathbb{E}\big[\big(z^\top f(z) - \operatorname{div} f(z)\big)^2\big] = \mathbb{E}\big[\|f(z)\|^2\big] + \mathbb{E}\operatorname{trace}\big[\{\nabla f(z)\}^2\big]. \tag{2.2}$$

The above result, in the twice differentiable case, was known to Stein [37], equation (8.6). If f is twice differentiable, the result follows by a sequence of integration by parts. The differentiability requirement was relaxed to only once weakly differentiable f in [7] where statistical applications of this formula to such once differentiable f are discussed.

2.1. *Linear approximation.* The goal of the present section is to derive normal approximations and CLT for the random variable (2.1). The intuition is as follows. We are looking for linear approximation of the random variable (2.1), of the form $z^\top \boldsymbol{\mu} \sim N(0, \|\boldsymbol{\mu}\|^2)$ for some deterministic $\boldsymbol{\mu} \in \mathbb{R}^n$. We rewrite (2.1) as

$$z^\top f(z) - \operatorname{div} f(z) = \underbrace{z^\top \boldsymbol{\mu}}_{\text{linear part}} + \underbrace{z^\top\big(f(z) - \boldsymbol{\mu}\big) - \operatorname{div} f(z)}_{\text{remainder}}. \tag{2.3}$$

The remainder term above is mean zero with second moment equal to $\mathbb{E}[\|f(z) - \boldsymbol{\mu}\|^2] + \mathbb{E}\operatorname{trace}[\{\nabla f(z)\}^2]$ by Proposition 2.1. This second moment is minimized for $\boldsymbol{\mu} = \mathbb{E}[f(z)]$, hence $z^\top \mathbb{E}[f(z)]$ gives the best linear approximation of ξ in (2.1). The following result provides conditions on f under which the remainder term is negligible in (2.3).

THEOREM 2.2. *Let $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ and f be a function $f : \mathbb{R}^n \to \mathbb{R}^n$, with each coordinate f_i being squared integrable and weakly differentiable with squared integrable gradient, that is, $\mathbb{E}[f_i(z)^2] + \mathbb{E}[\|\nabla f_i(z)\|^2] < +\infty$. Then $\xi = z^\top f(z) - \operatorname{div} f(z)$ satisfies*

$$\mathbb{E}\big[\big(\xi/\operatorname{Var}[\xi]^{1/2} - Z\big)^2\big] = \epsilon_1^2 + \big(1 - \big(1 - \epsilon_1^2\big)^{1/2}\big)^2 = \epsilon_1^2 + c_1\epsilon_1^4 \tag{2.4}$$

with $Z = z^\top \mathbb{E}[f(z)]/\|\mathbb{E}[f(z)]\| \sim N(0,1)$, deterministic real $1/4 \le c_1 \le 1$ and

$$\epsilon_1^2 \overset{\text{def}}{=} 1 - \frac{\|\mathbb{E}[f(z)]\|^2}{\operatorname{Var}[\xi]} \le \overline{\epsilon}_1^2 \le \frac{2\mathbb{E}[\|\nabla f(z)\|_F^2]}{\mathbb{E}[\|f(z)\|^2] + \mathbb{E}[\|\nabla f(z)\|_F^2]}, \tag{2.5}$$

where $\overline{\epsilon}_1^2 \overset{\text{def}}{=} 2\mathbb{E}[\|\{\nabla f(z)\}^s\|_F^2]/\{\|\mathbb{E}[f(z)]\|^2 + 2\mathbb{E}[\|\{\nabla f(z)\}^s\|_F^2]\}$. Consequently, $\sup_{t\in\mathbb{R}} |\mathbb{P}(\xi/\operatorname{Var}[\xi]^{1/2} \le t) - \mathbb{P}(Z \le t)| \le C(\epsilon_1^2 + c_1\epsilon_1^4)^{1/3}$ for $C = 1 + (2\pi)^{-1/2}$.

A direct consequence of Theorem 2.2 is $\epsilon_1^2 \le (2.4) \le 2\epsilon_1^2 \le 2\overline{\epsilon}_1^2$. Inequality (2.4) provides an upper bound on the 2-Wasserstein distance between $\xi/\operatorname{Var}[\xi]^{1/2}$ and $Z \sim N(0,1)$. When $\epsilon_1^2 \to 0$, it gives a stronger L_2 form of the CLT $\xi/\operatorname{Var}[\xi]^{1/2} \to^d N(0,1)$ in addition to the Kolmogorov distance bound in Theorem 2.2. The theorem follows from Proposition 2.1 and an application of the Gaussian Poincaré inequality.

PROOF OF THEOREM 2.2. Define $Z = z^\top \mathbb{E}[f(z)]/\|\mathbb{E}[f(z)]\|$ then $Z \sim N(0,1)$ and

$$\xi - \operatorname{Var}[\xi]^{1/2} Z = z^\top g(z) - \operatorname{div} g(z),$$

where $g(z) = f(z) - r\mathbb{E}f(z)$ and $r = (\mathrm{Var}[\xi]^{1/2}/\|\mathbb{E}f(z)\|)$. By Proposition 2.1 applied to g and a bias-variance decomposition,

$$\begin{aligned}
&\mathbb{E}\big[(\xi - \mathrm{Var}[\xi]^{1/2}Z)^2\big] \\
&\quad = \mathbb{E}\big\|f(z) - r\mathbb{E}\big[f(z)\big]\big\|^2 + \mathbb{E}\,\mathrm{trace}\big[\{\nabla f(z)\}^2\big] \\
&\quad = \mathbb{E}\big\|f(z) - \mathbb{E}\big[f(z)\big]\big\|^2 + \mathbb{E}\,\mathrm{trace}\big[\{\nabla f(z)\}^2\big] + \big\{\mathrm{Var}[\xi]^{1/2} - \big\|\mathbb{E}f(z)\big\|\big\}^2 \\
&\quad = \mathrm{Var}[\xi] - \big\|\mathbb{E}f(z)\big\|^2 + \big\{\mathrm{Var}[\xi]^{1/2} - \big\|\mathbb{E}f(z)\big\|\big\}^2
\end{aligned}$$

thanks to $(r-1)\|\mathbb{E}f(z)\| = \mathrm{Var}[\xi]^{1/2} - \|\mathbb{E}f(z)\|$. Thus, (2.4) follows from the definition of ϵ_1^2 in (2.5). Moreover, $\mathbb{E}[\|f(z) - \mathbb{E}[f(z)]\|^2] \le \mathbb{E}[\|\nabla f(z)\|_F^2]$ by the Gaussian Poincaré inequality and $\|\boldsymbol{M}\|_F^2 + \mathrm{trace}(\boldsymbol{M}^2) = 2\|\boldsymbol{M}^s\|_F^2$ for $\boldsymbol{M} \in \mathbb{R}^{n\times n}$. Hence, with $a = \|\mathbb{E}f(z)\|^2$, $b = \mathbb{E}\|f(z) - \mathbb{E}f(z)\|^2$, $c = \mathbb{E}\,\mathrm{trace}[\{\nabla f(z)\}^2]$ and $d = \mathbb{E}[\|\{\nabla f(z)\}^s\|_F^2]$ we have

$$\epsilon_1^2 = \frac{b+c}{a+b+c} \le \frac{2d}{a+2d} = \overline{\epsilon}_1^2 \le \frac{2\mathbb{E}[\|\nabla f(z)\|_F^2]}{a+2\mathbb{E}[\|\nabla f(z)\|_F^2]} \le \frac{2\mathbb{E}[\|\nabla f(z)\|_F^2]}{\mathbb{E}[\|f(z)\|^2 + \|\nabla f(z)\|_F^2]}$$

thanks to another Gaussian Poincaré inequality for the last inequality. Finally, $x^2/4 \le (1 - \sqrt{1-x})^2 \le x^2$ holds for all $x \in [0,1]$, which proves $c_1 \in [1/4, 1]$.

For any $\delta > 0$, by Markov's inequality $\mathbb{P}(\xi/\mathrm{Var}[\xi]^{1/2} \le t) - \mathbb{P}(Z \le t) \le \mathbb{P}(|\xi/\mathrm{Var}[\xi]^{1/2} - Z| > \delta) + \mathbb{P}(Z \in [t, t+\delta]) \le (\epsilon_1^2 + c_1\epsilon_1^4)/\delta^2 + \delta(2\pi)^{-1/2}$ since the standard normal pdf is uniformly bounded by $(2\pi)^{-1/2}$. Hence, with $\delta = (\epsilon_1^2 + c_1\epsilon_1^4)^{1/3}$, the above and a similar argument on $[t-\delta, t]$ provide the Kolmogorov distance bound. □

Normal approximation results such as Theorem 2.2 are flexible tools as they let us derive asymptotic normality results by mechanically computing gradients: By Theorem 2.2, it suffices to show that the expectation of $\|\nabla f(z)\|_F^2$ is negligible compared with that of $\|f(z)\|^2$ to obtain $\xi/\mathrm{Var}[\xi]^{1/2} \to^d N(0,1)$. Normal approximations involving derivatives have been studied for random variables with the more general form $W = g(z)$ for differentiable functions $g : \mathbb{R}^n \to \mathbb{R}$. The second-order Poincaré inequality of [19] bounds the total variation distance d_{TV} of $g(z)$ to the Gaussian distribution using the first and second derivatives of g: [19], Theorem 2.2, specialized to $W = g(z)$ with $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ states that

$$d_{\mathrm{TV}}\big\{W, N\big(\mu_0, \sigma_0^2\big)\big\} \le \big(2\sqrt{5}/\sigma_0^2\big)\mathbb{E}\big[\|\nabla g(z)\|^4\big]^{1/4}\mathbb{E}\big[\|\nabla^2 g(z)\|_{\mathrm{op}}^4\big]^{1/4}, \tag{2.6}$$

where $W = g(z)$, $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$, $\mu_0 = \mathbb{E}[W]$ and $\sigma_0^2 = \mathrm{Var}[W]$. Above, ∇g, $\nabla^2 g$ denote the gradient and Hessian matrix of g. Inequality (2.6) provides a CLT for $g(z)$ provided that the moments of the derivatives $\mathbb{E}[\|\nabla g(z)\|^4]^{1/4}$ and $\mathbb{E}[\|\nabla^2 g(z)\|_{\mathrm{op}}^4]^{1/4}$ are negligible compared to the variance $\sigma_0^2 = \mathrm{Var}[g(z)]$. Inequality (2.6) has been successfully applied to derive asymptotic normality of unregularized M-estimators when $p/n \to \gamma < 1$ and the M-estimation loss is twice differentiable [31]. However, the (2.6) based approach is not applicable for regularized estimators such as the Lasso and group Lasso that are only once differentiable functions of $(\boldsymbol{X}, \boldsymbol{y})$. In fact, by Proposition 4.1 below, the Lasso is not twice differentiable as $\mathrm{trace}[(\partial/\partial \boldsymbol{y})\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})]$ is integer-valued. In Theorem 2.2, while $\xi = z^\top f(z) - \mathrm{div}\, f(z)$ already involves the derivatives of f through the divergence, the ratio $\overline{\epsilon}_1^2$ that appears in the upper bound (2.5) only involves f and its gradient ∇f; the second derivatives of f need not exist. Section 3 uses Theorem 2.2 to provide asymptotic normality for debiasing estimators that are only once differentiable.

Variance estimate. It follows from Theorem 2.2 that random variables ξ of the form (2.1) are asymptotically normal under the condition $1 - \|\mathbb{E}[f(z)]\|^2/\operatorname{Var}[\xi] \to 0$, or under a somewhat stronger but more explicit condition $\overline{\epsilon}_1^2 \to 0$ as in (2.5). The following theorem provides consistent estimates of $\operatorname{Var}[\xi]$.

THEOREM 2.3. *Let f, z, ξ, ϵ_1^2 and $c_1 \in [1/4, 1]$ be as in Theorem 2.2. Then*

$$\begin{aligned}\mathbb{E}\big[\big(\|f(z)\|/\operatorname{Var}[\xi]^{1/2} - 1\big)^2\big] &\le \epsilon_1^2 - \mathbb{E}\big[\operatorname{trace}\big(\{\nabla f(z)\}^2\big)\big]/\operatorname{Var}[\xi] + c_1\epsilon_1^4 \\ &\le (1-\epsilon_1^2)\overline{\overline{\epsilon}}_1^2/(2-2\overline{\overline{\epsilon}}_1^2) + c_1\epsilon_1^4\end{aligned} \tag{2.7}$$

with $\overline{\overline{\epsilon}}_1^2 \overset{\text{def}}{=} 2\mathbb{E}[\|\nabla f(z)\|_F^2]/\{\|\mathbb{E}[f(z)]\|^2 + 2\mathbb{E}[\|\nabla f(z)\|_F^2]\} \ge \epsilon_1^2$. *Consequently,*

$$\|f(z)\|^2/\operatorname{Var}[\xi] \to^{\mathbb{P}} 1 \quad \text{and} \quad \xi/\|f(z)\| \to^{d} N(0,1) \tag{2.8}$$

when $\epsilon_1^2 + \overline{\overline{\epsilon}}_1^2 I\{\mathbb{E}[\operatorname{trace}(\{\nabla f(z)\}^2)] < 0\} \to 0$.

PROOF OF THEOREM 2.3. It follows from the Jensen inequality and (2.4) that

$$\begin{aligned}\mathbb{E}\big[\big(\|f(z)\|/\operatorname{Var}[\xi]^{1/2} - 1\big)^2\big] &\le \mathbb{E}\big[\|f(z)\|^2\big]/\operatorname{Var}[\xi] + 1 - 2\|\overline{\boldsymbol{\mu}}\|/\operatorname{Var}[\xi]^{1/2} \\ &= \epsilon_1^2 - \mathbb{E}\big[\operatorname{trace}\big(\{\nabla f(z)\}^2\big)\big]/\operatorname{Var}[\xi] + c_1\epsilon_1^4\end{aligned}$$

with $\overline{\boldsymbol{\mu}} = \mathbb{E}[\nabla f(z)]$ due to $\epsilon_1^2 = 1 - \|\overline{\boldsymbol{\mu}}\|^2/\operatorname{Var}[\xi]$. For the second inequality in (2.7),

$$\epsilon_1^2 - \mathbb{E}\big[\operatorname{trace}\big(\{\nabla f(z)\}^2\big)\big]/\operatorname{Var}[\xi] = \mathbb{E}\big[\|f(z) - \overline{\boldsymbol{\mu}}\|^2\big]/\operatorname{Var}[\xi] \le \mathbb{E}\big[\|\nabla f(z)\|_F^2\big]/\operatorname{Var}[\xi],$$

thanks to the Gaussian Poincaré inequality, and $(1-\epsilon_1^2)\overline{\overline{\epsilon}}_1^2/(2-2\overline{\overline{\epsilon}}_1^2)$ equals to the right-hand side above by the definition of $\overline{\overline{\epsilon}}_1^2$. □

2.2. *Quadratic approximation.* The decomposition (2.3) is especially useful if the linear part $z^\top\boldsymbol{\mu}$ with $\boldsymbol{\mu} = \mathbb{E}[f(z)]$ is a good approximation for $\xi = z^\top f(z) - \operatorname{div} f(z)$. In some cases, for example, if $f(z) = \boldsymbol{A}z$ for some square deterministic matrix $\boldsymbol{A}$, the decomposition (2.3) is uninformative. It is then natural to look for the best quadratic approximation of ξ in the sense of the L_2 orthogonal projection to

$$\mathscr{H}_{1,2} = \{\xi_{\boldsymbol{\mu},\boldsymbol{A}} = \boldsymbol{\mu}^\top z + z^\top\boldsymbol{A}z - \operatorname{trace}[\boldsymbol{A}] : \boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n\times n}\} = \mathscr{H}_1 \oplus \mathscr{H}_2,$$

where $\mathscr{H}_1 = \{\boldsymbol{\mu}^\top z : \boldsymbol{\mu} \in \mathbb{R}^n\}$ and $\mathscr{H}_2 = \{z^\top\boldsymbol{A}z - \operatorname{trace}[\boldsymbol{A}] : \boldsymbol{A} \in \mathbb{R}^{n\times n}\}$ are L_2 subspaces orthogonal to each other.

The calculation in (2.4) for $\mathscr{H}_1$ is generic in the following sense. If $\overline{\xi}$ is the L_2 projection of a random variable ξ in L_2, then the sine of the L_2-angle between $\overline{\xi}$ and ξ is $\epsilon = (\mathbb{E}[(\xi - \overline{\xi})^2]/\mathbb{E}[\xi^2])^{1/2} = (1 - \mathbb{E}[\overline{\xi}^2]/\mathbb{E}[\xi^2])^{1/2}$ and

$$\mathbb{E}\big[\big(\xi/\operatorname{Var}[\xi]^{1/2} - \overline{\xi}/\operatorname{Var}[\overline{\xi}]^{1/2}\big)^2\big] = 2\big(1 - \sqrt{1-\epsilon^2}\big) = \epsilon^2 + c\epsilon^4 \tag{2.9}$$

holds for some deterministic real $1/4 \le c \le 1$. Indeed, take $\epsilon = \sin\alpha$ with α being the L_2-angle between ξ and $\overline{\xi}$, so that (2.9) becomes $(2\sin(\alpha/2))^2 = 2(1-\cos(\alpha)) = \epsilon^2 + c\epsilon^4$ as in the proof of Theorem 2.2.

The next result extends Theorem 2.2 to the L_2 quadratic projections to $\mathscr{H}_2$ and $\mathscr{H}_{1,2}$, and also gives Theorem 2.2 the interpretation as the L_2 projection to $\mathscr{H}_1$.

THEOREM 2.4. *Let $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$, $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfy the assumption of Theorem 2.2, and $\xi = z^\top f(z) - \operatorname{div} f(z)$. For $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ let $\xi_{\boldsymbol{\mu},\boldsymbol{A}} = z^\top(\boldsymbol{\mu} + \boldsymbol{A}z) - \operatorname{trace} \boldsymbol{A}$. Let $\overline{\boldsymbol{\mu}} = \mathbb{E}[f(z)]$ and $\overline{\boldsymbol{A}} = \mathbb{E}[\nabla f(z)]$. Then $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}$ is the L_2 projection of ξ to $\mathscr{H}_{1,2}$ and*

$$\begin{aligned}\mathbb{E}\big[(\xi - \xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}})^2\big] &= \mathbb{E}\big[\big\|f(z) - \overline{\boldsymbol{\mu}}\big\|^2 - \|\overline{\boldsymbol{A}}\|_F^2\big] + \mathbb{E}\operatorname{trace}\big[\big\{\nabla f(z) - \overline{\boldsymbol{A}}\big\}^2\big] \\ &\le 2\mathbb{E}\big[\big\|\big\{\nabla f(z) - \overline{\boldsymbol{A}}\big\}^s\big\|_F^2\big].\end{aligned} \tag{2.10}$$

Consequently, $\xi_{\overline{\boldsymbol{\mu}},\mathbf{0}}$ is the projection of ξ and $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}$ to $\mathscr{H}_1$ with $\mathbb{E}[(\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}} - \xi_{\overline{\boldsymbol{\mu}},\mathbf{0}})^2] = 2\|\overline{\boldsymbol{A}}^s\|_F^2$ and $\xi_{\mathbf{0},\overline{\boldsymbol{A}}}$ is the projection of ξ and $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}$ to $\mathscr{H}_2$ with $\mathbb{E}[(\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}} - \xi_{\mathbf{0},\overline{\boldsymbol{A}}})^2] = \|\overline{\boldsymbol{\mu}}\|^2$.

For the projection $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}$ of ξ to $\mathscr{H}_{1,2}$, $\epsilon_{1,2}^2 \stackrel{\text{def}}{=} 1 - \mathbb{E}[\xi^2_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}]/\mathbb{E}[\xi^2]$ satisfies $\epsilon_{1,2}^2 \le \overline{\epsilon}_{1,2}^2 \stackrel{\text{def}}{=} 2\mathbb{E}[\|\{\nabla f(z) - \overline{\boldsymbol{A}}\}^s\|_F^2]/\{\|\overline{\boldsymbol{\mu}}\|^2 + 2\mathbb{E}[\|\{\nabla f(z)\}^s\|_F^2]\}$ and under the condition $\epsilon_{1,2}^2 = o(1)$,

$$\|\overline{\boldsymbol{A}}^s\|_{\text{op}}^2/\big(\|\overline{\boldsymbol{\mu}}\|_2^2 + \|\overline{\boldsymbol{A}}^s\|_F^2\big) \to 0 \quad \Leftrightarrow \quad \xi/\operatorname{Var}[\xi]^{1/2} \to^d N(0,1). \tag{2.11}$$

For the projection $\xi_{\mathbf{0},\overline{\boldsymbol{A}}}$ of ξ to $\mathscr{H}_2$, $\epsilon_2^2 = 1 - \mathbb{E}[\xi^2_{\mathbf{0},\overline{\boldsymbol{A}}}]/\mathbb{E}[\xi^2]$ satisfies $\epsilon_2^2 \le \overline{\epsilon}_2^2 \stackrel{\text{def}}{=} \{\|\overline{\boldsymbol{\mu}}\|^2 + 2\mathbb{E}[\|\{\nabla f(z) - \overline{\boldsymbol{A}}\}^s\|_F^2]\}/\{\|\overline{\boldsymbol{\mu}}\|^2 + 2\mathbb{E}[\|\{\nabla f(z)\}^s\|_F^2]\}$ and under the condition $\epsilon_2^2 = o(1)$,

$$\|\overline{\boldsymbol{A}}^s\|_{\text{op}}^2/\|\overline{\boldsymbol{A}}^s\|_F^2 \to 0 \quad \Leftrightarrow \quad \xi/\operatorname{Var}[\xi]^{1/2} \to^d N(0,1). \tag{2.12}$$

PROOF OF THEOREM 2.4. The function $g(z) = f(z) - \boldsymbol{\mu} - \boldsymbol{A}^\top z$ has gradient $\nabla g = \nabla f - \boldsymbol{A}$. Application of the second-order Stein's formula in Proposition 2.1 to g yields

$$\mathbb{E}\big[(\xi - \xi_{\boldsymbol{\mu},\boldsymbol{A}})^2\big] = \mathbb{E}\big[\big\|f(z) - \boldsymbol{\mu} - \boldsymbol{A}^\top z\big\|^2\big] + \mathbb{E}\operatorname{trace}\big[\big\{\nabla f(z) - \boldsymbol{A}\big\}^2\big] \stackrel{\text{def}}{=} I + II.$$

The first term is $I = \mathbb{E}[\|f(z) - \boldsymbol{\mu}\|^2] + \|\boldsymbol{A}\|_F^2 - 2\mathbb{E}[z^\top \boldsymbol{A}(f(z) - \boldsymbol{\mu})]$. By Stein's formula and the linearity of the trace, we have

$$\begin{aligned}\|\boldsymbol{A}\|_F^2 - 2\mathbb{E}\big[z^\top \boldsymbol{A}\big(f(z) - \boldsymbol{\mu}\big)\big] &= \|\boldsymbol{A}\|_F^2 - 2\mathbb{E}\operatorname{trace}\big(\nabla f(z)\boldsymbol{A}^\top\big) \\ &= \|\boldsymbol{A}\|_F^2 - 2\operatorname{trace}\big[\boldsymbol{A}^\top\overline{\boldsymbol{A}}\big] \\ &= -\|\overline{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_F^2.\end{aligned}$$

We also have $\mathbb{E}[\|\nabla f(z) - \overline{\boldsymbol{A}}\|_F^2] = \mathbb{E}[\|\nabla f(z)\|_F^2] - \|\overline{\boldsymbol{A}}\|_F^2$ so that

$$I = \mathbb{E}\big[\big\|f(z) - \boldsymbol{\mu}\big\|^2 - \big\|\nabla f(z)\big\|_F^2\big] + \mathbb{E}\big[\big\|\nabla f(z) - \overline{\boldsymbol{A}}\big\|_F^2\big] + \|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_F^2.$$

For the second term, using that $\mathbb{E}[\nabla f(z) - \overline{\boldsymbol{A}}] = 0$ we get

$$II = \mathbb{E}\operatorname{trace}\big[\big\{\nabla f(z) - \boldsymbol{A}\big\}^2\big] = \mathbb{E}\operatorname{trace}\big[\big\{\nabla f(z) - \overline{\boldsymbol{A}}\big\}^2\big] + \operatorname{trace}\big[\{\overline{\boldsymbol{A}} - \boldsymbol{A}\}^2\big].$$

Due to $\|\boldsymbol{M}\|_F^2 + \operatorname{trace}(\boldsymbol{M}^2) = 2\|\boldsymbol{M}^s\|_F^2$ for $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, it follows that

$$\begin{aligned}\mathbb{E}\big[(\xi - \xi_{\boldsymbol{\mu},\boldsymbol{A}})^2\big] &= \mathbb{E}\big[\big\|f(z) - \overline{\boldsymbol{\mu}}\big\|^2 - \|\overline{\boldsymbol{A}}\|_F^2\big] + \mathbb{E}\operatorname{trace}\big[\big\{\nabla f(z) - \overline{\boldsymbol{A}}\big\}^2\big] \\ &\quad + \|\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}\|^2 + 2\big\|(\boldsymbol{A} - \overline{\boldsymbol{A}})^s\big\|_F^2\end{aligned}$$

The optimality of $\boldsymbol{\mu} = \overline{\boldsymbol{\mu}}$ and $\boldsymbol{A} = \overline{\boldsymbol{A}}$ follows, so that $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}$ is the L_2 projection of ξ to $\mathscr{H}_{1,2}$. Also, the first line above gives the formula of $\mathbb{E}[(\xi - \xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}})^2]$ in (2.10), and the second line gives the formulas for the variances of $\xi_{\overline{\boldsymbol{\mu}},\mathbf{0}}$ and $\xi_{\mathbf{0},\overline{\boldsymbol{A}}}$. The upper bound in (2.10) follows from $\mathbb{E}[\|f(z) - \overline{\boldsymbol{\mu}}\|^2 - \|\overline{\boldsymbol{A}}\|_F^2] \le \mathbb{E}[\|\nabla f(z)\|_F^2] - \|\overline{\boldsymbol{A}}\|_F^2 = \mathbb{E}[\|\nabla f(z) - \overline{\boldsymbol{A}}\|_F^2]$ thanks to the Gaussian Poincaré inequality. Inequality (2.10) is equivalent to

$$\mathbb{E}\big[\xi^2\big] = \mathbb{E}\big[\xi^2_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}\big] + \mathbb{E}\big[(\xi - \xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}})^2\big] \le \|\overline{\boldsymbol{\mu}}\|^2 + 2\big\|\overline{\boldsymbol{A}}^s\big\|_F^2 + 2\mathbb{E}\big[\big\|\big\{\nabla f(z) - \overline{\boldsymbol{A}}\big\}^s\big\|_F^2\big],$$

which provides $\epsilon_{1,2}^2 \le \overline{\epsilon}_{1,2}^2$ and $\epsilon_2^2 \le \overline{\epsilon}_2^2$ by bounding from above the denominator in $\epsilon_{1,2}^2 = 1 - \mathbb{E}[\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}^2]/\mathbb{E}[\xi^2]$ and $\epsilon_2^2 = 1 - \mathbb{E}[\xi_{\mathbf{0},\overline{\boldsymbol{A}}}^2]/\mathbb{E}[\xi^2]$.

For (2.11), we write $\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}} = \sum_{j=1}^n \{a_j G_j + b_j(G_j^2 - 1)\}$ with i.i.d. $G_j \sim N(0,1)$, where $a_j = \boldsymbol{u}_j^\top \overline{\boldsymbol{\mu}}$ and $G_j = \boldsymbol{u}_j^\top \boldsymbol{z}$ with the eigenvalue decomposition $\overline{\boldsymbol{A}}^s = \sum_{j=1}^n b_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$. Assume without loss of generality that $\operatorname{Var}(a_j G_j + b_j(G_j^2 - 1)) = a_j^2 + 2b_j^2$ is nonincreasing in j and that $\operatorname{Var}[\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}]$ satisfies $\sum_{j=1}^n (a_j^2 + 2b_j^2) = 1$. The condition on the left-hand side of (2.11) implies that the integer $k_n \overset{\text{def}}{=} \lceil \|\overline{\boldsymbol{A}}^s\|_{\text{op}}^{-1} \rceil = \lceil (\max_j b_j)^{-1} \rceil$ satisfies $k_n \to +\infty$ and $\sum_{j=1}^{k_n} b_j^2 \ll 1 = \operatorname{Var}[\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}}]$, so that

$$\xi_{\overline{\boldsymbol{\mu}},\overline{\boldsymbol{A}}} = \sum_{j=1}^{k_n} a_j G_j + \sum_{j=k_n+1}^{n} \{a_j G_j + b_j(G_j^2 - 1)\} + o_{\mathbb{P}}(1).$$

Assuming that $\sum_{j=1}^{k_n} a_j^2 \to c$ for some $c \in [0,1]$ by extracting a subsequence if necessary, $k_n \to +\infty$ implies $\max_{j>k_n}(a_j^2 + 2b_j^2) = a_{k_n+1}^2 + 2b_{k_n+1}^2 \to 0$ so that the second term above is independent of the first and approximately $N(0, 1-c)$ by the Lyapunov CLT when $\sum_{j=1}^{k_n} a_j^2 \to c \le 1$. This proves that the LHS of (2.11) implies the RHS. Conversely, assume the asymptotic normality on the RHS so that $\sum_{j=1}^n \{a_j G_j + b_j(G_j^2 - 1)\} \to N(0,1)$. Let $W_j = a_j G_j + b_j(G_j^2 - 1)$ and $j_n \le n$. As W_{j_n} is an independent component of the sum, for any $(a_{j_n}, b_{j_n}) \to (a,b)$ along a subsequence with $a^2 + b^2 > 0$, we must have $b = 0$ because $W_{j_n} \to^d N(0, a^2 + 2b^2)$ by the Cramér–Lévi theorem and $W_{j_n} \to^d aG + b(G^2+1)$ for some $G \sim N(0,1)$. As $j_n \le n$ are arbitrary, this gives $\|\overline{\boldsymbol{A}}^s\|_{\text{op}} = \max_{j=1,\dots,n} b_j^2 \to 0$. $\square$

Variance estimate: *Quadratic case*. Theorem 2.4 provides the quadratic normal approximation of ξ under the condition $\overline{\epsilon}_{1,2}^2 \to 0$ with

$$\overline{\epsilon}_{1,2}^2 \ge \epsilon_{1,2}^2 = 1 - \big(\|\mathbb{E}[f(\boldsymbol{z})]\|^2 + 2\|\{\mathbb{E}[\nabla f(\boldsymbol{z})]\}^s\|_F^2\big)/\operatorname{Var}[\xi], \tag{2.13}$$

where $\overline{\epsilon}_{1,2}^2$ is defined using the upper bound $\|\mathbb{E} f(\boldsymbol{z})\|^2 + 2\mathbb{E}[\|\{\nabla f(\boldsymbol{z})\}^s\|_F^2] \ge \operatorname{Var}[\xi]$ established in (2.10) in the denominator on the right-hand side of (2.13).

THEOREM 2.5. *Let* $f, \boldsymbol{z}, \xi, \overline{\boldsymbol{\mu}} = \mathbb{E}[f(\boldsymbol{z})]$ *and* $\overline{\boldsymbol{A}} = \mathbb{E}[\nabla f(\boldsymbol{z})]$ *be as in Theorem 2.4 and* $\widehat{\operatorname{Var}[\xi]} = \|f(\boldsymbol{z})\|^2 + \operatorname{trace}[\{\nabla f(\boldsymbol{z})\}^2]$. *Then*

$$\mathbb{E}\big[\big|\widehat{\operatorname{Var}[\xi]}/\operatorname{Var}[\xi] - 1\big|\big] \le 2\overline{\overline{\epsilon}}_{1,2}^2 + 2\overline{\overline{\epsilon}}_{1,2} C_0 + C_0 \|\overline{\boldsymbol{A}}\|_{\text{op}}/\operatorname{Var}[\xi]^{1/2} \tag{2.14}$$

with $\overline{\overline{\epsilon}}_{1,2} \overset{\text{def}}{=} \{2\mathbb{E}[\|\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\|_F^2]/\operatorname{Var}[\xi]\}^{1/2}$ *and* $C_0 \overset{\text{def}}{=} \{(\|\overline{\boldsymbol{\mu}}\|^2 + 2\|\overline{\boldsymbol{A}}\|_F^2)/\operatorname{Var}[\xi]\}^{1/2}$. *Consequently, under the conditions* $\{\mathbb{E}[\|\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\|_F^2] + \|\overline{\boldsymbol{A}}\|_{\text{op}}\}/(\|\overline{\boldsymbol{\mu}}\|^2 + \|\overline{\boldsymbol{A}}^s\|_F^2) = o(1)$ *and* $\|\overline{\boldsymbol{A}}\|_F^2/(\|\overline{\boldsymbol{\mu}}\|^2 + \|\overline{\boldsymbol{A}}^s\|_F^2) = O(1)$,

$$\widehat{\operatorname{Var}[\xi]}/\operatorname{Var}[\xi] \to^{\mathbb{P}} 1 \quad \text{and} \quad \big(\widehat{\operatorname{Var}[\xi]}\big)^{-1/2} \xi \to^d N(0,1). \tag{2.15}$$

It follows from the second-order Stein formula in Proposition 2.1 that $\widehat{\operatorname{Var}[\xi]}$ is an unbiased estimator of $\operatorname{Var}[\xi]$. Moreover, when $\mathbb{E}[\|\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\|_F^2]$, $\|\overline{\boldsymbol{A}}\|_F$ and $\|\overline{\boldsymbol{A}}\|_{\text{op}}$ are all equivalent to their symmetric counterparts, $\mathbb{E}[\|\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\|_F^2] \asymp \mathbb{E}[\|\{\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\}^s\|_F^2]$, $\|\overline{\boldsymbol{A}}\|_F \asymp \|\overline{\boldsymbol{A}}^s\|_F$ and $\|\overline{\boldsymbol{A}}\|_{\text{op}} \asymp \|\overline{\boldsymbol{A}}^s\|_{\text{op}}$, the condition for (2.15) holds if and only if $\epsilon_{1,2}^2 + \|\overline{\boldsymbol{A}}^s\|_{\text{op}}^2/(\|\overline{\boldsymbol{\mu}}\|^2 + \|\overline{\boldsymbol{A}}^s\|_F^2) = o(1)$ for the quantities in (2.13) and (2.11). The proof is given in Appendix D.

3. Debiasing general convex regularizers. Our main application of the normal approximation in Theorem 2.2 concerns debiasing regularized estimators of the form

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b})\} \tag{3.1}$$

for convex $g : \mathbb{R}^p \to \mathbb{R}$ in the linear model (1.1). Throughout, let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ be the error vector, $\boldsymbol{a}_0 \in \mathbb{R}^p$ be a direction of interest, $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ be the target of statistical inference, and $\boldsymbol{u}_0$, z_0 and $\boldsymbol{Q}_0$ be as in (1.13) so that (1.14) holds.

3.1. *Assumption.* We say that g is μ-strongly convex with respect to the norm $\boldsymbol{b} \to \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{b}\|$ if its symmetric Bregman divergence is bounded from below as

$$(\tilde{\boldsymbol{b}} - \boldsymbol{b})^\top \big((\partial g)(\tilde{\boldsymbol{b}}) - (\partial g)(\boldsymbol{b})\big) \ge \mu \|\boldsymbol{\Sigma}^{1/2}(\tilde{\boldsymbol{b}} - \boldsymbol{b})\|^2 \tag{3.2}$$

for some $\mu > 0$. Here, the interpretation of (3.2) is its validity for all choices in the subdifferential $(\partial g)(\tilde{\boldsymbol{b}})$ and $(\partial g)(\boldsymbol{b})$. Condition (3.2) holds for any convex g for $\mu = 0$. If g is twice differentiable, (3.2) holds if and only if $\mu\boldsymbol{\Sigma}$ is a lower bound for the Hessian of g. However, (3.2) may also hold for nondifferentiable g, for example, the elastic-net penalty with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Our results require the following assumption.

ASSUMPTION 3.1. (i) *Let $\gamma > 0$, $\mu \in [0, \frac{1}{2}]$ be constants such that $\mu + (1-\gamma)_+ > 0$, that is, either $\mu > 0$ or $\gamma < 1$ must hold. Consider a sequence of regression problems (1.1) with $n, p \to +\infty$ and $p/n \le \gamma$. The penalty $g : \mathbb{R}^p \to \mathbb{R}$ in (3.1) is convex and (3.2) holds. The rows of $\boldsymbol{X}$ are i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with invertible $\boldsymbol{\Sigma}$ and the noise $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ is independent of $\boldsymbol{X}$.* (ii) *$\boldsymbol{a}_0 \in \mathbb{R}^p$ is a sequence of vectors normalized with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$.*

Note that if (3.2) holds for $\mu \ge 0$ it also holds for $\mu' = \min(\frac{1}{2}, \mu)$ and we may thus assume $\mu \in [0, \frac{1}{2}]$ without loss of generality. Strongly convex objective functions admit unique minimizers. Since $\gamma < 1$ implies $\mathbb{P}(\phi_{\min}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}) > 0) = 1$ (cf. Appendix A) and the objective function of the optimization problem (3.1) is $(\phi_{\min}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}/n) + \mu)$-strongly convex, Assumption 3.1 grants almost surely the existence and uniqueness of the minimizer (3.1).

3.2. *Gradient with respect to $\boldsymbol{y}$ and effective degrees-of-freedom.* Consider a penalized estimator (3.1) viewed as a function $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$. For every $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, the map $\boldsymbol{y} \mapsto \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is 1-Lipschitz (cf. Proposition 7.3). By Rademacher's theorem, for almost every $\boldsymbol{y}$ there exists a unique matrix $\widehat{\boldsymbol{H}} \in \mathbb{R}^{n\times n}$ such that

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y} + \boldsymbol{\eta}, \boldsymbol{X}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) + \widehat{\boldsymbol{H}}^\top\boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|), , \tag{3.3}$$

as in (1.7), that is, $\widehat{\boldsymbol{H}}$ is the gradient of the map $\boldsymbol{y} \mapsto \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$. Furthermore, $\widehat{\boldsymbol{H}}$ is symmetric with eigenvalues in $[0, 1]$; see Proposition 7.3 for the existence of $\widehat{\boldsymbol{H}}$ and its properties. While existence of $\widehat{\boldsymbol{H}}$ was assumed in (1.7) in the Introduction, for penalized estimators (3.1) the matrix $\widehat{\boldsymbol{H}}$ provably exists for almost every $\boldsymbol{y}$ by Proposition 7.3.

Table 1 provides closed-form expressions of $\widehat{\boldsymbol{H}}$ for specific penalty functions g. The proofs of these closed-form expressions will be given in Section 4. An advantage of defining $\widehat{\boldsymbol{H}}$ as the Fréchet derivative of the Lipschitz map $\boldsymbol{y} \mapsto \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is that this definition applies to any convex penalty g, even though for arbitrary penalty g we are unable to provide a closed-form expression for $\widehat{\boldsymbol{H}}$. Finally, define the effective degrees-of-freedom $\widehat{\mathsf{df}}$ of $\widehat{\boldsymbol{\beta}}$ by

$$\widehat{\mathsf{df}} = \operatorname{trace}[\widehat{\boldsymbol{H}}] \tag{3.4}$$

as in (1.10). Because $\widehat{\boldsymbol{H}}$ is symmetric with eigenvalues in $[0, 1]$ (cf. Proposition 7.3), $0 \le \widehat{\mathsf{df}} \le n$ holds almost surely. The matrix $\widehat{\boldsymbol{H}}$ and the scalar $\widehat{\mathsf{df}}$ play a major role in our analysis.

TABLE 1

Closed-form expressions $\widehat{\boldsymbol{H}}$ from equation (3.3) for specific convex penalty functions $g : \mathbb{R}^p \to \mathbb{R}$. For the Lasso and elastic-net, $\widehat{S} = \{j \in [p] : \widehat{\beta}_j \neq 0\}$. For the group Lasso, $\widehat{S}$ and $\boldsymbol{M}$ are given in Section 4.3

Penalty	$\widehat{\boldsymbol{H}} \in \mathbb{R}^{n\times n}$	Justification
$g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1$ (Lasso)	$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$	[44], Proposition 4.1
$g(\boldsymbol{b}) = \mu\|\boldsymbol{b}\|_2^2$ (Ridge)	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + n\mu\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top$	(1.12), Section 4.1
$g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1 + \mu\|\boldsymbol{b}\|_2^2$ (Elastic-Net)	$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + n\mu\boldsymbol{I}_{\|\widehat{S}\|})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$	[44], (28), [7], Section 3.5.3,
$g(\boldsymbol{b}) = \|\boldsymbol{b}\|_{\mathrm{GL}} = \sum_{k=1}^K \lambda_k\|\boldsymbol{b}_{G_k}\|_2$ (group Lasso (3.33))	$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$	[45], Proposition 4.2
$g(\boldsymbol{b})$ twice continuously differentiable	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + n\nabla^2 g(\widehat{\boldsymbol{\beta}}))^{-1}\boldsymbol{X}^\top$	Section 4.1
$g(\boldsymbol{b})$ arbitrary convex function	symmetric with eigenvalues in [0, 1]	Proposition 7.3

3.3. *Approximation for $\xi_0 = z_0^\top f(z_0) - \operatorname{div} f(z_0)$ and the debiased vector $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$.* Consider, for a fixed value of $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ the function $f_{(\boldsymbol{X}\boldsymbol{Q}_0,\boldsymbol{\varepsilon})} : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$f_{(\boldsymbol{X}\boldsymbol{Q}_0,\boldsymbol{\varepsilon})}(z_0) = f(z_0) = \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}. \tag{3.5}$$

For brevity, we will often omit the dependence on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ of f as discussed after (1.16). The Fréchet gradient $\nabla f(z_0)$, where it exists, is uniquely defined by

$$f_{(\boldsymbol{X}\boldsymbol{Q}_0,\boldsymbol{\varepsilon})}(z_0+\boldsymbol{\eta}) - f_{(\boldsymbol{X}\boldsymbol{Q}_0,\boldsymbol{\varepsilon})}(z_0) = \big[\nabla f(z_0)\big]^\top\boldsymbol{\eta} + o\big(\|\boldsymbol{\eta}\|\big) \tag{3.6}$$

and the divergence by $\operatorname{div} f(z_0) = \operatorname{trace}[\nabla f(z_0)]$. If $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p}(\|\boldsymbol{\varepsilon} - \widetilde{\boldsymbol{X}}(\boldsymbol{b}-\boldsymbol{\beta})\|^2/(2n) + g(\boldsymbol{b}))$ with $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}_0^\top$, then (3.6) is equivalent to

$$\big(\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}\big) - (\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) = \big[\nabla f(z_0)\big]^\top\boldsymbol{\eta} + o\big(\|\boldsymbol{\eta}\|\big). \tag{3.7}$$

By Stein's formula, we have conditionally on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ that almost surely

$$\mathbb{E}\big[\xi_0 | (\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})\big] = 0 \quad \text{for } \xi_0 = z_0^\top f(z_0) - \operatorname{div} f(z_0). \tag{3.8}$$

As in (1.18) for the general case discussed in the Introduction, (3.8) gives an unbiased estimating equation for $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$. The next lemma provides an expression for $\nabla f(z_0)$.

LEMMA 3.1. *Let Assumption 3.1(i) be fulfilled, $\boldsymbol{a}_0 \in \mathbb{R}^p$ and $\widehat{\boldsymbol{H}}$ be as in (3.3). Then*

$$\nabla f(z_0)^\top = (\boldsymbol{I}_n - \widehat{\boldsymbol{H}})^\top\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + \boldsymbol{w}_0(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \tag{3.9}$$

satisfies (3.6) for some random $\boldsymbol{w}_0 \in \mathbb{R}^n$ almost surely. If additionally $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$, then

$$\|\boldsymbol{w}_0\|^2 \le n^{-1}\min\big\{(4\mu)^{-1}, \phi_{\min}\big(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}/n\big)^{-1}\big\}. \tag{3.10}$$

Lemma 3.1 is proved in Section 7.1. Although we do not use this fact in any results, we mention here in passing that vector $\boldsymbol{w}_0$ in (3.9) is linear in $\boldsymbol{a}_0$ in the sense that $\boldsymbol{w}_0$ can be chosen of the form $\boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0$ for some matrix $\boldsymbol{W} \in \mathbb{R}^{n\times p}$. Indeed, the proof of Lemma 3.1 shows that the map $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}$ is Fréchet differentiable at almost every point by Rademacher's theorem. At such a point, with $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^p$, $t \in \mathbb{R}$, the linear combination $\boldsymbol{a}_3 = \boldsymbol{a}_1 + t\boldsymbol{a}_2$ and the perturbed design matrix $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\eta}(\boldsymbol{a}_1 + t\boldsymbol{a}_2)^\top$, linearity of the Fréchet derivative implies that

$$\begin{aligned}&\big(\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}\big) - (\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\\ &\quad = \big(\langle \boldsymbol{a}_1 + t\boldsymbol{a}_2, \boldsymbol{h}\rangle(\boldsymbol{I}_n - \widehat{\boldsymbol{H}})^\top + (\boldsymbol{w}_1 + t\boldsymbol{w}_2)(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top\big)\boldsymbol{\eta} + o\big(\|\boldsymbol{\eta}\|\big),\end{aligned}$$

TABLE 2
Closed-form expressions for $\boldsymbol{w}_0 \in \mathbb{R}^n$ in Lemma 3.1 for specific convex penalties $g : \mathbb{R}^p \to \mathbb{R}$

Penalty	Vector $\boldsymbol{w}_0 \in \mathbb{R}^n$ in Lemma 3.1	Justification
$g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1$ (Lasso)	$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}(\boldsymbol{a}_0)_{\widehat{S}}$	[8], Proposition 4.1
$g(\boldsymbol{b}) = \mu\|\boldsymbol{b}\|_2^2$ (Ridge)	$\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\mu \boldsymbol{I}_p)^{-1}\boldsymbol{a}_0$	Section 4.1
$g(\boldsymbol{b}) = \|\boldsymbol{b}\|_{\mathrm{GL}} = \sum_{k=1}^K \lambda_k \|\boldsymbol{b}_{G_k}\|_2$ (group Lasso (3.33))	$\boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}(\boldsymbol{a}_0)_{\widehat{S}}$	Proposition 4.2
$g(\boldsymbol{b})$ twice continuously differentiable	$\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\nabla^2 g(\widehat{\boldsymbol{\beta}}))^{-1}\boldsymbol{a}_0$	Section 4.1

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ denote the $\boldsymbol{w}_0$ from (3.9) for $\boldsymbol{a}_0 = \boldsymbol{a}_1$ and $\boldsymbol{a}_0 = \boldsymbol{a}_2$, respectively. Hence, with $\boldsymbol{w}_3 = \boldsymbol{w}_1 + t\boldsymbol{w}_2$, (3.9) holds for $(\boldsymbol{a}_0, \boldsymbol{w}_0) = (\boldsymbol{a}_3, \boldsymbol{w}_3)$. This proves that $\boldsymbol{w}_0$ is linear in $\boldsymbol{a}_0$, that is, it is of the form $\boldsymbol{w}_0 = \boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0$ for some matrix $\boldsymbol{W} \in \mathbb{R}^{n\times p}$. One way to construct $\boldsymbol{W}$ explicitly is the following: define the jth column of $\boldsymbol{W}$ as the vector $\boldsymbol{w}_0$ corresponding to $\boldsymbol{a}_0 = \boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is the jth canonical basis vector. The linearity proved above for any linear combination $\boldsymbol{a}_3$ then implies that (3.9) holds for $\boldsymbol{w}_0 = \boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0$ for any $\boldsymbol{a}_0 \in \mathbb{R}^p$. Inequality (3.10) provides an upper bound on the operator norm of $\boldsymbol{W}$. Linearity of $\boldsymbol{w}_0$ with respect to $\boldsymbol{a}_0$ and explicit matrices $\boldsymbol{W}$ can be seen for some penalty functions in Table 2. Such Fréchet differentiability with respect to $\boldsymbol{X}$ is used in [4, 6] to develop estimates of $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$ in linear models with Gaussian covariates similar to the present paper.

By taking the trace of (3.9), we obtain almost surely under Assumption 3.1,

$$\begin{aligned} -\xi_0 &= \operatorname{div} f(z_0) - \langle z_0, f(z_0)\rangle \\ &= (n - \widehat{\mathsf{df}})\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big) + \langle z_0 + \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle \\ &= (n - \widehat{\mathsf{df}})(\widehat{\theta} - \theta) \end{aligned} \tag{3.11}$$

for

$$\widehat{\theta} \stackrel{\mathrm{def}}{=} \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle + (n - \widehat{\mathsf{df}})^{-1}\langle z_0 + \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle. \tag{3.12}$$

In the present context of the regularized estimator $\widehat{\boldsymbol{\beta}}$ in (3.1) with its effective degrees-of-freedom $\widehat{\mathsf{df}}$ defined in (3.4) and under Assumption 3.1, the quantities ξ_0, $\widehat{\mathsf{df}}$, $\widehat{\theta}$ in the previous display coincide with the random variables with the same name in (1.18). By (3.8), equality $\mathbb{E}[\xi_0] = 0$ holds so that

$$\begin{aligned} 0 &= \mathbb{E}[-\xi_0] \\ &= \mathbb{E}\big[(n - \widehat{\mathsf{df}})(\widehat{\theta} - \theta)\big] \\ &= \mathbb{E}\big[(n - \widehat{\mathsf{df}})\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big) + \langle z_0 + \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle\big] \end{aligned} \tag{3.13}$$

by taking expectations in (3.11). This provides a first evidence that the correction $(n - \widehat{\mathsf{df}})^{-1}\langle z_0 + \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle$ indeed removes the bias, at least after multiplication of $(\widehat{\theta} - \theta)$ by $(n - \widehat{\mathsf{df}})$. Since $z_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0$ under the normalization (1.15), the unbiased estimating equation (3.13) is the specialization of (1.11) from the Introduction to the penalized estimator (3.1), for which we have the gradient expression (3.9) in terms of $\boldsymbol{w}_0$ and $\widehat{\boldsymbol{H}}$. If $\widehat{\boldsymbol{\beta}}$ is given by (3.1), by identifying the terms in (1.18) and (3.11) we see that the random variable $\widehat{A}$ defined in (1.10) of the Introduction is here $\widehat{A} = \langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle$ with $\boldsymbol{w}_0$ given by Lemma 3.1. The following lemma shows that this term is negligible.

LEMMA 3.2. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma, \mu)n^{-1/2}$ and*

$$\mathbb{E}\big[I_{\Omega_n}\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2 / \operatorname{Var}_0[\xi_0]\big] \le C_1(\gamma, \mu)n^{-1}. \tag{3.14}$$

The proof is given in Section 7.5. Since $\mathbb{P}(\Omega_n) \to 1$, inequality (3.14) implies $\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2 / \operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 0$. This motivates the definition

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} \stackrel{\mathrm{def}}{=} \widehat{\boldsymbol{\beta}} + (n - \widehat{\mathsf{df}})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{3.15}$$

The debiased estimate $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}\rangle$ in direction $\boldsymbol{a}_0$ is obtained from $\widehat{\theta}$ in (3.12) by dropping the smaller order term $(n - \widehat{\mathsf{df}})^{-1}\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle$. By Slutsky's theorem, (3.14) implies

$$\frac{\xi_0}{\operatorname{Var}_0[\xi_0]^{1/2}} \to^d F \quad \text{if and only if} \quad \frac{(n - \widehat{\mathsf{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle}{\operatorname{Var}_0[\xi_0]^{1/2}} \to^d F \tag{3.16}$$

for any candidate limiting distribution F. As $\mathbb{E}[\xi_0] = 0$, this suggests that the simpler correction in (3.15) also corrects the bias of $\widehat{\boldsymbol{\beta}}$. By Prohorov's theorem, there exists a subsequence and limiting distribution F such that (3.16) holds in this subsequence. While F is mean zero as $\xi_0/\operatorname{Var}_0[\xi_0]$ has mean zero and variance one, F has variance at most one by Fatou's lemma. However, the normality of F is unclear at this point.

To obtain more precise information on the limiting distribution and the deviations of ξ_0, the next subsections build estimate of its variance and derive asymptotic normality results by showing that $F = N(0, 1)$ for most directions $\boldsymbol{a}_0$. The next result provides a loose data-driven upper bound on the error $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}\rangle - \theta$.

THEOREM 3.3. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma, \mu)n^{-1/2}$ and*

$$\mathbb{E}\big[I_{\Omega_n}(n - \widehat{\mathsf{df}})^2\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle^2 / \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\big] \le C_2(\gamma, \mu). \tag{3.17}$$

Furthermore, $|\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle| = O_{\mathbb{P}}(1)\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/(n - \widehat{\mathsf{df}}) = O_{\mathbb{P}}(1)\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/n$.

Theorem 3.3 is proved in Section 7.6. If $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n = O_{\mathbb{P}}(\sigma^2)$, then $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/n = O_{\mathbb{P}}(1)\sigma/\sqrt{n}$ is of the same order as the width of confidence intervals based on the least-squares estimator as $n \to +\infty$ while p remains fixed. Theorem 3.3 shows that under this mild additional assumption on the prediction error $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/n$, the second term in (3.15) indeed corrects the bias, achieving $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle = O_{\mathbb{P}}(1)\sigma/\sqrt{n}$.

3.4. *Variance estimates.* By Proposition 2.1, the conditional variance $\operatorname{Var}_0[\xi_0]$ can be written as $\operatorname{Var}_0[\xi_0] = \mathbb{E}_0[V^*(\theta)]$ for

$$V^*(\theta) \stackrel{\mathrm{def}}{=} \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \operatorname{trace}\big[\{\nabla f(z_0)\}^2\big]. \tag{3.18}$$

We allow the variance estimate to depend on the unknown $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ as the resulting pivotal quantity, $-V^*(\theta)^{-1/2}\xi_0 = V^*(\theta)^{-1/2}(n - \widehat{\mathsf{df}})(\widehat{\theta} - \theta)$ via (3.11), would depend on θ anyway. While $V^*(\theta)$ itself can be used to estimate $\operatorname{Var}_0[\xi_0]$, its sign is unclear. The following simplified version of it, obtained by removing the smaller order terms in $V^*(\theta)$,

$$\begin{aligned} \widehat{V}(\theta) &\stackrel{\mathrm{def}}{=} \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \operatorname{trace}\big[(\widehat{\boldsymbol{H}} - \boldsymbol{I}_n)^2\big]\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big)^2 \\ &= \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\widehat{\boldsymbol{H}} - \boldsymbol{I}_n\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2, \end{aligned} \tag{3.19}$$

is nonnegative. This follows from Proposition 7.3 since $\boldsymbol{I}_n - \widehat{\boldsymbol{H}}$ is almost surely positive semidefinite. Lemma 3.4 below shows that the relative bias $\mathbb{E}_0[\widehat{V}(\theta)]/\operatorname{Var}_0[\xi_0] - 1$ converges to 0 in probability, that is, $\widehat{V}(\theta)$ is asymptotically unbiased for $\operatorname{Var}_0[\xi_0]$.

LEMMA 3.4. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma, \mu)n^{-1/2}$ and*

$$\mathbb{E}\left[I_{\Omega_n}\left|\frac{\mathbb{E}_0[\widehat{V}(\theta)]}{\operatorname{Var}_0[\xi_0]} - 1\right|\right] \le \mathbb{E}\left[I_{\Omega_n}\frac{\mathbb{E}_0[|\widehat{V}(\theta) - V^*(\theta)|]}{\operatorname{Var}_0[\xi_0]}\right] \le \frac{C_3(\gamma, \mu)}{n}. \tag{3.20}$$

An alternative variance estimate that does not depend on the unknown parameter θ is given by replacing θ in $\widehat{V}(\theta)$ by the point estimate $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} \rangle$ with $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ in (3.15):

$$\check{V}(\boldsymbol{a}_0) = \widehat{V}\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} \rangle\big) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \frac{\langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \rangle^2}{(n - \widehat{\mathrm{df}})^2}. \tag{3.21}$$

The next lemma provides $\check{V}(\boldsymbol{a}_0)/\widehat{V}(\theta) \to^{\mathbb{P}} 1$ and that $\check{V}(\boldsymbol{a}_0)$ is also asymptotically unbiased in the sense $\mathbb{E}_0[\check{V}(\boldsymbol{a}_0)]/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$. Lemmas 3.4 and 3.5 are proved in Section 7.5.

LEMMA 3.5. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\max\left\{ \mathbb{E}\left[I_{\Omega_n} \left| \frac{\check{V}(\boldsymbol{a}_0)^{1/2}}{\widehat{V}(\theta)^{1/2}} - 1 \right|^2 \right], \mathbb{E}\left[I_{\Omega_n} \left| \frac{\mathbb{E}_0[\check{V}(\boldsymbol{a}_0)]^{1/2}}{\mathbb{E}_0[\widehat{V}(\theta)]^{1/2}} - 1 \right|^2 \right] \right\} \le \frac{C_4(\gamma,\mu)}{n}. \tag{3.22}$$

3.5. *Asymptotic normality of debiased estimates.* Throughout this section, $\Phi(t) = \mathbb{P}(N(0,1) \le t)$ denotes the standard normal cdf. For a given penalty function $g : \mathbb{R}^p \to \mathbb{R}$, we define the deterministic oracle $\boldsymbol{\beta}^*$ and its associated noiseless prediction risk R_* by

$$\begin{aligned} \boldsymbol{\beta}^* &\stackrel{\mathrm{def}}{=} \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{b})\|^2/2 + g(\boldsymbol{b}), \\ \boldsymbol{h}^* &\stackrel{\mathrm{def}}{=} \boldsymbol{\beta}^* - \boldsymbol{\beta}, \\ R_* &\stackrel{\mathrm{def}}{=} \sigma^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2. \end{aligned} \tag{3.23}$$

Our first result provides asymptotic normality of the debiased estimate when the error $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$ of $\widehat{\boldsymbol{\beta}}$ in direction $\boldsymbol{a}_0$ is negligible compared to R_*.

THEOREM 3.6. *Let Assumption 3.1 be fulfilled. Let $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ be as in (3.15). Then, for any $\boldsymbol{a}_0$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h} \rangle^2/R_* \to^{\mathbb{P}} 0$,*

$$\sup_{t\in\mathbb{R}} \left[\left| \mathbb{P}\left(\frac{\xi_0}{V_0^{1/2}} \le t \right) - \Phi(t) \right| + \left| \mathbb{P}\left(\frac{\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} \rangle - \theta}{V_0^{1/2}/(n - \widehat{\mathrm{df}})} \le t \right) - \Phi(t) \right| \right] \to 0,$$

where V_0 denotes any of the four quantities: $\operatorname{Var}_0[\xi_0]$, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{V}(\theta)$ *or* $\check{V}(\boldsymbol{a}_0)$.

Theorem 3.6 is proved in Section 7.6. The theorem, as well as its variants below, are obtained by applying Theorem 2.2 conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ to the function $f(z_0)$ in (3.5). This argument relies on the normality of z_0 conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$, and thus the Gaussian design assumption. Here is an outline. Define

$$\delta_1^2(\boldsymbol{a}_0) \stackrel{\mathrm{def}}{=} \mathbb{E}_0[\|\nabla f(z_0)\|_F^2]/\{\mathbb{E}_0[\|f(z_0)\|^2] + \mathbb{E}_0[\|\nabla f(z_0)\|_F^2]\}, \tag{3.24}$$

where $\mathbb{E}_0$ and Var_0 are the conditional expectation and conditional variance given $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$. It is sufficient to show that $\delta_1^2(\boldsymbol{a}_0) \to^{\mathbb{P}} 0$ in order to prove asymptotic normality of $\xi_0/\operatorname{Var}_0[\xi_0]^{1/2}$ by Theorem 2.2 and of $\xi_0/\|f(z_0)\|$ by Theorem 2.3 since $\delta_1^2 = \delta_1^2(\boldsymbol{a}_0)$ satisfies $2\delta_1^2 \ge \max(\epsilon_1^2, \bar{\epsilon}_1^2)$ for the ϵ_1^2, $\bar{\epsilon}_1^2$ in Theorems 2.2 and 2.3. The proof makes rigorous the following informal bound:

$$\delta_1^2(\boldsymbol{a}_0) = \frac{\mathbb{E}_0[\|\nabla f(z_0)\|_F^2]}{\mathbb{E}_0[\|f(z_0)\|^2] + \mathbb{E}_0[\|\nabla f(z_0)\|_F^2]} \lesssim \frac{\mathbb{E}_0[\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \langle \boldsymbol{a}_0, \boldsymbol{h} \rangle^2]}{C_*^2(\gamma,\mu) n R_*} + O_{\mathbb{P}}(n^{-1/2})$$

for some constant $C_*(\gamma,\mu)$, by establishing a lower bound on $\|f(z_0)\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ for the denominator (Lemmas 7.4, 7.6 and 7.7), and by showing that the rank one term $\boldsymbol{w}_0(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top$

in (3.9) is negligible in the numerator. Finally, $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n$ always holds by Proposition 7.3 and $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*$ is shown to be uniformly integrable, so that the assumption $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ grants $\mathbb{E}[\delta_1^2(\boldsymbol{a}_0)] \to 0$. The next two results identify directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ holds.

THEOREM 3.7. *There exists an absolute constant $C^* > 0$ such that the following holds. Let Assumption 3.1 be fulfilled, $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ be as in (3.15). Then for any increasing sequence $a_p \to +\infty$ (e.g., $a_p = \log\log p$), the subset*

$$\overline{S} = \{\boldsymbol{v} \in S^{p-1} : \mathbb{E}[\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2] \le C^*/a_p\} \tag{3.25}$$

of the unit sphere S^{p-1} in $\mathbb{R}^p$ has relative volume $|\overline{S}|/|S^{p-1}| \ge 1 - 2e^{-p/a_p}$ and

$$\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \sup_{t\in\mathbb{R}} \left[\left| \mathbb{P}\left(\frac{\xi_0}{V_0^{1/2}} \le t \right) - \Phi(t) \right| + \left| \mathbb{P}\left(\frac{\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle}{V_0^{1/2}/(n - \widehat{\mathsf{df}})} \le t \right) - \Phi(t) \right| \right] \to 0 \tag{3.26}$$

where V_0 denotes any of the four quantities: $\mathrm{Var}_0[\xi_0]$, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{V}(\theta)$ *or* $\check{V}(\boldsymbol{a}_0)$. *Furthermore, with $\boldsymbol{e}_j \in \mathbb{R}^p$ the jth canonical basis vector and $\phi_{\mathrm{cond}}(\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}}$, the asymptotic normality in (3.26) uniformly holds over at least $(p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*)$ canonical directions in the sense that $J_p = \{j \in [p] : \boldsymbol{e}_j/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\| \in \boldsymbol{\Sigma}^{1/2}\overline{S}\}$ has cardinality $|J_p| \ge p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*$.*

Theorem 3.7 is proved in Section 7.6. For a given sequence of directions $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}S^{p-1}$, if $b_p = \mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2] \to 0$ then it follows by choosing $a_p = C^*/b_p$ that $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ for the $\overline{S}$ in (3.25) so that (3.26) implies that asymptotic normality holds for this sequence of $\boldsymbol{a}_0$. In other words, asymptotic normality holds for all $\boldsymbol{a}_0$ such that $\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2] \to 0$. Thus, a sequence of directions $\boldsymbol{a}_0$ for which asymptotic normality does not follow from Theorem 3.7 is a sequence such that $\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2]$ does not vanish, that is, the squared error $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2$ in direction $\boldsymbol{a}_0$ carries a constant fraction of the full prediction error $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2$. Such direction $\boldsymbol{a}_0$ must thus be very special, which is embodied by the exponentially small bound on the relative volume $|\overline{S} \setminus S^{p-1}|/|S^{p-1}|$.

THEOREM 3.8. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma, \mu)n^{-1/2}$ and*

$$\mathbb{E}\big[I_{\Omega_n}\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle + (n - \widehat{\mathsf{df}})^{-1}\boldsymbol{z}_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big)^2\big] \le R_* C_5(\gamma, \mu)/n \tag{3.27}$$

If additionally g is a seminorm, then $|\boldsymbol{z}_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})|/n = |\boldsymbol{a}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})|/n \le g(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)$ always holds by properties of the subdifferential of a norm. Consequently, if $g(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)^2/R_* \to 0$ then $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ and the conclusions of Theorem 3.6 hold.

Theorem 3.8 is proved in Section 7.6. The first part of the theorem says that the estimation error $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle$ is essentially $-\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle/(n - \widehat{\mathsf{df}})$ up to an error term of order $R_* O_{\mathbb{P}}(n^{-1/2})$, so that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ if and only if $(n - \widehat{\mathsf{df}})^{-2}\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/R_* \to^{\mathbb{P}} 0$. Combined with the fact that $(1 - \widehat{\mathsf{df}}/n)$ is bounded away from 0 by Lemma 7.4, this implies that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ if and only if $n^{-2}\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/R_* \to^{\mathbb{P}} 0$. The last part of the theorem relies on the property $g(\boldsymbol{u}) \le \sup_{\boldsymbol{s}\in\partial g(\boldsymbol{u})} \boldsymbol{u}^\top\boldsymbol{s}$ for any norm g where $\partial g(\boldsymbol{u})$ is the subdifferential of g at $\boldsymbol{u}$. This property also holds if g is a seminorm; however, it is unclear how to extend the last part of the above theorem if g is not a seminorm.

For the Lasso, the penalty function is $g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1$ and condition $g(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)^2/R_* \to 0$ becomes $\lambda^2\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1^2/R_* \to 0$. Typically, the tuning parameter λ is chosen as $\lambda \propto \sigma(2\log(p/k)/n)^{1/2}$ with $k = 1$ [11], among others, or $k = s_0$ [3, 5, 25, 30, 39], where $s_0 =$

$\|\boldsymbol{\beta}\|_0$. For such choices, the condition $\lambda^2\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1^2/R_* \to 0$ can be written as $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1 = (R_*/\sigma^2)^{1/2}o(\sqrt{n/\log(p/k)})$ and since $R_* \geq \sigma^2$, a sufficient condition is $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1 = o(\sqrt{n/\log(p/k)})$. If $\boldsymbol{a}_0 = \boldsymbol{e}_j$ is a vector of the canonical basis, the normalization (1.15) gives $(\boldsymbol{\Sigma}^{-1})_{jj} = 1$ and $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1$ is the ℓ_1 norm of the jth column of $\boldsymbol{\Sigma}^{-1}$. The condition $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1 = o(\sqrt{n/\log(p/k)})$ allows, for instance, the jth column of $\boldsymbol{\Sigma}^{-1}$ to have $o(\sqrt{n/\log(p/k)})$ constant entries. This assumption is weaker than that of some previous studies; for instance, [28] requires $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0\|_1 = O(1)$ for $\boldsymbol{a}_0 = \boldsymbol{e}_j$. The following example illustrates the benefit of picking a proper penalty level λ.

EXAMPLE 1. *Let $p/n \to \gamma < 1$ and $g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1$.*

(i) *For $\lambda = 0$, the Lasso and debiased Lasso are both identical to the least squares estimator and the debiasing correction proportional to $\boldsymbol{z}_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ is 0 since $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$, so that $\widehat{\theta} - \theta = \langle \boldsymbol{a}_0, \boldsymbol{h}\rangle = \boldsymbol{a}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$ in (3.12), $\widehat{\mathsf{df}} = p$, $\widehat{V}(\theta) \approx \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \sim \sigma^2\chi^2_{n-p}$ and $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2/(1-\gamma))$.*

(ii) *Suppose $|\widehat{S}|/n + \|\boldsymbol{X}\boldsymbol{h}\|^2/n + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 = o_{\mathbb{P}}(1)$ for suitable $\lambda \geq \sigma\sqrt{2\log(p/s_0)/n}$ as in [8, 50]. Then $\widehat{V}(\theta) = (1 + o_{\mathbb{P}}(1))n\sigma^2$ and $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$.*

3.6. *Confidence intervals.* Theorems 3.6 to 3.8 are valid for any choice of the variance estimate among $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{V}(\theta)$ in (3.19) and $\check{V}(\boldsymbol{a}_0)$ in (3.22) for directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ holds. For such direction $\boldsymbol{a}_0$, the choice $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ leads to the narrowest confidence interval for θ, namely

$$\mathbb{P}\big(\theta \in \big[\big\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}\big\rangle \pm z_{\alpha/2}(n - \widehat{\mathsf{df}})^{-1}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|\big]\big) \to 1 - \alpha, \tag{3.28}$$

where $[u \pm v]$ denotes the interval $[u - v, u + v]$, $\mathbb{P}(|N(0,1)| > z_{\alpha/2}) = \alpha$ and $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ is the debiased estimator in (3.15). The choice $\check{V}(\boldsymbol{a}_0)$ leads to intervals with larger multiplicative coefficient for $z_{\alpha/2}$, namely

$$\theta \in \left[\big\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}\big\rangle \pm z_{\alpha/2}\left(\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{(n - \widehat{\mathsf{df}})^2} + \frac{\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2\langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2}{(n - \widehat{\mathsf{df}})^4}\right)^{1/2}\right] \tag{3.29}$$

has probability converging to $1 - \alpha$ for directions $\boldsymbol{a}_0$ satisfying any of the above theorems. For such directions, the choice $\widehat{V}(\theta)$ justifies confidence intervals of the form (1.21) as

$$\big((n - \widehat{\mathsf{df}})\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle - \theta\big) + \langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle\big)^2 - \widehat{V}(\theta)z_{\alpha/2}^2 \leq 0 \tag{3.30}$$

holds with probability converging to $1 - \alpha$. Given the expression for $\widehat{V}(\theta)$ in (3.19), the left-hand side of (3.30) is a quadratic polynomial in θ with dominant coefficient $(n - \widehat{\mathsf{df}})^2 - z_{\alpha/2}\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2$. Since $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \leq n - \widehat{\mathsf{df}}$ almost surely by properties of $\widehat{\boldsymbol{H}}$ in Proposition 7.3 and $n - \widehat{\mathsf{df}} \geq C_*(\gamma, \mu)n$ for some constant $C_*(\gamma, \mu)$ with probability approaching one by Lemma 7.4, in the same event the dominant coefficient is positive. The intersection of events (3.30) and $\{n - \widehat{\mathsf{df}} \geq C_*(\gamma, \mu)n\}$ has probability converging to $1 - \alpha$ and in this event, $\theta \in [\Theta_1(z_{\alpha/2}), \Theta_2(z_{\alpha/2}))]$ where $\Theta_1(z_{\alpha/2})$, $\Theta_2(z_{\alpha/2})$ are the two real roots of the left-hand side of (3.30) as a quadratic function of θ.

3.7. *Variance spike.* One can pick any choice among the three variance estimates in Theorem 3.6 because it assumes $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ and this limit in probability implies both $\widehat{V}(\theta)/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$ and $\check{V}(\boldsymbol{a}_0)/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$. These limits in probability to 1 are made rigorous by (3.22) and by the lower bound $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \geq R_*n(C_*^2(\gamma, \mu) - O_{\mathbb{P}}(n^{-1/2}))$ obtained from Lemmas 7.4, 7.6 and 7.7 as explained in (7.38) of the proof.

The reason that the estimates $\widehat{V}(\theta)$ and $\check{V}(\boldsymbol{a}_0)$ were introduced is that the simpler estimate $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ is not asymptotically unbiased for $\mathrm{Var}_0[\xi_0]$ for directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*$ does not converge to 0 in probability: While the relative bias of $\widehat{V}(\theta)$ and $\check{V}(\boldsymbol{a}_0)$ provably converges to 0 in Lemma 3.4 and (3.22) for all directions $\boldsymbol{a}_0$, the same cannot be said for the simpler estimate $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$.

THEOREM 3.9. *Let Assumption 3.1 be fulfilled. Then the following are equivalent*:

(i) $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/\mathrm{Var}_0[\xi_0] \to^{\mathbb{P}} 1$,
(ii) $\mathbb{E}_0[\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]/\mathrm{Var}_0[\xi_0] \to^{\mathbb{P}} 1$,
(iii) $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$,
(iv) $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 0$,
(v) $\langle z_0, \boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/(n\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2) \to^{\mathbb{P}} 0$,
(vi) $\widehat{V}(\theta)/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$,
(vii) $\check{V}(\boldsymbol{a}_0)/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$.

Theorem 3.9 is proved in Section 7.5. It shows that for the directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 0$ does not hold, for example, directions such that the error $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2$ is of the same order as the average squared residual $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n$ (see Lemma 7.2 in Section 7.2), the simpler estimate $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ fails to account for a nonnegligible part of the variance $\mathrm{Var}_0[\xi_0]$ by item (i) above. The goal of the estimates $\widehat{V}(\theta)$ and $\check{V}(\boldsymbol{a}_0)$ is to repair this as $\mathbb{E}_0[\widehat{V}(\theta)]/\mathrm{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ and $\mathbb{E}_0[\check{V}(\boldsymbol{a}_0)]/\mathrm{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ hold for all directions by Lemmas 3.4 and 3.5, even for directions $\boldsymbol{a}_0$ such that (i)–(vii) above fail. Note that the quantity $\langle z_0, \boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/(n\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2)$ in item (v) is observable (i.e., does not depend on $\boldsymbol{\beta}$), so that (i)–(vii) are expected to hold when this quantity is sufficiently small.

For directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 0$ does not hold, we expect a variance spike, that is, an extra additive term in the variance estimate equal to $\|\boldsymbol{I}_n-\widehat{\boldsymbol{H}}\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2$ in $\widehat{V}(\theta)$ and to $\|\boldsymbol{I}_n-\widehat{\boldsymbol{H}}\|_F^2\langle z_0, \boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/(n-\widehat{\mathrm{df}})^2$ in $\check{V}(\boldsymbol{a}_0)$. The confidence interval (3.28) that does not take into account this variance spike is expected to be too narrow and to suffer from undercoverage for directions $\boldsymbol{a}_0$ with large $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n/\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$. The wider confidence interval (3.29) is expected to repair this, although for directions $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 0$ does not hold the current theory does not prove whether the asymptotic distribution is normal. The theoretical evidence that the variance spike occurs is grounded in the relative asymptotic unbiasedness of $\widehat{V}(\theta)$ in (3.20) and of $\check{V}(\boldsymbol{a}_0)$ in (3.22), combined with the negative result for the simpler variance estimate $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ via Theorem 3.9 as discussed above. Figure 1 in Section 5 illustrates the variance spike on simulations for the Lasso and direction $\boldsymbol{a}_0$ proportional to the first canonical basis vector.

The second term in the variance estimates (3.19) and (3.21) is necessary for certain $\boldsymbol{a}_0$ for the estimate $\widehat{\boldsymbol{\beta}}=\boldsymbol{0}$, which corresponds to (1.2) with penalty $g(\boldsymbol{0})=0$ and $g(\boldsymbol{b})=+\infty$ for $\boldsymbol{b}\neq\boldsymbol{0}$. For $\boldsymbol{\Sigma}=\boldsymbol{I}_p$, $\boldsymbol{a}_0=\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$ and $V_*=2\|\boldsymbol{\beta}\|^2+\sigma^2$,

$$\frac{-\xi_0}{\widehat{V}(\theta)}=\frac{-n\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle+\boldsymbol{z}_0^\top\boldsymbol{y}}{(\|\boldsymbol{y}\|^2+n\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle^2)^{1/2}}=\frac{(nV_*)^{-1/2}\sum_{i=1}^n(-\|\boldsymbol{\beta}\|+z_{0i}\epsilon_i+\|\boldsymbol{\beta}\|z_{0i}^2)}{(\|\boldsymbol{y}\|^2/n+\|\boldsymbol{\beta}\|^2)^{1/2}/V_*^{1/2}}. \tag{3.31}$$

By the CLT, the numerator of the rightmost quantity converges to $N(0,1)$ and in the denominator, $(\|\boldsymbol{y}\|^2/n+\|\boldsymbol{\beta}\|^2)/V_* \to^{\mathbb{P}} 1$ by the weak law of large numbers, so that (3.31) converges in law to $N(0,1)$ by Slutsky's theorem. On the other hand, if the variance estimate $\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ is used instead of $\widehat{V}(\theta)$ in the denominator, the CLT $-\xi_0/\|\boldsymbol{y}\| \to^d N(0, V_*/(\|\boldsymbol{\beta}\|^2+\sigma^2))$ still holds but the asymptotic variance is $1+(1+\sigma^2/\|\boldsymbol{\beta}\|^2)^{-1}>1$.

3.8. *Relaxing strong convexity when $p > n$.* The previous theorems are valid under Assumption 3.1: Either $\gamma < 1$ and g is an arbitrary convex function, or $\gamma \geq 1$ and g is required to be strongly convex with parameter μ. If $p/n > 1$ and the penalty g is not strongly convex, the techniques of the present paper still provide asymptotic normality results under additional assumptions as shown in the following result.

Consider either the Lasso

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1\} \tag{3.32}$$

for some $\lambda > 0$ or the group Lasso norm $\|\cdot\|_{\mathrm{GL}}$ and group Lasso $\widehat{\boldsymbol{\beta}}$ defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \|\boldsymbol{b}\|_{\mathrm{GL}}\}, \qquad \|\boldsymbol{b}\|_{\mathrm{GL}} = \sum_{k=1}^{K} \lambda_k \|\boldsymbol{b}_{G_k}\|_2, \tag{3.33}$$

where $(G_1, \ldots, G_K)$ is a partition of $\{1, \ldots, p\}$ into K nonoverlapping groups and $\lambda_1, \ldots, \lambda_K > 0$ are tuning parameters. If each G_k is a singleton and $\lambda_k = \lambda > 0$ for all $k = 1, \ldots, p$, then (3.33) reduces to the Lasso (3.32).

THEOREM 3.10. *Let $\gamma \geq 1$, $\kappa \in (0, 1)$ be constants independent of $\{n, p\}$. Consider a sequence of regression problems with $p/n \leq \gamma$ and invertible $\boldsymbol{\Sigma}$. Assume that the group Lasso estimator $\widehat{\boldsymbol{\beta}}$ in (3.33) satisfies*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}\|_0 \leq \kappa n/2\big) \to 1. \tag{3.34}$$

If $\boldsymbol{a}_0$ is such that $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$ and $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle^2/R_* \xrightarrow{\mathbb{P}} 0$ for the R_* in (3.23), then

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^{-1}\big((n - \widehat{\mathsf{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle + z_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big) \leq t\big) - \Phi(t)\big| \to 0. \tag{3.35}$$

Furthermore, for any a_p with $a_p \to \infty$ and $\overline{S}$ in (3.25), the relative volume bound given after (3.25) holds, and the asymptotic normality (3.35) holds uniformly over all $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ and uniformly over at least $(p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*)$ canonical directions in the sense that $J_p = \{j \in [p] : \boldsymbol{e}_j/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\| \in \boldsymbol{\Sigma}^{1/2}\overline{S}\}$ has cardinality $|J_p| \geq p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*$.

Theorem 3.10 is proved in Appendix B. The strong convexity requirement in Assumption 3.1 is relaxed and replaced by assuming the high-probability bound $\|\widehat{\boldsymbol{\beta}}\|_0 \leq \kappa n/2$ on the number of nonzero coordinates. Surprisingly, no conditions are required on the true regression vector $\boldsymbol{\beta}$ or on the tuning parameters, although these quantities affect whether $\mathbb{P}(\|\widehat{\boldsymbol{\beta}}\|_0 \leq \kappa n/2) \to 1$ is satisfied. Figure 2 in Section 6 illustrates Theorem 3.10 on simulated data.

Comparison with existing works on the Lasso. The Lasso is largely the most studied initial estimator in previous literature on debiasing and asymptotic normality, so it provides a level playing field to compare our method with existing results. In the approximate message passing (AMP) literature, which includes most existing works in the $n/p \to \gamma$ regime, for example, [21, 23, 27] or more recently [33, 40, 42, 43], it is assumed that $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and that the empirical distribution $G_{n,p}(t) = p^{-1}\sum_{j=1}^{p} I\{\sqrt{n}\beta_j \leq t\}$ converges in distribution and in the second moment to some "prior" G as $n, p \to +\infty$. Assume these conditions and consider the jth component $\widehat{\beta}_j^{(\mathrm{de-bias})}$ of $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ in (3.15) for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, that is,

$$\widehat{\beta}_j^{(\mathrm{de-bias})} = \widehat{\beta}_j + \langle \boldsymbol{X}\boldsymbol{e}_j, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle/(n - \widehat{\mathsf{df}}).$$

Then the Lasso has the interpretation as its soft thresholded debiased version,

$$\widehat{\beta}_j = \eta\big(\widehat{\beta}_j^{(\mathrm{de-bias})}; \lambda/(1-\widehat{\mathrm{df}}/n)\big) \quad \text{where } \eta(u;t) = \mathrm{sgn}(u)\big(|u|-t\big)_+$$

and the main thrust of the AMP theory is that the joint empirical distribution of the debiased errors and the true coefficients,

$$H_{n,p}(u,t) = p^{-1}\sum_{j=1}^{p} I\big\{\sqrt{n}\widehat{\beta}_j^{(\mathrm{de-bias})} - \sqrt{n}\beta_j \le u, \sqrt{n}\beta_j \le t\big\},$$

converges in distribution and the second moment to the limit H with independent $N(0,\tau_0)$ and G components, where τ_0 is characterized by a system of nonlinear equations with 2 or 3 unknowns. These nonlinear equations depend on the loss (here, the ℓ_2 loss), the penalty (here, the ℓ_1-norm), the distribution of the noise, as well as the prior distribution that governs the empirical distribution of the coefficients of $\boldsymbol{\beta}$. We note that these works typically assume that $\boldsymbol{X}$ has $N(0,1/n)$ entries, so that their coefficient vector is equivalent to our $\sqrt{n}\boldsymbol{\beta}$. For instance, [33] characterizes the limit of the empirical distribution of $(\sqrt{n}\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}, \sqrt{n}\boldsymbol{\beta})$ in terms of two parameters, $\{\tau_*(\lambda), \kappa_*(\lambda)\}$, that are defined as solutions of the nonlinear equations in [33], Proposition 3.1; see also [17], Proposition 4.3, for similar results applicable to permutation-invariant penalty. This approach presents some drawbacks: For instance, it requires the convergence of the empirical distribution $G_{n,p}$ to a limit (which can be viewed as a prior), it yields the limiting distribution for the joint empirical distribution $H_{n,p}$ of the estimation errors and the unknown coefficients but not for a fixed coordinate.

The above Theorem 3.10 for the Lasso differs from this previous literature in major ways. First, it provides a limiting distribution for the de-biased version of $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}\rangle$ for a single, fixed direction $\boldsymbol{a}_0$: Theorem 3.10 does not involve the empirical distributions of $\sqrt{n}\boldsymbol{\beta}$, $\sqrt{n}\widehat{\boldsymbol{\beta}}$ or its debiased version. This contrasts with previous literature on the $n/p \to \gamma$ regime where the confidence interval guarantee holds on average over the coefficients $\{1,\dots,p\}$ [21, 23, 27, 40]. This improvement is important in practice: if the practitioner is interested in the effect of a specific effect $j_0 \in \{1,\dots,p\}$, it is important to construct confidence intervals with strict type I error control for β_{j_0}, as opposed to a controlled type I error that only holds on average over all coefficients. Another feature of the results in this paper is that there is no need to assume a prior on the coefficients of $\boldsymbol{\beta}$ in the limit.

Surprisingly, Theorems 3.6, 3.7 and 3.10 and their proofs completely bypass solving the nonlinear equations that appear in the aforementioned works as the nonlinearity is directly treated here with the normal approximation in Theorem 2.2. Asymptotic normality in Theorems 3.6, 3.7 and 3.10 is obtained for a fixed direction $\boldsymbol{a}_0$ (or a fixed coordinate $j \in \{1,\dots,p\}$ when $\boldsymbol{a}_0 = \boldsymbol{e}_j$), and the correlations in $\boldsymbol{\Sigma}$ are handled with a direct approach. Since the first version of this paper was made public, extensions of works cited in the two previous paragraphs were developed [18, 32] to to obtain, for $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$, asymptotic normality results in an averaged sense over $\{1,\dots,p\}$. It is unclear at this point whether these methods can yield asymptotic normality for a fixed coordinate instead of in an averaged sense.

4. Examples. We now present three penalty functions for which closed-form expressions for $\widehat{\boldsymbol{H}}$ and $\boldsymbol{w}_0$ are available. In this section, when computing gradients with respect to $\boldsymbol{z}_0$ in order to find closed-form expressions for $\boldsymbol{w}_0$ in (3.9), we consider $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$ fixed as in (3.7). Explicitly, $\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{z}_0)^\top$ is uniquely defined as

$$\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} = \big[\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{z}_0)\big]^\top \boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|), \tag{4.1}$$

where $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{X}(\boldsymbol{b}-\boldsymbol{\beta}) - \boldsymbol{\varepsilon}\|^2/(2n) + g(\boldsymbol{b})$ and $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \|(\boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}_0^\top)(\boldsymbol{b}-\boldsymbol{\beta}) - \boldsymbol{\varepsilon}\|^2/(2n) + g(\boldsymbol{b})$. When computing gradients with respect to $\boldsymbol{y}$ in order to find closed-form expressions for $\widehat{\boldsymbol{H}}$ in (3.3), we view $\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X}) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}\|^2/(2n) + g(\boldsymbol{b})$ as

a function of $(\boldsymbol{y}, \boldsymbol{X})$ and if $\boldsymbol{y} \mapsto \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is differentiable at $\boldsymbol{y}$ for a fixed $\boldsymbol{X}$ then

$$\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}, \boldsymbol{X}) - \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \big[(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})\big](\widetilde{\boldsymbol{y}} - \boldsymbol{y}) + o(\|\widetilde{\boldsymbol{y}} - \boldsymbol{y}\|), \tag{4.2}$$

where $(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^{p\times n}$ is the Jacobian. Once the Jacobian $(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is computed, $\widehat{\boldsymbol{H}}$ in (3.3) is given by $\widehat{\boldsymbol{H}}^\top = \boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$. We use the Jacobian notation $(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ when computing the derivatives with respect to $\boldsymbol{y}$ to avoid confusion with the gradient $\nabla\boldsymbol{\beta}(z_0)$ in (4.1).

4.1. *Twice continuously differentiable penalty.* The simplest example for which closed-form expressions for $\widehat{\boldsymbol{H}}$, $\widehat{\mathsf{df}}$, $\boldsymbol{w}_0$ can be obtained is that of twice continuously differentiable and strongly convex penalty g. If g is strongly convex, Lemma 7.1 proves that the Fréchet derivative of $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ with respect to $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ exist for almost every $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ by Rademacher's theorem. At a point $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ where the derivative exist, we obtain a closed-form expression for the gradient (3.9) as follows. The KKT conditions of the optimization problem (3.1) read $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^\top(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}) = n\nabla g(\widehat{\boldsymbol{\beta}})$. Differentiation with respect to z_0 for a fixed $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ as in (4.1) gives

$$\{\boldsymbol{X}^\top\boldsymbol{X} + n[\nabla^2 g(\widehat{\boldsymbol{\beta}})]\}(\nabla\widehat{\boldsymbol{\beta}}(z_0))^\top = \boldsymbol{a}_0(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top - \boldsymbol{X}^\top\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle.$$

By the product rule, this provides the derivative of $f(z_0) = \boldsymbol{X}\boldsymbol{h} - \boldsymbol{\varepsilon}$, namely

$$\nabla f(z_0)^\top = \boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X} + n\nabla g(\widehat{\boldsymbol{\beta}})\big)^{-1}\big[\boldsymbol{a}_0(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top - \langle \boldsymbol{a}_0, \boldsymbol{h}\rangle\boldsymbol{X}^\top\big] + \boldsymbol{I}_n\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle. \tag{4.3}$$

Regarding $\widehat{\boldsymbol{H}}$ involving differentiation with respect to $\boldsymbol{y}$, the Lipschitz condition of the map $\boldsymbol{y} \mapsto \widehat{\boldsymbol{\beta}}$ for strongly convex g follows from (7.19) in the proof of Proposition 7.3. Hence, the Jacobian in (4.2) exists almost everywhere, and differentiation of the KKT conditions for fixed $\boldsymbol{X}$ gives $(\boldsymbol{X}^\top\boldsymbol{X} + n\nabla^2 g(\widehat{\boldsymbol{\beta}}))(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \boldsymbol{X}^\top$ so that

$$\widehat{\boldsymbol{H}} = \big(\boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})\big)^\top = \boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X} + n\nabla^2 g(\widehat{\boldsymbol{\beta}})\big)^{-1}\boldsymbol{X}^\top.$$

Identity (4.3) combined with this expression for $\widehat{\boldsymbol{H}}$ provides (3.9) with

$$\boldsymbol{w}_0 = \boldsymbol{X}\{\boldsymbol{X}^\top\boldsymbol{X} + n\nabla g(\widehat{\boldsymbol{\beta}})\}^{-1}\boldsymbol{a}_0.$$

4.2. *Lasso.* Consider the Lasso $\widehat{\boldsymbol{\beta}}$ in (3.32). For $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ with continuous distribution such as Gaussian under consideration here, almost surely $\widehat{\boldsymbol{\beta}}$ is unique and

$$\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = \lambda\,\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{\widehat{S}}), \qquad \|\boldsymbol{X}_{\widehat{S}^c}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\|_\infty < \lambda, \quad \mathrm{rank}(\boldsymbol{X}_{\widehat{S}}) = |\widehat{S}|, \tag{4.4}$$

for the Lasso as in [44, 48] and [7], Proposition 3.9, so that the Jacobian of the mapping $(z_0, \boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0) \to \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ with respect to z_0 and $\boldsymbol{\varepsilon}$ can be computed directly by differentiating the KKT condition as in [7, 8, 44]. The following proposition provides closed-form expressions for the gradients for the Lasso estimator, which are valid almost surely and require no assumption on the sparsity of $\boldsymbol{\beta}$ or the penalty level.

PROPOSITION 4.1. *Let $\lambda > 0$ and consider the Lasso $\widehat{\boldsymbol{\beta}}$ in (3.32). Let $\widehat{S} = \{j \in [p] : \widehat{\beta}_j \neq 0\}$. For almost every $(\overline{\boldsymbol{\varepsilon}}, \overline{\boldsymbol{X}}) \in \mathbb{R}^{n\times(p+1)}$, there exists a neighborhood of $(\overline{\boldsymbol{\varepsilon}}, \overline{\boldsymbol{X}})$ in which the map $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \widehat{S}$ is constant, $|\widehat{S}| \le n$, $\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}}$ is invertible and the map $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ is Lipschitz. In this neighborhood, almost surely $[\nabla\widehat{\boldsymbol{\beta}}(z_0)^\top]_{\widehat{S}^c} = \boldsymbol{0} \in \mathbb{R}^{n\times|\widehat{S}^c|}$,*

$$[\nabla\widehat{\boldsymbol{\beta}}(z_0)^\top]_{\widehat{S}} = (\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}\big((\boldsymbol{a}_0)_{\widehat{S}}(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})^\top - \boldsymbol{X}_{\widehat{S}}^\top\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle\big) \in \mathbb{R}^{n\times|\widehat{S}|},$$

$\widehat{\boldsymbol{H}} = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$, $\widehat{\mathsf{df}} = |\widehat{S}|$ *and* (3.9) *holds with* $\boldsymbol{w}_0 = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}})^{-1}(\boldsymbol{a}_0)_{\widehat{S}}$.

PROOF OF PROPOSITION 4.1. Proposition 3.9 in [7] proves, for almost every $(\boldsymbol{\varepsilon}, \boldsymbol{X})$, the uniqueness of $\widehat{\boldsymbol{\beta}}$ and (4.4). Let $(\overline{\boldsymbol{\varepsilon}}, \overline{\boldsymbol{X}}) \in \mathbb{R}^{n\times(p+1)}$ be a point at which (4.4) holds. Let $\overline{\boldsymbol{y}} = \overline{\boldsymbol{X}}\boldsymbol{\beta} + \overline{\boldsymbol{\varepsilon}}$, $\overline{S} = \operatorname{supp}(\widehat{\boldsymbol{\beta}}(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}))$ and $\boldsymbol{s} = \overline{\boldsymbol{X}}^\top(\overline{\boldsymbol{y}} - \overline{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}))/(n\lambda)$. At $(\boldsymbol{y}, \boldsymbol{X}) = (\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$, the unique solution of (4.4) is given by the analytic expression

$$\widehat{\boldsymbol{\beta}}_{\overline{S}} = (\boldsymbol{X}_{\overline{S}}^\top \boldsymbol{X}_{\overline{S}})^{-1}(\boldsymbol{X}_{\overline{S}}^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_{\overline{S}}), \quad \widehat{\boldsymbol{\beta}}_{\overline{S}^c} = \boldsymbol{0}_{\overline{S}^c}.$$

Moreover, the above expression gives the unique solution of (4.4) in an open neighborhood of $(\overline{\boldsymbol{\varepsilon}}, \overline{\boldsymbol{X}})$ in $\mathbb{R}^{n\times(p+1)}$ in which $\widehat{S} = \overline{S}$, $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\widehat{S}}) = \boldsymbol{s}_{\overline{S}}$ and $\operatorname{rank}(\boldsymbol{X}_{\widehat{S}}) = |\overline{S}|$ are constants. Differentiating this expression immediately yields the formulas for $\widehat{\boldsymbol{H}}$, $\widehat{\mathrm{df}}$ and $[\nabla\widehat{\boldsymbol{\beta}}(z_0)^\top]_{\widehat{S}^c}$. For $[\nabla\widehat{\boldsymbol{\beta}}(z_0)^\top]_{\widehat{S}}$, differentiating both sides of $\boldsymbol{X}_{\overline{S}}^\top(\boldsymbol{X}_{\overline{S}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}) = -n\lambda \boldsymbol{s}_{\overline{S}}$ yields

$$(\boldsymbol{a}_0)_{\overline{S}}(\boldsymbol{X}_{\overline{S}}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})^\top + \boldsymbol{X}_{\overline{S}}^\top \boldsymbol{a}_0^\top \boldsymbol{h} + \boldsymbol{X}_{\overline{S}}^\top \boldsymbol{X}_{\overline{S}}[\nabla\widehat{\boldsymbol{\beta}}(z_0)^\top]_{\overline{S}} = \boldsymbol{0}$$

due to $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + z_0\boldsymbol{a}_0^\top$. Finally, the formula for $\boldsymbol{w}_0$ follows from $(\partial/\partial z_0)(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}) = \boldsymbol{X}(\partial/\partial z_0)\widehat{\boldsymbol{\beta}} + \boldsymbol{I}_n\boldsymbol{a}_0^\top\boldsymbol{h}$ and simple algebra. □

4.3. *Group Lasso.* Consider a partition $(G_1, \dots, G_K)$ of $\{1, \dots, p\}$ and the group Lasso estimator in (3.33). Let $\widehat{B} = \{k \in [K] : \|\widehat{\boldsymbol{\beta}}_{G_k}\| \neq 0\}$ be the set of active groups and $\widehat{S} = \bigcup_{k\in\widehat{B}} G_k$ the union of all active groups. Define the block diagonal matrix $\boldsymbol{M} = \operatorname{diag}((\boldsymbol{M}_{G_k,G_k})_{k\in\widehat{B}}) \in \mathbb{R}^{|\widehat{S}|\times|\widehat{S}|}$ by

$$\boldsymbol{M}_{G_k,G_k} = n\lambda_k\|\widehat{\boldsymbol{\beta}}_{G_k}\|^{-1}(\boldsymbol{I}_{G_k} - \|\widehat{\boldsymbol{\beta}}_{G_k}\|^{-2}\widehat{\boldsymbol{\beta}}_{G_k}\widehat{\boldsymbol{\beta}}_{G_k}^\top), \quad \boldsymbol{M} \in \mathbb{R}^{|\widehat{S}|\times|\widehat{S}|}. \tag{4.5}$$

The following proposition provides closed-form expressions for the gradients for the group Lasso estimator and related quantities $\widehat{\boldsymbol{H}}$ and $\boldsymbol{w}_0$ in terms of $\widehat{S}$ and $\boldsymbol{M}$. Its proof is given in Appendix C. Note that the formula for $\widehat{\boldsymbol{H}}$ was known [45].

PROPOSITION 4.2. *The following holds for for almost every* $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}}) \in \mathbb{R}^{n\times(1+p)}$. *The set* $\overline{B} = \{k \in [K] : \|\overline{\boldsymbol{\beta}}_{G_k}\| > 0\}$ *of active groups is the same for all minimizers* $\overline{\boldsymbol{\beta}}$ *of* (3.33) *at* $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$ *and* $\widehat{B} = \overline{B}$ *for all* $(\boldsymbol{y}, \boldsymbol{X})$ *in a sufficiently small neighborhood of* $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$. *If additionally* $\overline{\boldsymbol{X}}_{\overline{S}}^\top\overline{\boldsymbol{X}}_{\overline{S}}$ *is invertible where* $\overline{S} = \bigcup_{k\in\overline{B}} G_k$ *then the map* $(\boldsymbol{y}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ *is Lipschitz in a sufficiently small neighborhood of* $(\overline{\boldsymbol{y}}, \overline{\boldsymbol{X}})$. *In this neighborhood, we have*

$$[\nabla\widehat{\boldsymbol{\beta}}(z_0)]_{\widehat{S}^c} = 0, \qquad [\nabla\widehat{\boldsymbol{\beta}}(z_0)]_{\widehat{S}}^\top = (\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}[(\boldsymbol{a}_0)_{\widehat{S}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top - \langle \boldsymbol{a}_0, \boldsymbol{h}\rangle \boldsymbol{X}_{\widehat{S}}^\top],$$

$$\widehat{\boldsymbol{H}} = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}\boldsymbol{X}_{\widehat{S}}^\top \text{ and (3.9) holds with } \boldsymbol{w}_0 = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}(\boldsymbol{a}_0)_{\widehat{S}}.$$

5. Simulations: Lasso and variance spike. Figure 1 illustrates the variance spike phenomenon of Section 3.7 for the Lasso and $\boldsymbol{a}_0$ proportional to the first canonical basis vector. The data is generated as follows: $(s, n, p, \sigma^2) = (200, 750, 1000, 1.0)$, coefficient vector $\boldsymbol{\beta}$ is s-sparse with $\beta_1 = 20$, $\beta_j = \pm 1$ for $j = 2, \dots, s$ (independent random signs), $\beta_j = 0$ for $j > s$; inverse covariance matrix $\boldsymbol{\Sigma}^{-1} = \boldsymbol{I}_p + 0.9s^{-1/2}(\boldsymbol{e}_1\operatorname{sgn}(\boldsymbol{\beta})^\top + \operatorname{sgn}(\boldsymbol{\beta})\boldsymbol{e}_1^\top)$, direction $\boldsymbol{a}_0 = \boldsymbol{e}_1/(\boldsymbol{e}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1)^{1/2}$ for $\boldsymbol{e}_1 \in \mathbb{R}^p$ the first canonical basis vector. 512 repetitions were used and $(\boldsymbol{\Sigma}, \boldsymbol{\beta})$ are the same across these repetitions. We see that $V_0 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ yields an empirical standard deviation (std) substantially larger than 1 (second column), whereas using $V_0 = \hat{V}(\theta)$ repairs this with an std close to 1 (third column). This choice of $(\boldsymbol{\Sigma}, \boldsymbol{\beta})$ is a minor modification of [8], Example 2.1 and Figure 2: it is constructed so that $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle^2$ captures a substantial fraction of the full prediction error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$.

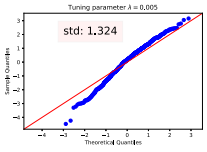
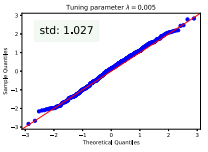
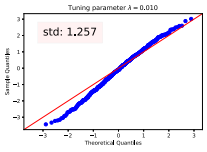
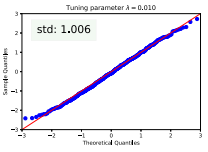
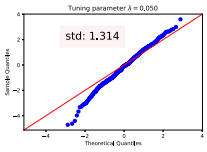
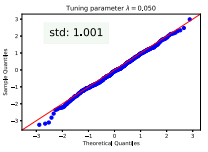
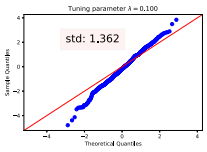
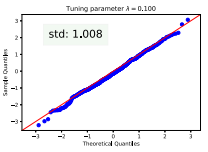
λ	$\frac{(n-\widehat{\mathrm{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de\text{-}bias})}-\boldsymbol{\beta}\rangle}{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|}$	$\frac{(n-\widehat{\mathrm{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de\text{-}bias})}-\boldsymbol{\beta}\rangle}{\hat{V}(\theta)^{1/2}}$	$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2$	$\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle^2$
0.005			3.23 ±0.45	0.32 ±0.11
0.01			2.62 ±0.41	0.40 ±0.12
0.05			4.39 ±0.86	1.81 ±0.39
0.1			11.70 ±2.40	5.6 ±1.14

FIG. 1. *Standard normal QQ-plots of* $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle / V_0^{1/2}$ *with* $V_0 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$ *(second column) and with* $V_0 = \hat{V}(\theta) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + (n - \widehat{\mathrm{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle^2$ *(third column), prediction error* $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$ *(fourth column) and squared estimation error in direction* $\boldsymbol{a}_0$ *(fifth column) for the Lasso (3.32) for each* $\lambda \in \{0.005, 0.01, 0.05, 0.1\}$ *with the data-generating process described in Section 5.*

6. Simulations: Group Lasso. Figure 2 illustrates Theorem 3.10 for the group Lasso (3.33) with standard normal QQ-plots of $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle (n - \widehat{\mathrm{df}})/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|$ for $(n, p, \sigma^2) = (600, 900, 2)$ and the group Lasso (3.33) with 30 nonoverlapping groups each of size 30, where all λ_k in (3.33) are equal to a single parameter λ. The unknown coefficient vector $\boldsymbol{\beta}$ is the same across all 256 repetitions and has 240 nonzero coefficients, all equal to 1 and belonging to 8 groups (so that the group sparsiy of $\boldsymbol{\beta}$ is 8, and within these 8 groups all coefficients are equal to 1). The design covariance $\boldsymbol{\Sigma}$ is generated once as $\boldsymbol{\Sigma} = \boldsymbol{W}/(5p)$ where $\boldsymbol{W}$ has Wishart distribution with covariance $\boldsymbol{I}_p$ and $5p$ degrees-of-freedom. This choice of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is the same across all 256 repetitions. The direction of interest is $\boldsymbol{a}_0 = \boldsymbol{e}_1/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_1\|$ where $\boldsymbol{e}_1 \in \mathbb{R}^p$ is the first canonical basis vector. The first 8 plots above are standard normal QQ-plots across 256 repetitions for 8 different choices of λ. The ninth plot shows, for each λ, boxplots of $\hat{\tau}^2 = (1 - \widehat{\mathrm{df}}/n)^{-2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n$ across the 256 repetitions. This $\hat{\tau}$ is proportional to the length of the corresponding confidence interval (3.28) so that the smallest confidence interval (3.28) is achieved for $\lambda = 0.138$.

7. Proof of the main results in Section 3. In order to prove Theorems 3.6 and 3.7, we apply the bound on the normal approximation in Theorem 2.2. We recall here some notation used throughout the proof. Let $\widehat{\boldsymbol{\beta}}$ be the estimator (3.1), $\widehat{\boldsymbol{H}}$ the gradient of $\boldsymbol{y} \to \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ as in (1.8), $\boldsymbol{a}_0 \in \mathbb{R}^p$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$, z_0 and $\boldsymbol{Q}_0$ as defined in (1.13),

$$\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle, \qquad f(z_0) = \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y}, \qquad \xi_0 = z_0^\top f(z_0) - \operatorname{div} f(z_0). \tag{7.1}$$

Vector $\boldsymbol{w}_0 \in \mathbb{R}^n$ is given by Lemma 3.1. The oracle $\boldsymbol{\beta}^*$ and its associated noiseless prediction risk R_* are given by (3.23). Throughout, $\mathbb{E}_0$ denotes the conditional expectation given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$ and Var_0 the conditional variance given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$.

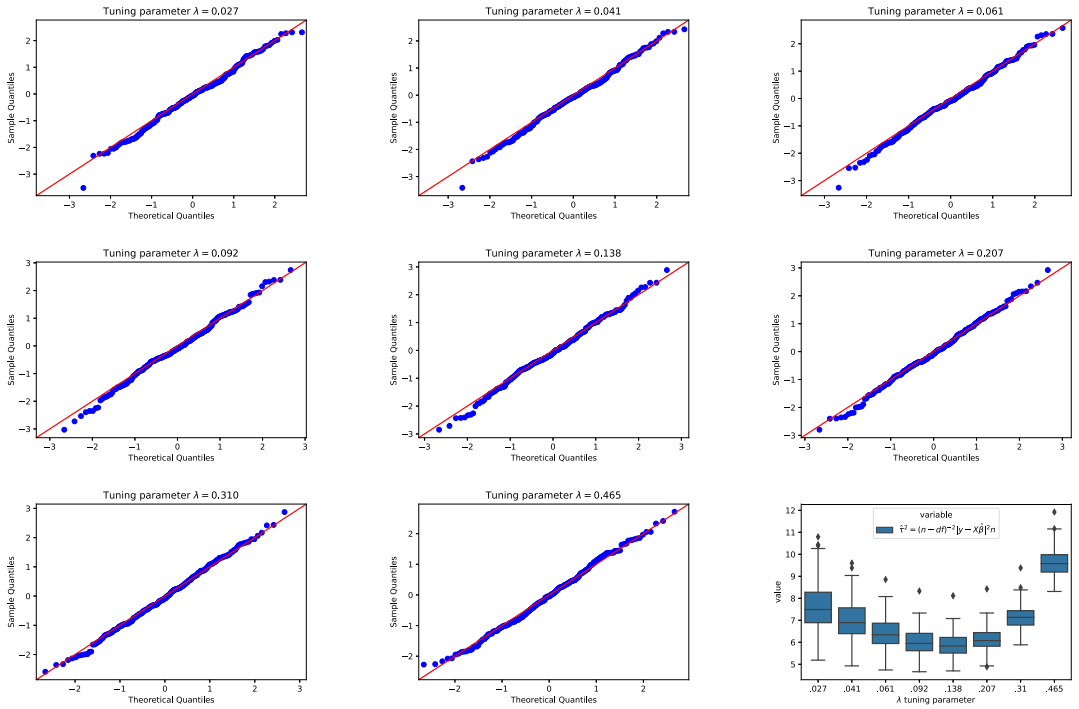


FIG. 2. *Standard normal QQ-plots of* $\langle a_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle (n - \widehat{\mathrm{df}})/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|$ *for for the group Lasso* (3.33). *The data-generating process is described in Section* 6.

7.1. *Lipschitzness of regularized least-squares and existence of* $\boldsymbol{w}_0$. By Rademacher's theorem, a Lipschitz function $U \to \mathbb{R}$ for some open set $U \subset \mathbb{R}^q$ is Fréchet differentiable almost everywhere in U. The following lemma is the device that verifies the Lipschitz condition for the mappings $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ and $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \boldsymbol{X}\boldsymbol{h} - \boldsymbol{\varepsilon}$ in certain open set U, and consequently, differentiability almost everywhere in U.

LEMMA 7.1. *Let* $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{X}$ *and* $\widetilde{\boldsymbol{X}}$ *be two design matrices of size* $n \times p$, *and* $\boldsymbol{\varepsilon}$ *and* $\widetilde{\boldsymbol{\varepsilon}}$ *two noise vectors in* $\mathbb{R}^n$. *Let* $g : \mathbb{R}^p \to \mathbb{R}$ *be convex such that minimizers*

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{\|\boldsymbol{\varepsilon} + \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{b})\|^2}{2n} + g(\boldsymbol{b}) \right\}, \qquad \widetilde{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{\|\widetilde{\boldsymbol{\varepsilon}} + \widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{b})\|^2}{2n} + g(\boldsymbol{b}) \right\}$$

exist. Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{h} - \boldsymbol{\varepsilon}$, $\widetilde{\boldsymbol{h}} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}} - \widetilde{\boldsymbol{\varepsilon}}$. *Let also* $D_g(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) = (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})^\top \{(\partial g)(\widetilde{\boldsymbol{\beta}}) - (\partial g)(\widehat{\boldsymbol{\beta}})\}$ *where* $(\partial g)(\widetilde{\boldsymbol{\beta}}) = n^{-1}\widetilde{\boldsymbol{X}}^\top(\widetilde{\boldsymbol{\varepsilon}} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}})$ *is the subdifferential at* $\widetilde{\boldsymbol{\beta}}$ *given by the optimality condition of the above optimization problem and similarly for* $(\partial g)(\widehat{\boldsymbol{\beta}})$, *with* $D_g(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}) \geq 0$ *by the monotonicity of the subdifferential. Then*

$$\begin{aligned} & n D_g(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) + \|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\|^2 \\ &= (\widetilde{\boldsymbol{h}} - \boldsymbol{h})^\top (\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top \boldsymbol{f} + \big(\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon} + (\boldsymbol{X} - \widetilde{\boldsymbol{X}})\boldsymbol{h}\big)^\top (\boldsymbol{f} - \widetilde{\boldsymbol{f}}) \\ &= \operatorname{trace}\big[(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top \big(\boldsymbol{f}\widetilde{\boldsymbol{h}}^\top - \widetilde{\boldsymbol{f}}\boldsymbol{h}^\top\big)\big] + (\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon})^\top (\boldsymbol{f} - \widetilde{\boldsymbol{f}}). \end{aligned} \tag{7.2}$$

If g *is coercive* (*i.e.*, $g(\boldsymbol{x}) \to +\infty$ *as* $\|\boldsymbol{x}\| \to +\infty$) *then the map* $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}$ *is Holder continuous with coefficient* $1/2$ *on every compact. We also have*

$$\begin{aligned} & n D_g(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) + \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})\|^2/2 + \|\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})\|^2/2 \\ &= (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^\top \big(\boldsymbol{X}^\top \boldsymbol{\varepsilon} - \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{\varepsilon}}\big) + (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^\top \big(\boldsymbol{X}^\top \boldsymbol{X} - \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}\big)(\boldsymbol{h} + \widetilde{\boldsymbol{h}})/2 \end{aligned} \tag{7.3}$$

$$\le \|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|\|\boldsymbol{\varepsilon}-\widetilde{\boldsymbol{\varepsilon}}\|\|\boldsymbol{X}+\widetilde{\boldsymbol{X}}\|_{\mathrm{op}}/2 \tag{7.4}$$
$$+\|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|\|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\|_{\mathrm{op}}\big[\big(\|\boldsymbol{\varepsilon}+\widetilde{\boldsymbol{\varepsilon}}\|/2+\big(\|\boldsymbol{X}\|_{\mathrm{op}}+\|\widetilde{\boldsymbol{X}}\|_{\mathrm{op}}\big)\big)\big(\|\widetilde{\boldsymbol{h}}\|+\|\boldsymbol{h}\|\big)/2\big].$$

If either g is strongly convex or if there exists a constant $\overline{\kappa}>0$ and a bounded neighborhood $\mathcal{N}$ of $(\overline{\boldsymbol{\varepsilon}},\overline{\boldsymbol{X}})$ such that $\|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|\overline{\kappa}\le\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\|$ for all $\{(\boldsymbol{\varepsilon},\boldsymbol{X}),(\widetilde{\boldsymbol{\varepsilon}},\widetilde{\boldsymbol{X}})\}\subset\mathcal{N}$ then the map $(\boldsymbol{\varepsilon},\boldsymbol{X})\mapsto\widehat{\boldsymbol{\beta}}$ is Lipschitz in $\mathcal{N}$.

PROOF OF LEMMA 7.1. The KKT conditions for $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ provide

$$n(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^\top(\partial g)(\widehat{\boldsymbol{\beta}})=(\widetilde{\boldsymbol{h}}-\boldsymbol{h})^\top\boldsymbol{X}^\top\boldsymbol{f},$$
$$n(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^\top(\partial g)(\widetilde{\boldsymbol{\beta}})=(\boldsymbol{h}-\widetilde{\boldsymbol{h}})^\top\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{f}}.$$

Summing and adding $\|\boldsymbol{f}-\widetilde{\boldsymbol{f}}\|^2=(\widetilde{\boldsymbol{\varepsilon}}-\boldsymbol{\varepsilon})^\top(\boldsymbol{f}-\widetilde{\boldsymbol{f}})+(\boldsymbol{X}\boldsymbol{h}-\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}})^\top(\boldsymbol{f}-\widetilde{\boldsymbol{f}})$ on both sides,

$$nD_g(\widehat{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}})+\|\boldsymbol{f}-\widetilde{\boldsymbol{f}}\|^2=\widetilde{\boldsymbol{h}}^\top(\boldsymbol{X}-\widetilde{\boldsymbol{X}})^\top\boldsymbol{f}+\boldsymbol{h}^\top(\widetilde{\boldsymbol{X}}-\boldsymbol{X})^\top\widetilde{\boldsymbol{f}}+(\widetilde{\boldsymbol{\varepsilon}}-\boldsymbol{\varepsilon})^\top(\boldsymbol{f}-\widetilde{\boldsymbol{f}})$$

so that (7.2) holds.

By optimality of $\widehat{\boldsymbol{\beta}}$, $\|\boldsymbol{f}\|^2/(2n)+g(\widehat{\boldsymbol{\beta}})\le\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\|^2/(2n)$. If g is coercive, this implies that for every compact $K\subset\mathbb{R}^{n\times(1+p)}$, $\|\boldsymbol{f}\|+\|\boldsymbol{h}\|$ and $\|\widetilde{\boldsymbol{f}}\|+\|\widetilde{\boldsymbol{h}}\|$ are bounded by a constant depending only on g, $\boldsymbol{\beta}$, n, K if $\{(\boldsymbol{\varepsilon},\boldsymbol{X}),(\widetilde{\boldsymbol{\varepsilon}},\widetilde{\boldsymbol{X}})\}\subset K$. In this case, (7.2) implies that $\|\widetilde{\boldsymbol{f}}-\boldsymbol{f}\|^2\le(\|\widetilde{\boldsymbol{X}}-\boldsymbol{X}\|_F+\|\boldsymbol{\varepsilon}-\widetilde{\boldsymbol{\varepsilon}}\|)C(g,\boldsymbol{\beta},n,K)$ for some other constant depending on g, $\boldsymbol{\beta}$, n, K only. This implies Holder continuity of $(\boldsymbol{\varepsilon},\boldsymbol{X})\mapsto\boldsymbol{\varepsilon}-\boldsymbol{X}\boldsymbol{h}$ with Holder coefficient $1/2$ on every compact.

For (7.3) and (7.4), the KKT conditions for $\widehat{\boldsymbol{\beta}}$ yield

$$n(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^\top(\partial g)(\widehat{\boldsymbol{\beta}})+\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\|^2/2=(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\big(\boldsymbol{\varepsilon}-\boldsymbol{X}(\boldsymbol{h}+\widetilde{\boldsymbol{h}})/2\big).$$

Summing the above and its $\widetilde{\boldsymbol{\beta}}$ counterpart yields the equality (7.3). Writing $\boldsymbol{X}^\top\boldsymbol{\varepsilon}-\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\varepsilon}}=(\widetilde{\boldsymbol{X}}+\boldsymbol{X})^\top(\boldsymbol{\varepsilon}-\widetilde{\boldsymbol{\varepsilon}})/2+(\boldsymbol{X}-\widetilde{\boldsymbol{X}})^\top(\widetilde{\boldsymbol{\varepsilon}}+\boldsymbol{\varepsilon})/2$ and similarly $\boldsymbol{X}^\top\boldsymbol{X}-\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}=(\boldsymbol{X}+\widetilde{\boldsymbol{X}})^\top(\boldsymbol{X}-\widetilde{\boldsymbol{X}})/2+(\boldsymbol{X}-\widetilde{\boldsymbol{X}})^\top(\boldsymbol{X}+\widetilde{\boldsymbol{X}})/2$, inequality (7.4) follows. To prove the Lipschitz condition in $\mathcal{N}$, we note that for a fixed value of $(\boldsymbol{\varepsilon},\boldsymbol{X},\boldsymbol{h})$, the right-hand side of (7.4) is linear in $\|\widetilde{\boldsymbol{h}}\|$ while the left-hand side is quadratic in $\|\widetilde{\boldsymbol{h}}\|$ thanks to either strong convexity of g or the assumption on $\overline{\kappa}$. This implies that $\|\widetilde{\boldsymbol{h}}\|$ is bounded uniformly for all $(\widetilde{\boldsymbol{\varepsilon}},\widetilde{\boldsymbol{X}})$ in $\mathcal{N}$. Since $\boldsymbol{\varepsilon}$, $\boldsymbol{X}$, $\widetilde{\boldsymbol{\varepsilon}}$, $\widetilde{\boldsymbol{X}}$, $\|\widetilde{\boldsymbol{h}}\|$, $\|\boldsymbol{h}\|$ are all bounded in $\mathcal{N}$, (7.4) divided by $\|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|$ provides the desired Lipschitz property. □

LEMMA 3.1. *Let Assumption 3.1(i) be fulfilled, $\boldsymbol{a}_0\in\mathbb{R}^p$ and $\widehat{\boldsymbol{H}}$ be as in (3.3). Then*

$$\nabla f(z_0)^\top=(\boldsymbol{I}_n-\widehat{\boldsymbol{H}})^\top\langle\boldsymbol{a}_0,\boldsymbol{h}\rangle+\boldsymbol{w}_0(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \tag{3.9}$$

satisfies (3.6) for some random $\boldsymbol{w}_0\in\mathbb{R}^n$ almost surely. If additionally $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\|=1$, then

$$\|\boldsymbol{w}_0\|^2\le n^{-1}\min\big\{(4\mu)^{-1},\phi_{\min}\big(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}/n\big)^{-1}\big\}. \tag{3.10}$$

PROOF OF LEMMA 3.1. Under Assumption 3.1, Lemma 7.1 implies that the map $(\boldsymbol{\varepsilon},\boldsymbol{X})\mapsto\boldsymbol{f}=\boldsymbol{X}\boldsymbol{h}-\boldsymbol{\varepsilon}$ is Lipschitz in an open neighborhood of almost every point, and thus $\widehat{\boldsymbol{H}}$ and $\nabla f(z_0)$ are defined as Fréchet derivatives almost surely in (3.3) and (3.6), respectively. To prove (3.9), that is, that the range of $\nabla f(z_0)-\langle\boldsymbol{a}_0,\boldsymbol{h}\rangle(\boldsymbol{I}_n-\widehat{\boldsymbol{H}})$ is the linear span of $\boldsymbol{f}$, we study the directional derivative in a direction $\boldsymbol{\eta}\in\mathbb{R}^n$. For two pairs $(\boldsymbol{\varepsilon},\boldsymbol{X})$ and $(\widetilde{\boldsymbol{\varepsilon}},\widetilde{\boldsymbol{X}})$ with $\widetilde{\boldsymbol{X}}=\boldsymbol{X}+t\boldsymbol{\eta}\boldsymbol{a}_0^\top=\boldsymbol{X}\boldsymbol{Q}_0+(t\boldsymbol{\eta}+z_0)\boldsymbol{a}_0^\top$ and $\widetilde{\boldsymbol{\varepsilon}}=\boldsymbol{\varepsilon}+t\boldsymbol{\eta}\langle\boldsymbol{a}_0,\boldsymbol{h}\rangle$, consider the solutions $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ defined in Lemma 7.1 and $\boldsymbol{\phi}_t=\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\widetilde{\boldsymbol{\varepsilon}}$ with $\boldsymbol{\phi}_0=\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})-\boldsymbol{\varepsilon}=\boldsymbol{f}$. When the map $(\boldsymbol{\varepsilon},\boldsymbol{X})\mapsto\boldsymbol{f}$ is Fréchet differentiable at $(\boldsymbol{\varepsilon},\boldsymbol{X})$,

$$\lim_{t\to0+}(\boldsymbol{\phi}_t-\boldsymbol{\phi}_0)/t=\big(\nabla f(z_0)-\langle\boldsymbol{a}_0,\boldsymbol{h}\rangle(\boldsymbol{I}_n-\widehat{\boldsymbol{H}})\big)^\top\boldsymbol{\eta} \tag{7.5}$$

by the chain rule and the linearity of the Fréchet derivative, noting that $(\partial/\partial\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}) = \boldsymbol{I}_n - \widehat{\boldsymbol{H}}$. For this specific choice of $(\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}})$, we have

$$(\widetilde{\boldsymbol{X}} - \boldsymbol{X})\boldsymbol{h} + \boldsymbol{\varepsilon} - \widetilde{\boldsymbol{\varepsilon}} = 0. \tag{7.6}$$

It follows that the second term in the first line of (7.2) is zero, so that (7.2) gives

$$\mu n \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})\|^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}_t\|^2 \le |\langle \boldsymbol{a}_0, \boldsymbol{h} - \widetilde{\boldsymbol{h}}\rangle t\boldsymbol{\eta}^\top \boldsymbol{f}|$$

due to $\widetilde{\boldsymbol{X}} - \boldsymbol{X} = t\boldsymbol{\eta}\boldsymbol{a}_0^\top$. Consequently, $\phi_t - \phi_0 = 0$ when $\boldsymbol{\eta}^\top \boldsymbol{f} = 0$. This and (7.5) give (3.9). Moreover, for $\boldsymbol{f} \neq \boldsymbol{0}$, $\boldsymbol{w}_0 = \lim_{t\to 0}(\boldsymbol{\phi}_t - \boldsymbol{\phi}_0)/t$ for $\boldsymbol{\eta} = -\boldsymbol{f}/\|\boldsymbol{f}\|^2$, so that (3.10) is an upper bound for $\lim_{t\to 0+} \|\boldsymbol{\phi}_t - \boldsymbol{\phi}_0\|/t$ in the case of $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1 = -\boldsymbol{\eta}^\top \boldsymbol{f}$ where

$$\mu n \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})\|^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}_t\|^2 \le |t| \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|$$

by the previous display. For $\mu > 0$, the above inequality gives $\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}_t\|^2 \le t^2(4\mu n)^{-1}$ using $uv \le u^2/4 + v^2$. For $\mu = 0$,

$$\phi_{\min}(\widetilde{\boldsymbol{W}}) \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2 \le \|\widetilde{\boldsymbol{X}}^\top(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2 = \|\boldsymbol{\phi}_t - \boldsymbol{\phi}_0\|^2$$

with $\phi_{\min}(\widetilde{\boldsymbol{W}})$ being the smallest eigenvalue of $\widetilde{\boldsymbol{W}} = \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}\boldsymbol{\Sigma}^{-1/2}$. Hence, (3.10) holds in either cases. □

7.2. *Loss equivalence to oracle estimators*. To apply Theorem 2.2 with respect to z_0 to f in (7.1), we will need to control expectations involving $\|\boldsymbol{w}_0\|$, $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle$, $\|\boldsymbol{X}\boldsymbol{h}\|$ and $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|$. To this end, define the random variables F_+ and F by

$$F_+ \stackrel{\text{def}}{=} (\|\boldsymbol{g}\|^2/n) \vee (\|\boldsymbol{\varepsilon}\|^2/(\sigma^2 n)) \vee (\|\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}^*\|^2/(nR_*)) \vee 1 \tag{7.7}$$

with $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{h}^*/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$ and the $\boldsymbol{h}^*$ and R_* in (3.23), and

$$F \stackrel{\text{def}}{=} 2/[1 \wedge \max\{\mu, \phi_{\min}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}^\top \boldsymbol{X}/n)\boldsymbol{\Sigma}^{-1/2})\}]. \tag{7.8}$$

We note that the three random vectors $\boldsymbol{\varepsilon}/\sigma$, $\boldsymbol{g}$ and $(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}^*)/{R_*}^{1/2}$ have $N(\boldsymbol{0}, \boldsymbol{I}_n)$ distribution, so that F_+ is of the form $F_+ = \max_{i=1,2,3} W_i/n$ where each W_i has the χ^2_n distribution. Thus, by Proposition A.1 and properties of the χ^2_n distribution,

$$\mathbb{E}[F_+^4] \vee \mathbb{E}[F^{10}] \le C(\gamma, \mu), \qquad \mathbb{E}[(F_+ - 1)^2] \le 3\,\mathrm{Var}[\chi^2_n]/n^2 = 6/n. \tag{7.9}$$

It follows from (1.15), Lemma 7.2 below and (3.10) that almost surely

$$\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 \le \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 \vee (\|\boldsymbol{X}\boldsymbol{h}\|^2/n) \le F_+ F^2 R_*, \qquad \|\boldsymbol{w}_0\|^2 \le F/(2n), \tag{7.10}$$

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n \le 2F_+ + 2F_+F^2R_* \le 4F_+F^2R_*, \tag{7.11}$$

for the $\boldsymbol{w}_0$ in Lemma 3.1. The moment inequalities in (7.9) and the almost sure bounds (7.10)–(7.11) allow us to control expectations involving $\|\boldsymbol{w}_0\|$, $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle$, $\|\boldsymbol{X}\boldsymbol{h}\|$ and $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|$ throughout the proofs. The following lemma provides the first inequality in (7.10).

LEMMA 7.2 (Deterministic lemma). *Consider the linear model* (1.1) *and a convex penalty* $g(\cdot)$. *Let* $\widehat{\boldsymbol{\beta}}$ *in* (3.1) *and* $\boldsymbol{\beta}^*$, $\boldsymbol{h}^*$, R_* *be defined in* (3.23). *Suppose the penalty satisfies* $\boldsymbol{u}^\top\{(\partial g)(\boldsymbol{u} + \boldsymbol{\beta}^*) - (\partial g)(\boldsymbol{\beta}^*)\} \ge \mu\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}\|^2 \forall \boldsymbol{u} \in \mathbb{R}^p$ *with* $\mu \in [0, 1/2]$. *Let* F_+ *be defined in* (7.7) *and* F *any random variable satisfying*

$$1 \le F/2 \quad \textit{and either} \quad \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2/(n\|\boldsymbol{X}\boldsymbol{h}\|^2) \le F/2 \quad \textit{or} \quad \mu^{-1} = F/2, \tag{7.12}$$

for instance (7.8). *Then*

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 \le F^2 \max(\overline{\sigma}^2, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2) \le F_+F^2R_*, \tag{7.13}$$

$$\|\boldsymbol{X}\boldsymbol{h}\|^2/n \le \max\{2F\overline{\sigma}^2, \overline{\sigma}^2 + F^2\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2\} \le F_+F^2R_*, \tag{7.14}$$

where $\overline{\sigma}^2 = F_+\sigma^2 + (F_+ - 1)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2 = (F_+ - 1)R_* + \sigma^2$.

PROOF OF LEMMA 7.2. The KKT conditions for $\widehat{\boldsymbol{\beta}}$, that is, $n\partial g(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, yield

$$\begin{aligned}
2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top(\partial g)(\widehat{\boldsymbol{\beta}}) &= 2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n \\
&= \big(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 - \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2\big)/n \\
&= \big(\|\boldsymbol{X}\boldsymbol{h}^*\|^2 - \|\boldsymbol{X}\boldsymbol{h}\|^2 - \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 + 2\boldsymbol{\varepsilon}^\top \boldsymbol{X}(\boldsymbol{h} - \boldsymbol{h}^*)\big)/n \\
&\le \big(\|\boldsymbol{X}\boldsymbol{h}^*\|^2 - \|\boldsymbol{X}\boldsymbol{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2\big)/n.
\end{aligned}$$

Similarly, the KKT conditions $-\boldsymbol{\Sigma}\boldsymbol{h}^* = \partial g(\boldsymbol{\beta}^*)$ for $\boldsymbol{\beta}^*$ yield

$$2(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^\top(\partial g)(\boldsymbol{\beta}^*) + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \le \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 - \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2. \tag{7.15}$$

Summing the two above displays yields

$$\begin{aligned}
&(1 + 2\mu)\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \\
&\quad\le (F_+ - 1)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 - \|\boldsymbol{X}\boldsymbol{h}\|^2/n + F_+\sigma^2 \qquad (7.16)\\
&\quad= \overline{\sigma}^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 - \|\boldsymbol{X}\boldsymbol{h}\|^2/n.
\end{aligned}$$

For $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\| \ge F\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$, by the triangle inequality

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2(1 - 1/F)^2 \le \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \tag{7.17}$$

provides a lower bound on the left-hand side of (7.16) so that

$$\begin{aligned}
\overline{\sigma}^2 &\ge \begin{cases} \{(1 - 1/F)^2 + 2/F - 1)\}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 & \text{if } \|\boldsymbol{X}\boldsymbol{h}\|^2/n \ge (2/F)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2, \\ \{(1 - 1/F)^2(1 + 2\mu) - 1\}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 & \text{if } \|\boldsymbol{X}\boldsymbol{h}\|^2/n < (2/F)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2, \end{cases} \qquad (7.18)\\
&\ge F^{-2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2
\end{aligned}$$

due to $F = 2/\mu \ge 4$ in the second case. This gives (7.13). For $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\| \ge F\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$ by (7.16), (7.17) and (7.18) we have

$$\|\boldsymbol{X}\boldsymbol{h}\|^2/n \le \overline{\sigma}^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2\{1 - (1 - 1/F)^2\} \le \overline{\sigma}^2 + F^2\overline{\sigma}^2\{1 - (1 - 1/F)^2\} = 2F\overline{\sigma}^2,$$

and for $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\| < F\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$ we have $\|\boldsymbol{X}\boldsymbol{h}\|^2/n \le \overline{\sigma}^2 + F^2\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2$, and thus (7.14) holds. □

7.3. *Existence and properties of* $\widehat{\boldsymbol{H}}$ *and* $\widehat{\mathsf{df}}$.

PROPOSITION 7.3. *Let* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *be any fixed design matrix, and* $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + g(\boldsymbol{b})\}$. *Then the following statements hold*:

(i) $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}))\| \le \|\boldsymbol{y} - \widetilde{\boldsymbol{y}}\|$ *for all* $\boldsymbol{y}, \widetilde{\boldsymbol{y}} \in \mathbb{R}^n$, *that is,* $\boldsymbol{y} \mapsto \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y})$ *is* 1*-Lipschitz. Its gradient* $\widehat{\boldsymbol{H}}$ *exists almost everywhere by Rademacher's theorem, that is, for almost every* $\boldsymbol{y}$ *there exists* $\widehat{\boldsymbol{H}} \in \mathbb{R}^{n\times n}$ *with* $\|\widehat{\boldsymbol{H}}\|_{\mathrm{op}} \le 1$ *such that* $\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) + \widehat{\boldsymbol{H}}^\top\boldsymbol{\eta} + o(\|\boldsymbol{\eta}\|)$.

(ii) *For almost every* $\boldsymbol{y}$, *matrix* $\widehat{\boldsymbol{H}}$ *is symmetric with eigenvalues in* $[0, 1]$. *Consequently, with* $\widehat{\mathsf{df}} = \mathrm{trace}(\widehat{\boldsymbol{H}})$ *as degrees-of-freedom,* $(n - \widehat{\mathsf{df}})(1 - \widehat{\mathsf{df}}/n) \le \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n - \widehat{\mathsf{df}}$.

PROOF. A proof of (i) is given in [2]. For completeness, the argument is the following: by (7.3) with $\widetilde{\boldsymbol{X}} = \boldsymbol{X}$, $\widetilde{\boldsymbol{\varepsilon}} = \widetilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ we have

$$nD_g(\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}), \widehat{\boldsymbol{\beta}}(\boldsymbol{y})) + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}))\|^2 \le (\boldsymbol{y} - \widetilde{\boldsymbol{y}})^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})). \tag{7.19}$$

Using $D_g(\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}), \widehat{\boldsymbol{\beta}}(\boldsymbol{y})) \geq 0$ by monotonicity of the subdifferential and the Cauchy–Schwarz inequality yields the desired Lipschitz property. For (ii), define

$$\begin{aligned} u(\boldsymbol{y}) &= \big(\|\boldsymbol{y}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y})\|^2\big)/2 - ng\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{y})\big) \\ &= \sup_{\boldsymbol{b}\in\mathbb{R}^p} \big\{\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{b} - \|\boldsymbol{X}\boldsymbol{b}\|^2/2 - ng(\boldsymbol{b})\big\}. \end{aligned}$$

The function $u : \mathbb{R}^n \to \mathbb{R}$ is convex in $\boldsymbol{y}$ as a supremum of affine functions, and $\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y})$ is a subgradient of u at $\boldsymbol{y}$. Alexandrov's theorem as stated in [35], Theorem D.2.1, states that any convex $u : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at $\boldsymbol{y}$ for almost every $\boldsymbol{y}$ in the following sense: u is Fréchet differentiable at $\boldsymbol{y}$ with gradient $\nabla u(\boldsymbol{y})$ and there exists a symmetric positive semidefinite matrix $\boldsymbol{S}$ such that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\widetilde{\boldsymbol{y}} \in \mathbb{R}^n$,

$$\|\widetilde{\boldsymbol{y}} - \boldsymbol{y}\| \leq \delta \quad \text{implies} \quad \sup_{\boldsymbol{v}\in\partial u(\widetilde{\boldsymbol{y}})} \|\boldsymbol{v} - \nabla u(\boldsymbol{y}) - \boldsymbol{S}(\widetilde{\boldsymbol{y}} - \boldsymbol{y})\| \leq \varepsilon\|\widetilde{\boldsymbol{y}} - \boldsymbol{y}\|.$$

By (i) and the definition of $\widehat{\boldsymbol{H}}$, for almost every $\boldsymbol{y}$ it holds that $\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) + \widehat{\boldsymbol{H}}^\top(\widetilde{\boldsymbol{y}} - \boldsymbol{y}) + o(\|\widetilde{\boldsymbol{y}} - \boldsymbol{y}\|)$. Combining these two results and taking $\boldsymbol{v} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})$, we get that $\boldsymbol{S} = \widehat{\boldsymbol{H}}$ for almost every $\boldsymbol{y}$. □

LEMMA 7.4. *Let Assumption 3.1 be fulfilled with $n \geq 2$. Then there exists an event Ω_0 independent of $(\boldsymbol{z}_0, \boldsymbol{\varepsilon})$ such that*

$$\Omega_0 \subset \big\{n - \widehat{\mathsf{df}} \geq \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \geq C_*(\gamma, \mu)n\big\} \quad \text{with} \quad \begin{cases} \mathbb{P}(\Omega_0^c) = 0 & \text{if } \gamma < 1, \\ \mathbb{P}(\Omega_0^c) \leq e^{-n/2} & \text{if } \gamma \geq 1, \end{cases}$$

where $C_*(\gamma, \mu) \in (0, 1)$ depends on $\{\gamma, \mu\}$ only.

PROOF OF LEMMA 7.4. If $\gamma < 1$, the choice $C_*(\gamma, \mu) = (1 - \gamma)$ works with probability one because $\operatorname{rank}(\widehat{\boldsymbol{H}}) \leq \operatorname{rank}(\boldsymbol{X}) \leq p \leq \gamma n$ and $\|\widehat{\boldsymbol{H}}\|_{\mathrm{op}} \leq 1$.

If $\gamma \geq 1$, then we have $\mu > 0$ in Assumption 3.1. Let $\Omega_0 = \{\|\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}} \leq \sqrt{p} + 2\sqrt{n}\}$. By [20], Theorem II.13, $\mathbb{P}(\Omega_0) \geq 1 - e^{-n/2}$ due to $\boldsymbol{X}\boldsymbol{Q}_0\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I}_p - (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0)(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0)^\top$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$. Next, we hold $\boldsymbol{X}$ fixed and study the derivatives of $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{y}$. Let $\widehat{\boldsymbol{\beta}}(\boldsymbol{y})$ be as in Proposition 7.3. Let $\boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{z}_0\boldsymbol{z}_0^\top/\|\boldsymbol{z}_0\|^2$ be the projection onto $\{\boldsymbol{z}_0\}^\perp$ so that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{Q}_0$. Let $\boldsymbol{y}, \widetilde{\boldsymbol{y}}$ be such that $\boldsymbol{z}_0^\top(\boldsymbol{y} - \widetilde{\boldsymbol{y}}) = 0$, or equivalently $\boldsymbol{P}(\boldsymbol{y} - \widetilde{\boldsymbol{y}}) = \boldsymbol{y} - \widetilde{\boldsymbol{y}}$. By (7.19) and (3.2),

$$\begin{aligned} n\mu\|\boldsymbol{\Sigma}^{1/2}\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})\big)\|^2 + \|\boldsymbol{X}\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})\big)\|^2 &\leq (\boldsymbol{y} - \widetilde{\boldsymbol{y}})^\top \boldsymbol{X}\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})\big) \\ &= (\boldsymbol{y} - \widetilde{\boldsymbol{y}})^\top \boldsymbol{P}\boldsymbol{X}\big(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}})\big). \end{aligned}$$

On Ω_0, $\mu(\sqrt{\gamma} + 2)^{-2}\|\boldsymbol{X}\boldsymbol{Q}_0(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}))\|^2 \leq n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}))\|^2$. Combined with the above display, this implies $(1 + \mu(\sqrt{\gamma} + 2)^{-2})\|\boldsymbol{P}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) - \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{y}}))\| \leq \|\boldsymbol{P}(\boldsymbol{y} - \widetilde{\boldsymbol{y}})\|$. With $\widetilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{P}\boldsymbol{\eta}$ and by definition of $\widehat{\boldsymbol{H}}$, we have $L^{-1}\|\boldsymbol{P}\widehat{\boldsymbol{H}}\boldsymbol{P}\boldsymbol{\eta}\| \leq \|\boldsymbol{P}\boldsymbol{\eta}\| + o(\|\boldsymbol{\eta}\|)$ for $L = (1 + \mu(\sqrt{\gamma} + 2)^{-2})^{-1}$, hence $\|\boldsymbol{P}\widehat{\boldsymbol{H}}\boldsymbol{P}\|_{\mathrm{op}} \leq L$. Since $\operatorname{rank}(\boldsymbol{P}) = n - 1$, by Cauchy's interlacing theorem $\boldsymbol{I}_n - \widehat{\boldsymbol{H}}$ has at least $n - 1$ eigenvalues no smaller than $1 - L > 0$. Finally, since $\boldsymbol{I}_n - \widehat{\boldsymbol{H}}$ is symmetric with eigenvalues in $[0, 1]$ by Proposition 7.3,

$$n - \widehat{\mathsf{df}} = \operatorname{trace}[\boldsymbol{I}_n - \widehat{\boldsymbol{H}}] \geq \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \geq (n - 1)(1 - L)^2 \geq nC_*$$

with $C_* = (1 - L)^2/2$ thanks to $n \geq 2$. □

7.4. *Lower bound on* $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n$. The following lemmas are useful to bound from below the denominator in (3.24).

LEMMA 7.5. *Let Assumption 3.1 be fulfilled. Then* $\mathbb{E}[\xi_0^2] \le C_6(\gamma,\mu)nR_*$ *and*

$$\mathbb{E}\big[\big((1-\widehat{\mathsf{df}}/n)\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + n^{-1}\langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle\big)^2\big]/R_* \le C_7(\gamma,\mu)n^{-1}, \tag{7.20}$$

$$\mathbb{E}\big[I_{\Omega_0}\big(\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + (n-\widehat{\mathsf{df}})^{-1}\langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle\big)^2\big]/R_* \le C_8(\gamma,\mu)n^{-1}, \tag{7.21}$$

where Ω_0 *is the event from Lemma 7.4.*

PROOF OF LEMMA 7.5. By (3.9) combined with (3.10), (7.11) and (7.9),

$$\mathbb{E}\big[\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2\big] \le \mathbb{E}\big[\|\boldsymbol{w}_0\|^2\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\big] \le \mathbb{E}\big[(F/(2n))4nF_+F^2\big] \le C_9(\gamma,\mu)R_* \tag{7.22}$$

Similarly, by definition of $V^*(\theta)$ in (3.18), $\mathbb{E}[\xi_0^2] = \mathbb{E}[V^*(\theta)] = \mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\nabla f(z_0)\|_F^2]$. Using (7.10)–(7.11), we have $\mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \le 4nR_*\mathbb{E}[F_+F^2]$ and

$$\begin{aligned}\mathbb{E}\big[\|\nabla f(z_0)\|_F^2\big] &\le \mathbb{E}\big[2\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 + 2\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2\big]\\ &\le n\mathbb{E}\big[2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2\big] + C_{10}(\gamma,\mu)R_*\\ &\le nR_*C_{11}(\gamma,\mu)\end{aligned} \tag{7.23}$$

thanks to $\nabla f(z_0)$ in (3.9), and $(a+b)^2 \le 2(a^2+b^2)$ for the first inequality, $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n$ by Proposition 7.3 and (7.22) for the second inequality, and (7.10)–(7.9) for the third inequality. This provides $\mathbb{E}[\xi_0^2] \le C_{12}(\gamma,\mu)nR_*$. Next, (7.20) holds due the bound (7.22) and the relationship in (3.11) between ξ_0, $\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle$ and the integrand in left-hand side of (7.20). Then (7.21) follows from (7.20) and $I_{\Omega_0}(1-\widehat{\mathsf{df}}/n)^{-2} \le C_*(\gamma,\mu)^{-2}$ by Lemma 7.4. □

LEMMA 7.6. *Let* $\widehat{\boldsymbol{\beta}}$ *be as in (3.1) for convex* g *and let* $\boldsymbol{\beta}^*$, $\boldsymbol{h}^*$, R_* *be as in (3.23). Then*

$$(1-\widehat{\mathsf{df}}/n)^2/8 \le \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*) + \Delta_n^a + \Delta_n^b + \Delta_n^c \tag{7.24}$$

$$\le V^*(\theta)/(nR_*) + \Delta_n^d + \Delta_n^a + \Delta_n^b + \Delta_n^c \tag{7.25}$$

where $V^*(\theta)$ *is defined in (3.18) and* $\Delta_n^a, \dots, \Delta_n^d$ *are nonnegative terms defined as*

$$\Delta_n^a \stackrel{\text{def}}{=} \sigma^2\big|(1-\widehat{\mathsf{df}}/n) - \boldsymbol{\varepsilon}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/\big(n\sigma^2\big)\big|^2/R_*, \tag{7.26}$$

$$\Delta_n^b \stackrel{\text{def}}{=} (F_+ - 1)_+\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*), \tag{7.27}$$

$$\Delta_n^c \stackrel{\text{def}}{=} \big|(1-\widehat{\mathsf{df}}/n)\langle \boldsymbol{a}_*, \boldsymbol{h}\rangle - \boldsymbol{g}^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})/n\big|^2/R_* \tag{7.28}$$

$$\Delta_n^d \stackrel{\text{def}}{=} n^{-1}|2\boldsymbol{w}_0^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle|/R_*.$$

where $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{h}^*/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$ *and* $\boldsymbol{a}_* = \boldsymbol{\Sigma}\boldsymbol{h}^*/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|$.

PROOF OF LEMMA 7.6. By the triangle inequality and definitions of Δ_n^a, Δ_n^c,

$$(1-\widehat{\mathsf{df}}/n)\sigma \le (\boldsymbol{\varepsilon}/\sigma)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n + \big(\Delta_n^a R_*\big)^{1/2}, \tag{7.29}$$

$$(1-\widehat{\mathsf{df}}/n)\langle \boldsymbol{a}_*, \boldsymbol{h}\rangle \le \big(\boldsymbol{g}^\top(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\big)/n + \big(\Delta^c R_*\big)^{1/2}, \tag{7.30}$$

$$(1-\widehat{\mathsf{df}}/n)\big(\sigma^2 + \boldsymbol{h}^\top\boldsymbol{\Sigma}\boldsymbol{h}^*\big) \le \big(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}^*\big)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n + \big(\Delta_n^a\big)^{1/2}R_* + \big(\Delta_n^c\big)^{1/2}R_*,$$

where the last line follows from the weighted sum $\sigma(7.29) + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|(7.30)$ and using $\sigma \vee \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\| \le {R_*}^{1/2}$ for the last two terms. By the KKT conditions of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$,

$$(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^\top \partial g(\boldsymbol{\beta}^*) = (\boldsymbol{h} - \boldsymbol{h}^*)^\top \boldsymbol{\Sigma}\boldsymbol{h}^*,$$
$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \partial g(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n.$$

Summing these equalities and using the monotonicity of the subdifferential yields

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}^*\|^2 + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n \\
&\quad \le \boldsymbol{h}^\top \boldsymbol{\Sigma}\boldsymbol{h}^* + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n \\
&\quad = \boldsymbol{h}^\top \boldsymbol{\Sigma}\boldsymbol{h}^* + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n.
\end{aligned} \tag{7.31}$$

Combining (7.31) multiplied by $1 - \widehat{\mathsf{df}}/n$ with the line after (7.30) gives

$$\begin{aligned}
&(1 - \widehat{\mathsf{df}}/n)(R_* + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n) \\
&\quad \le (2 - \widehat{\mathsf{df}}/n)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n + (\Delta_n^a)^{1/2}R_* + (\Delta_n^c)^{1/2}R_* \\
&\quad \le 2\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/n + 2(\max\{\Delta_n^a, \Delta_n^c\})^{1/2}R_*
\end{aligned}$$

using the Cauchy–Schwarz inequality and $(2 - \widehat{\mathsf{df}}/n) \le 2$ for the last inequality. Using $(2a + 2b)^2 \le 8(a^2 + b^2)$ for the right-hand side with $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n^2{R_*}^2) \le \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*) + \Delta_n^b$ completes the proof of (7.24). The second inequality, (7.25), then follows from (3.9), (3.18) and

$$\begin{aligned}
&\operatorname{trace}\big[\big((\boldsymbol{I}_n - \widehat{\boldsymbol{H}})\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + \boldsymbol{w}_0(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big)^2\big] \\
&\quad = \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 + \big(\boldsymbol{w}_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big)^2 + 2\boldsymbol{w}_0^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle,
\end{aligned}$$

which implies $V^*(\theta) - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \ge -n\Delta_n^d R_*$. □

LEMMA 7.7. *Define* $\Delta_n \overset{\text{def}}{=} \Delta_n^a + \Delta_n^b + \Delta_n^c + \Delta_n^d$ *where* $\Delta_n^a, \dots, \Delta_n^d$ *are defined in Lemma* 7.6. *Under Assumption* 3.1, *we have* $\mathbb{E}[\Delta_n] \le C(\gamma, \mu)n^{-1/2}$.

PROOF OF LEMMA 7.7. We bound each of Δ_n^a, Δ_n^b, Δ_n^c, Δ_n^d separately. We have $\Delta_n^b \le (F_+ - 1)4F_+F^2$ by (7.11) so that $\mathbb{E}[\Delta_n^b] \le C_{13}(\gamma, \mu)n^{-1/2}$ by virtue of (7.9). For Δ_n^a, we have $\Delta_n^a = n^{-2}\sigma^{-2}|\sigma^2(n - \widehat{\mathsf{df}}) - \boldsymbol{\varepsilon}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})|^2/R_*$. By the second-order Stein formula (Proposition 2.1) with respect to $\boldsymbol{\varepsilon}$ conditionally on $\boldsymbol{X}$,

$$\mathbb{E}[\Delta_n^a] = n^{-2}\mathbb{E}\big[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/R_* + \sigma^2 \operatorname{trace}(\{\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\}^2)/R_*\big] \le n^{-1}\mathbb{E}[4F_+F^2 + 1],$$

where we used $\operatorname{trace}(\{\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\}^2) \le n$ from Proposition 7.3 and (7.11) for the inequality. Thanks to (7.9), this shows that $\mathbb{E}[\Delta_n^a] \le n^{-1}C_{14}(\gamma, \mu)$. Similarly, for Δ_n^d in (7.25), $\Delta_n^d \le 2n^{-1}\|\boldsymbol{w}_0\|\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\||\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle|/R_* \le n^{-1}2(F/2)^{1/2}2F_+F^2$, hence $\mathbb{E}[\Delta_n^d] \le n^{-1}C_{15}(\gamma, \mu)$ by (7.11) and (7.9). For Δ_n^c, we have $\boldsymbol{g} = \boldsymbol{z}_0$ for $\boldsymbol{a}_0 = \boldsymbol{a}_*$ so that $\mathbb{E}[\Delta_n^c] \le C_{16}(\gamma, \mu)n^{-1}$ by (7.20). □

7.5. *Event* Ω_n. With Ω_0, $C_*(\gamma, \mu)$ in Lemma 7.4 and Δ_n in Lemma 7.7, let

$$\Omega_n = \Omega_0 \cap \{\mathbb{E}_0[\Delta_n] \vee \Delta_n \le C_*^2(\gamma, \mu)/16\}. \tag{7.32}$$

By the union bound, Markov's inequality and the bound on $\mathbb{E}[\Delta_n]$ in Lemma 7.7,

$$\mathbb{P}(\Omega_n^c) \le \mathbb{P}(\Omega_0^c) + C_{17}(\gamma, \mu)n^{-1/2} \le C_0 n^{-1/2} \tag{7.33}$$

thanks to $\mathbb{P}(\Omega_0^c) \le e^{-n/2}$ in Lemma 7.4. By (7.24),

$$\Omega_n \subset \{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \ge R_* n C_*^2(\gamma,\mu)/16\}. \tag{7.34}$$

Since Ω_0^c is independent of $\boldsymbol{z}_0$, taking the condition expectation $\mathbb{E}_0$ of (7.25) in Ω_0 gives

$$\Omega_n \subset \{\mathrm{Var}_0[\xi_0] \wedge \mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2] \ge R_* n C_*^2(\gamma,\mu)/16\}. \tag{7.35}$$

Proofs of Lemmas 3.2, 3.4 and 3.5 and Theorem 3.9.

LEMMA 3.2. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\mathbb{E}[I_{\Omega_n}\langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2 / \mathrm{Var}_0[\xi_0]] \le C_{18}(\gamma,\mu)n^{-1}. \tag{3.14}$$

PROOF OF LEMMA 3.2. With Ω_n in (7.32), (3.14) follows from (7.35) and (7.22). □

LEMMA 3.4. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\mathbb{E}\left[I_{\Omega_n}\left|\frac{\mathbb{E}_0[\widehat{V}(\theta)]}{\mathrm{Var}_0[\xi_0]} - 1\right|\right] \le \mathbb{E}\left[I_{\Omega_n}\frac{\mathbb{E}_0[|\widehat{V}(\theta) - V^*(\theta)|]}{\mathrm{Var}_0[\xi_0]}\right] \le \frac{C_{19}(\gamma,\mu)}{n}. \tag{3.20}$$

PROOF OF LEMMA 3.4. Let Ω_n be as in (7.32). The first inequality in (3.20) follows from the triangle inequality. By (7.35), we have $\mathbb{E}[I_{\Omega_n}\mathbb{E}_0[|\widehat{V}(\theta) - V^*(\theta)|]/\mathrm{Var}_0[\xi_0]] \le \mathbb{E}[|\widehat{V}(\theta) - V^*(\theta)|]16/(nR_*C_*^2(\gamma,\mu))$. With $V^*(\theta)$, $\widehat{V}(\theta)$ in (3.18)–(3.19) and $\nabla f(\boldsymbol{z}_0)^\top$ in (3.9),

$$V^*(\theta) - \widehat{V}(\theta) = \langle \boldsymbol{w}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2 + 2\boldsymbol{w}_0^\top(\boldsymbol{I}_n - \widehat{\boldsymbol{H}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle.$$

Using $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_{\mathrm{op}} \le 1$ from Proposition 7.3 and (7.10)–(7.11), we find by the Cauchy–Schwarz inequality $|V^*(\theta) - \widehat{V}(\theta)| \le (2 + 2\sqrt{2})R_*F_+F^3$. The proof of (3.20) is complete by virtue Holder's inequality and the moment bounds (7.9). □

LEMMA 3.5. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\max\left\{\mathbb{E}\left[I_{\Omega_n}\left|\frac{\check{V}(\boldsymbol{a}_0)^{1/2}}{\widehat{V}(\theta)^{1/2}} - 1\right|^2\right], \mathbb{E}\left[I_{\Omega_n}\left|\frac{\mathbb{E}_0[\check{V}(\boldsymbol{a}_0)]^{1/2}}{\mathbb{E}_0[\widehat{V}(\theta)]^{1/2}} - 1\right|^2\right]\right\} \le \frac{C_{20}(\gamma,\mu)}{n}. \tag{3.22}$$

PROOF OF LEMMA 3.5. By the triangle inequality for the Euclidean norm in $\mathbb{R}^2$,

$$|\check{V}(\boldsymbol{a}_0)^{1/2} - \widehat{V}(\theta)^{1/2}| \le \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F|\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle| + (n - \widehat{\mathsf{df}})^{-1}|\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle|. \tag{7.36}$$

Let Ω_n be as in (7.32). Using $\widehat{V}(\theta) \ge \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, the lower bound (7.34) and $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n$ by Proposition 7.3,

$$\begin{aligned}&\mathbb{E}[I_{\Omega_n}|\check{V}(\boldsymbol{a}_0)^{1/2}/\widehat{V}(\theta)^{1/2} - 1|^2]\\ &\quad\le \mathbb{E}[I_{\Omega_n}(\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + (n - \widehat{\mathsf{df}})^{-1}\langle \boldsymbol{z}_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle)^2]16/(R_*C_*^2(\gamma,\mu))\end{aligned}$$

so that $\Omega_n \subset \Omega_0$ and (7.21) completes the proof for the first term in the maximum in (3.22). For the second term in the maximum, by the triangle inequality for the norm $\mathbb{E}_0[(\cdot)^2]^{1/2}$, we have $|\mathbb{E}_0[\check{V}(\boldsymbol{a}_0)]^{1/2} - \mathbb{E}_0[\widehat{V}(\theta)]^{1/2}| \le \mathbb{E}_0[|\check{V}(\boldsymbol{a}_0)^{1/2} - \widehat{V}(\theta)^{1/2}|^2]^{1/2}$. The proof is completed by using again (7.36), the lower bound (7.35) on $\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]$ in Ω_n and the same argument as for the first term in the maximum. □

THEOREM 3.9. *Let Assumption 3.1 be fulfilled. Then the following are equivalent*:

(i) $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/\mathrm{Var}_0[\xi_0] \to^{\mathbb{P}} 1$,

(ii) $\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$,
(iii) $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$,
(iv) $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2 n/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 0$,
(v) $\langle z_0, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle^2/(n\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2) \to^{\mathbb{P}} 0$,
(vi) $\widehat{V}(\theta)/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$,
(vii) $\check{V}(\boldsymbol{a}_0)/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \to^{\mathbb{P}} 1$.

PROOF OF THEOREM 3.9. (v) $\Leftrightarrow$ (iv) is due to $C_*(\gamma,\mu)n \le \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n$ in Ω_n by Lemma 7.4 and Proposition 7.3 combined with (3.17).

(iv) $\Leftrightarrow$ (vi) follows from $C_*(\gamma,\mu)n \le \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \le n$ in Ω_n.

(vi) $\Leftrightarrow$ (vii) is proved in Lemma 3.5.

(iii) $\Rightarrow$ (i), (iii) $\Rightarrow$ (vi) and (iii) $\Rightarrow$ (vii) are shown in the proof of Theorem 3.6.

(iv) $\Rightarrow$ (iii) follows from $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*) = O_{\mathbb{P}}(1)$ by (7.11) and (7.9).

(iii) $\Rightarrow$ $\delta_1^2 \to^{\mathbb{P}} 0$ was shown in the proof of Theorem 3.6, and $\delta_1^2 \to^{\mathbb{P}} 0$ implies $\mathbb{E}[\|\nabla f(z_0)\|_F^2]/\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2] \to^{\mathbb{P}} 0$ and $\mathbb{E}[\operatorname{trace}[(\nabla f(z_0))^2]]/\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2] \to^{\mathbb{P}} 0$ so that (iii) $\Rightarrow$ (ii) holds.

By Lemma 3.4, (ii) implies that $\mathbb{E}_0[\widehat{V}(\theta)]/\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2] - 1 = \mathbb{E}_0[\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2]/\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]$ converges to 0 in probability. Since $\mathbb{E}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]/(nR_*) = O_{\mathbb{P}}(1)$ by (7.11) and (7.9), combined with $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F^2 \ge C_*^2(\gamma,\mu)n$ in Ω_0 by Lemma 7.4, this implies $I_{\Omega_0}\mathbb{E}_0[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2]/R_* = \mathbb{E}_0[I_{\Omega_0}\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2]/R_* \to^{\mathbb{P}} 0$ as Ω_0 is independent of z_0. Thus, (ii) implies (iii) by Markov's inequality with respect to $\mathbb{E}_0$.

Finally, to show (i) $\Leftrightarrow$ (ii), we have by the Gaussian Poincaré inequality

$$\operatorname{Var}_0\big[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\big] \le \mathbb{E}_0\big[\big\|\big[\nabla f(z_0)\big](\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big\|^2\big] \le \mathbb{E}_0\big[\|\nabla f(z_0)\|_{\mathrm{op}}^2\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\big].$$

With $\nabla f(z_0)$ in (3.9) and the bounds (7.10)–(7.11), we have $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \le 4nF_+F^2R_*$ and $\|\nabla f(z_0)\|_{\mathrm{op}}^2 \le 2(2F_+F^3 + F_+F^2)R_*$ thanks to $\|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_{\mathrm{op}} \le 1$ by Proposition 7.3. Combined with the lower bound (7.35) on $\operatorname{Var}_0[\xi_0]$ and the moment upper bounds (7.9), we obtain $\mathbb{E}[I_{\Omega_n}\operatorname{Var}_0[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2]/\operatorname{Var}_0[\xi_0]^2] \le C_{21}(\gamma,\mu)n^{-1/2}$, which gives (vii) $\Leftrightarrow$ (i). □

7.6. *Proofs of Theorem 3.3 and asymptotic normality results.*

THEOREM 3.3. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\mathbb{E}\big[I_{\Omega_n}(n - \widehat{\mathsf{df}})^2\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle^2/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\big] \le C_{22}(\gamma,\mu). \tag{3.17}$$

Furthermore, $|\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle| = O_{\mathbb{P}}(1)\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/(n - \widehat{\mathsf{df}}) = O_{\mathbb{P}}(1)\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/n$.

PROOF OF THEOREM 3.3. Let Ω_n be as in (7.32). Since $I_{\Omega_n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^{-2} \le 16/(C_*^2(\gamma,\mu)nR_*)$ by (7.34), using (7.21) completes the proof of (3.17). For the second part, random variables bounded in L_2 are stochastically bounded so that (3.3) provides $|\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle| = O_{\mathbb{P}}(1)\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/(n - \widehat{\mathsf{df}})$, and $I_{\Omega_0}(1 - \widehat{\mathsf{df}}/n)^{-1} \le C_*(\gamma,\mu)^{-1}$ for Ω_0 in Lemma 7.4 provides $(1 - \widehat{\mathsf{df}}/n) = O_{\mathbb{P}}(1)$. □

THEOREM 3.6. *Let Assumption 3.1 be fulfilled. Let $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ be as in (3.15). Then, for any $\boldsymbol{a}_0$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$ such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$,*

$$\sup_{t\in\mathbb{R}}\left[\left|\mathbb{P}\left(\frac{\xi_0}{V_0^{1/2}} \le t\right) - \Phi(t)\right| + \left|\mathbb{P}\left(\frac{\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}\rangle - \theta}{V_0^{1/2}/(n - \widehat{\mathsf{df}})} \le t\right) - \Phi(t)\right|\right] \to 0,$$

where V_0 denotes any of the four quantities: $\operatorname{Var}_0[\xi_0]$, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{V}(\theta)$ *or* $\check{V}(\boldsymbol{a}_0)$.

PROOF OF THEOREM 3.6. Let Ω_n be as in (7.32). Let δ_1^2 be the quantity in (3.24), omitting the dependence in $\boldsymbol{a}_0$ as it is clear from context. Since $\delta_1^2 \le 1$ by definition, $\mathbb{E}[\delta_1^2] \le \mathbb{E}[\Omega_n \delta_1^2] + \mathbb{P}(\Omega_n^c)$. In Ω_n, (7.35) provides a lower bound on the denominator of δ_1^2 so that $\mathbb{E}[I_{\Omega_n}\delta_1^2] \le \mathbb{E}[\|\nabla f(z_0)\|_F^2]16/(nR_*C_*^2(\gamma,\mu))$. By (7.23) and the bound (7.33) on $\mathbb{P}(\Omega_n^c)$, we obtain

$$\mathbb{E}\big[\delta_1^2\big] \le \mathbb{E}\big[I_{\Omega_n}\delta_1^2\big] + \mathbb{P}\big(\Omega_n^c\big) \le C_{23}(\gamma,\mu)\big(\mathbb{E}\big[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*\big] + n^{-1}\big) + C_0(\gamma,\mu)n^{-1/2} \tag{7.37}$$

Furthermore, $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*$ is bounded in L_2 thanks to $\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^4/{R_*}^2] \le \mathbb{E}[F_+^2F^4] \le C_{24}(\gamma,\mu)$ by (7.10) and (7.9). Since a sequence of random variables uniformly bounded in L_2 is uniformly integrable, the assumption $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ implies $\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*] \to 0$, and thus $\mathbb{E}[\delta_1^2] \to 0$. This completes the proof that $\mathbb{E}[\delta_1^2] \to 0$ and that $\xi_0/\operatorname{Var}_0[\xi_0]^{1/2} \to^d N(0,1)$ by Theorem 2.2. Next, by (3.16), $(n-\widehat{\mathsf{df}})\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle/\operatorname{Var}_0[\xi_0]^{1/2} \to^d N(0,1)$ also holds. It remains to prove $V_0/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ for all four possible choices for V_0. By (2.7), $\mathbb{E}[\delta_1^2] \to 0$ implies $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$, while

$$0 \le \frac{\widehat{V}(\theta)}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2} - 1 = \frac{\|\widehat{\boldsymbol{H}} - \boldsymbol{I}_n\|_F^2\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/(nR_*)}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*)}. \tag{7.38}$$

Proposition 7.3 provides $\|\widehat{\boldsymbol{H}} - \boldsymbol{I}_n\|_F^2 \le n$ so that the numerator converges to 0 in probability thanks to assumption $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$. The denominator is bounded from below by $C_*^2(\gamma,\mu)/16$ in Ω_n by (7.34) and $\mathbb{P}(\Omega_n) \to 1$. This proves $\widehat{V}(\theta)/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ and $\check{V}(\boldsymbol{a}_0)/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ follows by Lemma 3.5. Slutsky's theorem completes the proof as $V_0/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$ for all four possible choices for V_0. As $\Phi(t)$ is continuous, convergence in Kolmogorov distance is equivalent to convergence in distribution. □

THEOREM 3.7. *There exists an absolute constant $C^* > 0$ such that the following holds. Let Assumption 3.1 be fulfilled, $\widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})}$ be as in (3.15). Then for any increasing sequence $a_p \to +\infty$ (e.g., $a_p = \log\log p$), the subset*

$$\overline{S} = \{\boldsymbol{v} \in S^{p-1} : \mathbb{E}[\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2] \le C^*/a_p\} \tag{3.25}$$

of the unit sphere S^{p-1} in $\mathbb{R}^p$ has relative volume $|\overline{S}|/|S^{p-1}| \ge 1 - 2e^{-p/a_p}$ and

$$\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \sup_{t\in\mathbb{R}} \left[\left|\mathbb{P}\left(\frac{\xi_0}{V_0^{1/2}} \le t\right) - \Phi(t)\right| + \left|\mathbb{P}\left(\frac{\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}}^{(\mathrm{de-bias})} - \boldsymbol{\beta}\rangle}{V_0^{1/2}/(n-\widehat{\mathsf{df}})} \le t\right) - \Phi(t)\right|\right] \to 0 \tag{3.26}$$

where V_0 denotes any of the four quantities: $\operatorname{Var}_0[\xi_0]$, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{V}(\theta)$ *or* $\check{V}(\boldsymbol{a}_0)$. *Furthermore, with $\boldsymbol{e}_j \in \mathbb{R}^p$ the jth canonical basis vector and $\phi_{\mathrm{cond}}(\boldsymbol{\Sigma}) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}}$, the asymptotic normality in (3.26) uniformly holds over at least $(p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*)$ canonical directions in the sense that $J_p = \{j \in [p] : \boldsymbol{e}_j/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\| \in \boldsymbol{\Sigma}^{1/2}\overline{S}\}$ has cardinality $|J_p| \ge p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*$.*

PROOF OF THEOREM 3.7. We construct a subset $\overline{S}$ of the sphere such that $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ uniformly over all $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$. Let $\boldsymbol{v}$ be uniformly distributed on the unit Euclidean sphere S^{p-1}, independently of $(\boldsymbol{X}, \boldsymbol{y})$, and denote by ν its probability measure. The vector $\sqrt{p}\boldsymbol{v}$ is sub-Gaussian in $\mathbb{R}^p$ [47], Theorem 3.4.6, in the sense that for any nonzero vector $\boldsymbol{u} \in \mathbb{R}^p$, $\int \exp\{(\sqrt{p}\boldsymbol{v}^\top\boldsymbol{u})^2/(C^*\|\boldsymbol{u}\|^2)\}\,d\nu(\boldsymbol{v}) \le 2$ for some absolute constant $C^* > 0$. By Jensen's inequality and Fubini's theorem,

$$\int \exp\left\{\mathbb{E}\left[\frac{(\boldsymbol{v}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{h})^2}{C^*\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2}\right]\right\} d\nu(\boldsymbol{v}) \le \mathbb{E}\left[\int \exp\left\{\frac{(\sqrt{p}\boldsymbol{v}^\top\boldsymbol{h})^2}{C^*\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2}\right\} d\nu(\boldsymbol{v})\right]$$
$$\le 2.$$

Hence, by Markov's inequality, for any positive x,

$$\nu\big(\{\boldsymbol{v} \in S^{p-1} : \mathbb{E}\big[(\boldsymbol{v}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{h})^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2\big] > C^* x/p\}\big) \le 2e^{-x}.$$

Setting $x = p/a_p$, we obtain that the subset $\overline{S} \subset S^{p-1}$ defined by (3.25) has relative volume at least $|\overline{S}|/|S^{p-1}| \ge 1 - 2e^{-p/a_p}$, and for all $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$,

$$\mathbb{E}\big[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2\big] \le C^*/a_p. \tag{7.39}$$

Furthermore, the set $\overline{S} \cap \{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|, j \in [p]\}$ has cardinality at least $p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma})a_p/C^*$ due to

$$\sum_{j=1}^{p} \frac{1}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|^2}\mathbb{E}\bigg[\frac{\langle \boldsymbol{e}_j, \boldsymbol{h}\rangle^2}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2}\bigg] \le \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\mathbb{E}\bigg[\frac{\|\boldsymbol{h}\|^2}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2}\bigg] \le \phi_{\mathrm{cond}}(\boldsymbol{\Sigma}).$$

To show that $\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \mathbb{E}[\delta_1^2(\boldsymbol{a}_0)] \to 0$, thanks to (7.37) it is enough to prove that $\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*] \to 0$ uniformly over $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$. By the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}\big[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*\big] &= \mathbb{E}\big[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|/R_*\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|\big] \\ &\le \mathbb{E}\big[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^4/{R_*}^2\big]^{1/2}\big(C^*/a_p\big)^{1/2}\end{aligned}$$

for any $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ thanks to (7.39), while $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^4/{R_*}^2] \le \mathbb{E}[F_+^2F^4] \le C_{25}(\gamma,\mu)$ by (7.10) and (7.9). This implies $\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*] \to 0$ and $\sup_{\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}} \mathbb{E}[\delta_1^2(\boldsymbol{a}_0)] \to 0$ hold.

By Theorem 2.2, this shows that $\xi_0/\operatorname{Var}_0[\xi_0]^{1/2} \to^d N(0,1)$ uniformly over $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$. Since the bounds (3.14), (7.38) are all uniform over all $\boldsymbol{a}_0$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$, Slutsky's theorem implies $V_0/\operatorname{Var}_0[\xi_0] \to^{\mathbb{P}} 1$, $\xi_0/V_0^{1/2} \to^d N(0,1)$ and $\{(n - \widehat{\mathsf{df}})\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle + z_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\}/V_0^{1/2} \to^d N(0,1)$ uniformly over $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ for all four possible choices for V_0, and convergence in Kolmogorov distance follows from convergence in distribution. □

THEOREM 3.8. *Under Assumption 3.1, there exists Ω_n with $\mathbb{P}(\Omega_n^c) \le C_0(\gamma,\mu)n^{-1/2}$ and*

$$\mathbb{E}\big[I_{\Omega_n}\big(\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle + (n - \widehat{\mathsf{df}})^{-1}z_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big)^2\big] \le R_*C_{26}(\gamma,\mu)/n \tag{3.27}$$

If additionally g is a seminorm, then $|z_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})|/n = |\boldsymbol{a}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})|/n \le g(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)$ always holds by properties of the subdifferential of a norm. Consequently, if $g(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)^2/R_* \to 0$ then $\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_* \to^{\mathbb{P}} 0$ and the conclusions of Theorem 3.6 hold.

PROOF OF THEOREM 3.8. The first statement of the theorem follows from Lemma 7.5. Finally, if g is a norm then the KKT conditions of $\widehat{\boldsymbol{\beta}}$, $|z_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})| = n|(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)^\top\partial g(\widehat{\boldsymbol{\beta}})| \le ng(\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_0)$ since for a norm $g(\boldsymbol{u}) = \sup_{\boldsymbol{v} \in \partial g(\boldsymbol{u})}\langle \boldsymbol{u}, \boldsymbol{v}\rangle$. □

APPENDIX A: INTEGRABILITY OF $\phi_{\min}^{-1}(\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}/\sqrt{n})$ WHEN $p/n \to \gamma \in (0,1)$

In our regression model with Gaussian covariates, the matrix $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ has i.i.d. $N(0,1)$ entries, and the inverse of its smallest singular value enjoys the following integrability property as $n, p \to +\infty$ with $p/n \to \gamma \in (0,1)$.

PROPOSITION A.1. *Let $n > p$ and let $\boldsymbol{G}$ be a matrix with n rows, p columns and i.i.d. $N(0,1)$ entries. Then $\boldsymbol{G}^\top\boldsymbol{G}$ is a Wishart matrix and if $n, p \to +\infty$ with $p/n \to \gamma \in (0,1)$ we have for any constant k not growing with n, p,*

$$\lim_{p/n \to \gamma} \mathbb{E}\big[\phi_{\min}(\boldsymbol{G}^\top\boldsymbol{G}/n)^{-k}\big] = (1 - \sqrt{\gamma})^{-2k}$$

PROOF. Throughout the proof, $p = p_n$ is an implicit function of n; we omit the subscript for brevity. Since $S_n = \phi_{\min}(\boldsymbol{G}^\top \boldsymbol{G}/n) \to (1-\sqrt{\gamma})^2$ almost surely (cf. [36]), it is enough to show that the sequence of random variables $(S_n^{-k})_{n\geq n_0}$ is uniformly integrable for some $n_0 > 0$, that is, that $\sup_{n\geq n_0} \mathbb{E}[S_n^{-k} I_{\{S_n<\epsilon\}}] \to 0$ as $\epsilon \to 0$. For uniform integrability, we use the following argument from [22], Section 5. The matrix $\boldsymbol{G}^\top \boldsymbol{G}$ is a Wishart matrix and the density of $L = \phi_{\min}(\boldsymbol{G}^\top \boldsymbol{G})$ satisfies for $\lambda \geq 0$,

$$f_L(\lambda) \leq \frac{\sqrt{\pi} 2^{-(n-p+1)/2} \Gamma(\frac{n+1}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{n-p+1}{2})\Gamma(\frac{n-p+2}{2})} \lambda^{(n-p-1)/2} e^{-\lambda/2} = \frac{\sqrt{\pi}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{n-p+2}{2})} f_{\chi^2_{n-p+1}}(\lambda);$$

cf. [22], Section 5. The density of $S_n = L/n = \phi_{\min}(\boldsymbol{G}^\top \boldsymbol{G}/n)$ that we are interested in is given by $f_{S_n}(x) = n f_L(nx)$ for $x \geq 0$. Hence, if $0 < \epsilon < (1-\gamma)/2$,

$$\mathbb{E}[S_n^{-k} I_{\{S<\epsilon\}}] \leq \left[\frac{\sqrt{\pi}\Gamma(\frac{n+1}{2})(\frac{n}{2})^{(n-p+1)/2}}{\Gamma(\frac{p}{2})\Gamma(\frac{n-p+1}{2})\Gamma(\frac{n-p+2}{2})}\right] \int_0^\epsilon x^{(n-p-1)/2-k} e^{-nx/2}\, dx.$$

The mode of the integrand over $[0, +\infty)$ is $x_n^* = 1 - p/n - 1/n - 2k/n$. Thanks to $\epsilon < (1-\gamma)/2$, there exists some $n_1 \geq 1$ such that for all $n \geq n_1$,

$$n - p - 1 - 2k \geq n(1-\gamma)/2, \tag{A.1}$$

$(1-\gamma)/2$ is smaller than the mode x_n^* and the integral above is bounded by $\epsilon^{(n-p-k+1)/2} \times e^{-n\epsilon/2}$. Let Λ_n denote the bracket of the previous display. Then using Stirling's formula $\Gamma(x+1) \asymp \sqrt{2\pi x} e^{-x} x^x$, we have for some constants $n_2, C_2(\gamma) > 0$ possibly depending on γ,

$$\sup_{n\geq n_2} \frac{\log(\Lambda_n)}{(n-p+1)/2} \leq C_2(\gamma)$$

because the main terms (coming from x^x in Stirling's formula) cancel each other. Then for any $n \geq n_1 \vee n_2$,

$$\begin{aligned} \mathbb{E}[S_n^{-k} I_{\{S_n<\epsilon\}}] &\leq \big(\exp(C_2(\gamma))\epsilon\big)^{(n-p+1)/2} \epsilon^{-k} e^{-n\epsilon/2} \\ &\leq \big(\exp(C_2(\gamma))\epsilon\big)^{(n-p+1)/2-k} e^{kC_2(\gamma)-n\epsilon/2}. \end{aligned} \tag{A.2}$$

For $n \geq n_1$, (A.1) holds and if $\epsilon < (\exp C_2(\gamma))^{-1}$ we have

$$\sup_{n\geq n_1\vee n_2} \mathbb{E}[S_n^{-k} I_{\{S_n<\epsilon\}}] \leq \big(\exp(C_2(\gamma))\epsilon\big)^{(n_1(1-\gamma)/4)} e^{kC_2(\gamma)}$$

which converges to 0 as $\epsilon \to 0$. This shows uniform integrability of the sequence and proves the claim. □

APPENDIX B: PROOF: $p > n$ WITHOUT STRONG CONVEXITY

LEMMA B.1. *Let $\boldsymbol{\beta} \in \mathbb{R}^p$ and assume that $p/n \leq \gamma$. Then for any $\kappa < 1$,*

$$\mathbb{P}\left(\inf_{t\in\mathbb{R}, \boldsymbol{u}\in\mathbb{R}^p : \|\boldsymbol{u}\|_0 \leq \kappa n, \boldsymbol{u}-t\boldsymbol{\beta}\neq\boldsymbol{0}} \left(\frac{\|\boldsymbol{X}(\boldsymbol{u}-t\boldsymbol{\beta})\|^2}{n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{u}-t\boldsymbol{\beta})\|^2}\right) > \varphi(\gamma,\kappa)^2\right) \to 1,$$

for some constant $\varphi(\gamma,\kappa) > 0$ depending only on γ, κ.

Lemma B.1 and its proof are straightforward extensions of [12], Proposition 2.10, which treats the case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{\beta} = \boldsymbol{0}$.

PROOF. If $V \subset \mathbb{R}^p$ is a subspace of dimension $d = \lfloor \kappa n \rfloor + 1$ and $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$, then by (A.2) with $k = 0$, $\epsilon \in (0, (1 - d/n)/2)$ and n large enough,

$$\mathbb{P}\Big(\inf_{\boldsymbol{v} \in \boldsymbol{\Sigma}^{1/2} V : \|\boldsymbol{v}\| = 1} \|\boldsymbol{G}\boldsymbol{v}\|^2/n < \epsilon \Big) \le \exp\big(C_2(\kappa') \log(\epsilon)(n - p + 1)/2 - n\epsilon/2 \big)$$

for constant $\kappa' = (\kappa + 1)/2$ thanks to $1 > \kappa' \ge d/n$. Applying this bound to the subspace $V_B = \{\boldsymbol{u} - t\boldsymbol{\beta}, (\boldsymbol{u}, t) \in \mathbb{R}^{p+1} : \boldsymbol{u}_{B^c} = \boldsymbol{0}\}$ for $B \subset [p]$ with $|B| \le \kappa n$ and using the union bound,

$$\begin{aligned}\mathbb{P}\Big(\inf_{t \in \mathbb{R}, \boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_0 \le \kappa n} \Big(\frac{\|\boldsymbol{X}(\boldsymbol{u} - t\boldsymbol{\beta})\|^2}{n \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{u} - t\boldsymbol{\beta})\|^2} \Big) < \epsilon \Big) &\le \binom{p}{\lfloor \kappa n \rfloor} e^{C_2(\kappa') \log(\epsilon)(n - d + 1)/2 - n\epsilon/2} \\ &\le e^{n \log(e\gamma) + C_2(\kappa') \log(\epsilon)(n - d + 1)/2 - n\epsilon/2}\end{aligned}$$

using $\binom{p}{q} \le e^{q \log(ep/q)} \le e^{n \log(ep/n)}$ with $q = \lfloor \kappa n \rfloor \le n$ and $p/n \le \gamma$. Since $d \le \kappa n + 1$, choosing $\epsilon = 1 \wedge \exp(C_2(\kappa')^{-1}(1 - \kappa)^{-1} 2 \log(e\gamma))$ the right-hand side of the previous display is bounded from above by $e^{-n\epsilon/2}$. This value of ϵ provides $\varphi(\gamma, \kappa)^2$. $\square$

THEOREM 3.10. *Let $\gamma \ge 1$, $\kappa \in (0, 1)$ be constants independent of $\{n, p\}$. Consider a sequence of regression problems with $p/n \le \gamma$ and invertible $\boldsymbol{\Sigma}$. Assume that the group Lasso estimator $\widehat{\boldsymbol{\beta}}$ in (3.33) satisfies*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}\|_0 \le \kappa n/2 \big) \to 1. \tag{3.34}$$

If $\boldsymbol{a}_0$ is such that $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_0\| = 1$ and $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle^2 / R_* \xrightarrow{\mathbb{P}} 0$ for the R_* in (3.23), then

$$\sup_{t \in \mathbb{R}} \big| \mathbb{P}\big(\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^{-1} \big((n - \widehat{\mathrm{df}}) \langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle + z_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \big) \le t \big) - \Phi(t) \big| \to 0. \tag{3.35}$$

Furthermore, for any a_p with $a_p \to \infty$ and $\overline{S}$ in (3.25), the relative volume bound given after (3.25) holds, and the asymptotic normality (3.35) holds uniformly over all $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ and uniformly over at least $(p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma}) a_p / C^*)$ canonical directions in the sense that $J_p = \{j \in [p] : \boldsymbol{e}_j / \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\| \in \boldsymbol{\Sigma}^{1/2}\overline{S}\}$ has cardinality $|J_p| \ge p - \phi_{\mathrm{cond}}(\boldsymbol{\Sigma}) a_p / C^*$.

PROOF OF THEOREM 3.10. As in the rest of the paper, $f(z_0) = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and we wish to apply Theorem 2.2 to z_0 conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$. Instead of applying Theorem 2.2 to f, and in order to avoid certain events of small probability where the sparse eigenvalues of $\boldsymbol{X}$ are not well behaved, we will apply it to a different function. Consider F_+ in (7.7) and the events,

$$\begin{aligned}\Omega_L &= \big\{ \|\widehat{\boldsymbol{\beta}}\|_0 \le \kappa n/2 \big\}, \\ \Omega_\chi &= \big\{ F_+ < 2, (F_+ - 1)_+ < 4\sqrt{\log(n)/n} \big\}, \\ \Omega_E &= \Big\{ \min_{t \in \mathbb{R}, \boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_0 \le \kappa n} \frac{\|\boldsymbol{X}(\boldsymbol{u} - t\boldsymbol{\beta})\|}{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{u} - t\boldsymbol{\beta})\|} > \varphi\sqrt{n}, \big\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\big\|_{\mathrm{op}} < \sqrt{n}(2 + \sqrt{\gamma}) \Big\},\end{aligned}$$

where $\varphi = \varphi(\gamma, \kappa)$ is the constant from Lemma B.1. Finally, let Ω_{KKT} be the event (C.1) that the KKT conditions of $\widehat{\boldsymbol{\beta}}$ hold strictly, and set

$$\Omega \stackrel{\mathrm{def}}{=} \Omega_L \cap \Omega_E \cap \Omega_{\mathrm{KKT}} \cap \Omega_\chi.$$

We have $\mathbb{P}(\Omega_L) \to 1$ by (3.34) and standard concentration bounds for χ^2_n random variables [29], Lemma 1, give $\mathbb{P}(\Omega_\chi) \to 1$. Lemma B.1 and [20], Theorem II.13, provide $\mathbb{P}(\Omega_E) \to 1$ and (C.1) gives $\mathbb{P}(\Omega_{\mathrm{KKT}}) = 1$. These bounds imply $\mathbb{P}(\Omega) \to 1$ by the union bound.

As the only randomness of the problem comes from $(\boldsymbol{\varepsilon}, \boldsymbol{X})$, we may choose the underlying probability space as $\mathbb{R}^n \times \mathbb{R}^{n\times p}$, so that Ω, Ω_L, Ω_E, Ω_{KKT} are subsets of $\mathbb{R}^n \times \mathbb{R}^{n\times p}$. We next prove that Ω is open as a subset of $\mathbb{R}^n \times \mathbb{R}^{n\times p}$. Indeed, because the KKT conditions are strict in Ω, Ω is a disjoint union of sets of the form

$$\Omega_L \cap \Omega_E \cap \Omega_\chi \cap \{\|\widehat{\boldsymbol{\beta}}_{G_k}\| > 0, k \in B\} \cap \{\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\| < n\lambda_k, k \in B^c\} \tag{B.1}$$

over all possible active groups $B \subset \{1, \dots, K\}$. The sets Ω_E, Ω_χ are open as the inequalities are strict. In Ω_E, the function $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ is locally Lipschitz by Lemma 7.1, hence continuous. By continuity, the preimage of the open set $(0, +\infty)$ by the function $\Omega_E \to \mathbb{R}$, $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \|\widehat{\boldsymbol{\beta}}_{G_k}\|$ is open by continuity, and the preimage of the open set $(-\infty, n\lambda_k)$ by the function $\Omega_E \to \mathbb{R}$, $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|$ is also open, again by continuity. This shows that the set (B.1) is open for any fixed $B \subset \{1, \dots, K\}$ so that Ω is open as the union of sets of the form (B.1) over all $B \subset \{1, \dots, K\}$ satisfying $\sum_{k\in B} |G_k| \le \kappa n/2$. This proves that $\Omega \subset \mathbb{R}^n \times \mathbb{R}^{n\times p}$ is open.

For $F = 2\max\{1, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2/(n\|\boldsymbol{X}\boldsymbol{h}\|^2)\}$ in Lemma 7.2, (7.12) is satisfied so that (7.13)–(7.14) hold. In Ω, we thus have $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 \vee (\|\boldsymbol{X}\boldsymbol{h}\|^2/n) \le F_+F^2R_* \le 8\varphi^{-2}R_*$ and $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|/\sqrt{n} \le F_+^{1/2}\sigma + \sqrt{8}\varphi^{-1}{R_*}^{1/2} \le 3\sqrt{2}\varphi^{-1}R_*$. Furthermore, $\|\boldsymbol{w}_0\|_2^2 \le \varphi^{-1}/n$ in Ω_E thanks to $|\widehat{S}| \le \kappa n/2$ and the explicit expression for $\boldsymbol{w}_0$ in Proposition 4.2. In summary, we have in Ω,

$$\begin{aligned} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2 \vee \big(\|\boldsymbol{X}\boldsymbol{h}\|^2/n\big) &\le 8\varphi^{-2}R_*, \\ \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/n &\le 18\varphi^{-2}R_*, \\ \|\boldsymbol{w}_0\|^2 &\le \varphi^{-2}/n \end{aligned} \tag{B.2}$$

which replace (7.10)–(7.11) in the present context. By the deterministic inequality (7.29), in Ω we have $\widehat{\mathsf{df}} \le |\widehat{S}| \le \kappa n/2$ since $\widehat{\boldsymbol{H}}$ is rank at most $|\widehat{S}|$ with operator norm at most one, so that

$$I_\Omega(1 - \kappa/2)^2/8 \le \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*) + \Delta_n^a + \Delta_n^b + \Delta_n^c. \tag{B.3}$$

Let $(\boldsymbol{\varepsilon}, \boldsymbol{X})$, $(\boldsymbol{\varepsilon}, \widetilde{\boldsymbol{X}})$ both in Ω, let $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}$ and let $\boldsymbol{h}$, $\widetilde{\boldsymbol{h}}$, $\boldsymbol{f}$, $\widetilde{\boldsymbol{f}}$ be as in Lemma 7.1. Thanks to event Ω_E and the fact that $|\widehat{S}| \le \kappa n/2$, and similarly for $\widetilde{\boldsymbol{\beta}}$, we have $\varphi^2\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2 \le \|\boldsymbol{X}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2/(2n) + \|\widetilde{\boldsymbol{X}}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2/(2n)$. Thus, by (7.3),

$$n\varphi^2\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|^2 \le (\widetilde{\boldsymbol{h}} - \boldsymbol{h})^\top(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top\boldsymbol{\varepsilon} + (\boldsymbol{h} - \widetilde{\boldsymbol{h}})^\top(\boldsymbol{X}^\top\boldsymbol{X} - \widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}})(\boldsymbol{h} + \widetilde{\boldsymbol{h}})/2.$$

Summing this inequality with the first line in (7.2), we find

$$\begin{aligned} &n\varphi^2\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{h} - \widetilde{\boldsymbol{h}})\|^2 + \|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\|^2 \\ &\quad\le (\widetilde{\boldsymbol{h}} - \boldsymbol{h})^\top(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top\boldsymbol{\varepsilon} + (\boldsymbol{h} - \widetilde{\boldsymbol{h}})^\top(\boldsymbol{X}^\top\boldsymbol{X} - \widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}})(\boldsymbol{h} + \widetilde{\boldsymbol{h}})/2 \\ &\qquad + (\widetilde{\boldsymbol{h}} - \boldsymbol{h})^\top(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top\boldsymbol{f} + \boldsymbol{h}^\top(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top(\boldsymbol{f} - \widetilde{\boldsymbol{f}}). \end{aligned} \tag{B.4}$$

Thanks to the bounds in (B.2), this implies $\|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\| \le L\|(\boldsymbol{X} - \widetilde{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}$ if $\{(\boldsymbol{\varepsilon}, \boldsymbol{X}), (\boldsymbol{\varepsilon}, \widetilde{\boldsymbol{X}})\} \subset \Omega$, where $L = C_{27}(\gamma, \kappa){R_*}^{1/2}$.

For a given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$, we define $U_0 = \{z_0 \in \mathbb{R}^n : (\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0 + z_0\boldsymbol{a}_0^\top) \in \Omega\}$. In U_0, the function $f(z_0) = \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}$ is L-Lipschitz. By Kirszbraun's theorem, there exists a function $F : \mathbb{R}^n \to \mathbb{R}^n$ that is an extension of f, that is, $F(z_0) = f(z_0)$ for $z_0 \in U_0$, and such that F is L-Lipschitz in the whole $\mathbb{R}^n$. Note that both function F and f implicitly depend on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_0)$. Since Ω is open, U_0 is also open, and thus conditionally on $(\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon})$,

$$\nabla f(z_0) = \nabla F(z_0), \quad \text{for all } z_0 \in U_0. \tag{B.5}$$

(Without the openness of Ω established above, equality of the gradients would be unclear).

Since $F : \mathbb{R}^n \to \mathbb{R}^n$ is such that $F(z_0) = f(z_0)$ in Ω, by (B.3),

$$\begin{aligned}(1-\kappa/2)^2 I_\Omega &\le I_\Omega\big[\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(nR_*) + \Delta_n^a + \Delta_n^b + \Delta_n^c\big] \\ &= I_\Omega\big[\big\|F(z_0)\big\|^2/(nR_*) + \Delta_n^a + \Delta_n^b + \Delta_n^c\big].\end{aligned} \tag{B.6}$$

Taking conditional expectations and multiplying both sides by $\delta_1^2 \overset{\text{def}}{=} \mathbb{E}_0[\|\nabla F(z_0)\|_F^2]/\{\mathbb{E}_0[\|\nabla F(z_0)\|_F^2] + \mathbb{E}_0[\|F(z_0)\|^2]\}$, we find

$$\delta_1^2(1-\kappa/2)^2\mathbb{E}_0[I_\Omega] \le \mathbb{E}_0\big[\big\|\nabla F(z_0)\big\|_F^2/(nR_*)\big] + \mathbb{E}_0\big[I_\Omega\big(\Delta_n^a + \Delta_n^b + \Delta_n^c\big)\big],$$

due to $\delta_1^2\mathbb{E}_0[I_\Omega\|F(z_0)\|^2] \le \mathbb{E}_0[\|\nabla F(z_0)\|_F^2]$ for the first term and $\delta_1^2 \le 1$ for the second. Using $\delta_1^2 \le 1$ and $1 = I_\Omega + I_{\Omega^c}$,

$$\begin{aligned}\mathbb{E}\big[\delta_1^2\big](1-\kappa/2)^2 \le{}& \mathbb{E}\big[I_\Omega\big\|\nabla F(z_0)\big\|_F^2/(nR_*)\big] + \mathbb{E}\big[I_\Omega\big(\Delta_n^a + \Delta_n^b + \Delta_n^c\big)\big] \\ &+ \mathbb{P}\big[\Omega^c\big]\big(1 + L^2/R_*\big),\end{aligned}$$

where we used that $\|\nabla F(z_0)\|_F^2 \le n\|\nabla F(z_0)\|_{\text{op}}^2 \le nL^2$ in Ω^c since F is L-Lipschitz. We now prove that the three terms on the right-hand side converge to 0. For the third term, $L^2/R_* \le C_{28}(\gamma,\kappa)$ and $\mathbb{P}(\Omega^c) \to 0$ as Ω has probability approaching one. For the first term, since F is L-Lipschitz, $\|\nabla F(z_0)\|_F^2 \le nL^2$ almost surely so that the sequence of random variables $I_\Omega\|\nabla F(z_0)\|_F^2/(R_*n)$ is uniformly integrable. Thanks to uniform integrability, if we can prove $\|\nabla F(z_0)\|_F^2/(R_*n) \to^{\mathbb{P}} 0$, then $\mathbb{E}[I_\Omega\|\nabla F(z_0)\|_F^2/(R_*n)] \to 0$ holds. We use that $I_\Omega\nabla f(z_0) = I_\Omega\nabla F(z_0)$ by (B.5), and that in Ω the gradients of f with respect to z_0 are given in Proposition 4.2, so that by (B.2),

$$\begin{aligned}I_\Omega\big\|\nabla F(z_0)\big\|_F/(R_*n)^{1/2} &= I_\Omega\big\|\nabla f(z_0)\big\|_F/(R_*n)^{1/2} \\ &\le I_\Omega\big[\|\boldsymbol{w}_0\|\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\| + \|\boldsymbol{I}_n - \widehat{\boldsymbol{H}}\|_F\big|\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle\big|\big]/(R_*n)^{1/2} \\ &\le C_{29}(\gamma,\kappa)\big(n^{-1/2} + \big|\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle\big|/{R_*}^{1/2}\big),\end{aligned}$$

which converges to 0 in probability thanks to assumption $\langle \boldsymbol{a}_0, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle^2/R_* \overset{\mathbb{P}}{\to} 0$. Thanks to uniform integrability, this proves $\mathbb{E}[I_\Omega\|\nabla F(z_0)\|_F^2/(\sigma^2 n)] \to 0$. It remains to show $\mathbb{E}[I_\Omega(\Delta_n^a + \Delta_n^b + \Delta_n^c)] \to 0$. By definition of Δ_n^b in (7.27), thanks to Ω_χ and (B.2) we have $I_\Omega\Delta_n^b \le 18\varphi^{-2}(F_+ - 1) \le C_{30}(\gamma,\kappa)\sqrt{\log(n)/n}$. For Δ_n^a in (7.26), let $\boldsymbol{\Pi} : \mathbb{R}^n \to \mathbb{R}^n$ be the convex projection onto the Euclidean ball of radius $\sqrt{18\varphi^{-2}R_*}$, then $\boldsymbol{\Pi}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ in Ω by (B.2) so that

$$\begin{aligned}\mathbb{E}\big[I_\Omega\Delta_n^a\big] &= \mathbb{E}\big[I_\Omega\big\{(1 - \widehat{\mathsf{df}}/n) - \boldsymbol{\varepsilon}^\top\boldsymbol{\Pi}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/\big(n\sigma^2\big)\big\}^2\big]\sigma^2/R_* \\ &\le \mathbb{E}\big[\big\|\boldsymbol{\Pi}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\big\|^2\big]/\big(n^2R_*\big) + \sigma^2/(nR_*) \\ &\le 18\varphi^{-2}/n + 1/n\end{aligned} \tag{B.7}$$

by applying Proposition 2.1 to the function $\boldsymbol{\varepsilon} \mapsto \boldsymbol{\Pi}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, which is 1-Lipschitz as the composition of two 1-Lipschitz functions (cf. Proposition 7.3(i)). For Δ_n^c in (7.28), let $\boldsymbol{g}$, $\boldsymbol{a}_*$ by as in Lemma 7.6 and set $\boldsymbol{u}_* = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}_*$, $\boldsymbol{Q}_* = \boldsymbol{I}_p - \boldsymbol{u}_*\boldsymbol{a}_*^\top$ and note that $(\boldsymbol{g}, \boldsymbol{u}_*, \boldsymbol{Q}_*) = (z_0, \boldsymbol{u}_0, \boldsymbol{Q}_0)$ for $\boldsymbol{a}_0 = \boldsymbol{a}_*$. Let also $\boldsymbol{w}_*$ be the $\boldsymbol{w}_0$ from Proposition 4.2 for $\boldsymbol{a}_0 = \boldsymbol{a}_*$. As above for $\boldsymbol{a}_0$, for a fixed $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_*)$ the function $\boldsymbol{g} \mapsto \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is L-Lipschitz in $U_* = \{\boldsymbol{g} \in \mathbb{R}^n : (\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_* + \boldsymbol{g}\boldsymbol{a}_*^\top) \in \Omega\}$ by (B.4) for the value of L given after (B.4). Furthermore, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 \le 18n\varphi^{-2}R_*$ in Ω. By Kirszbraun's theorem, there exists an extension $F_* : \mathbb{R}^n \to \mathbb{R}^n$ implicitly depending on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_*)$ such that $F_*(\boldsymbol{g}) = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ in Ω

and $\|F_*(\boldsymbol{g})\|^2 \le 18n\varphi^{-2}R_*$ by composing the extension given by Kirszbraun's theorem by the convex projection onto the Euclidean ball of radius $(18n\varphi^{-2}R_*)^{1/2}$. By Proposition 2.1, with respect to $\boldsymbol{g}$ conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_*)$,

$$\mathbb{E}\big[I_\Omega\big((n-\widehat{\mathrm{df}})\langle \boldsymbol{a}_*, \boldsymbol{h}\rangle + \langle \boldsymbol{w}_* + \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle\big)^2\big] = \mathbb{E}\big[I_\Omega\big(\operatorname{div} F_*(\boldsymbol{g}) + \langle \boldsymbol{g}, F_*(\boldsymbol{g})\rangle\big)^2\big] \le 18n\varphi^{-2}R_* + nL^2.$$

For the value of L given after (B.4) and using the bound (B.2) to control $\langle \boldsymbol{w}_*, \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\rangle$ in Ω, this gives $\mathbb{E}[I_\Omega \Delta_n^c] \le C_{31}(\gamma,\kappa)n^{-1}$.

This proves $(1-\kappa/2)^2\mathbb{E}[\delta_1^2] \to 0$. Consequently, $\Xi_0 = z_0^\top F(z_0) - \operatorname{div} F(z_0)$ satisfies $|\sup_t |\mathbb{P}(\Xi_0/\|F(z_0)\| \le t) - \Phi(t)| \to 0$ by (2.8). Since $\xi_0 = z_0^\top f(z_0) - \operatorname{div} f(z_0)$ is equal to Ξ_0 on the event Ω because F is an extension of f, we have $|\mathbb{P}(\xi_0/\|f(z_0)\| \le t) - \mathbb{P}(\Xi_0/\|F(z_0)\| \le t)| \le 2\mathbb{P}(\Omega^c) \to 0$, so that $\sup_t |\mathbb{P}(\xi_0/\|f(z_0)\| \le t) - \Phi(t)| \to 0$ as well. The conclusion (3.35) is obtained by controlling the term $\boldsymbol{w}_0^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|$ by $\|\boldsymbol{w}_0\|$, which is bounded as in (B.2) in Ω.

It remains to show that (3.35) holds uniformly over all $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ and to derive the properties of $\overline{S}$. The proof of the relative volume bound on $\overline{S}$ and the lower bound on the cardinality of $\{j \in [p] : \boldsymbol{e}_j/\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\| \in \boldsymbol{\Sigma}^{1/2}\overline{S}\}$ is the same as in the proof of Theorem 3.7 given around (7.39), and for $\boldsymbol{a}_0 \in \boldsymbol{\Sigma}^{1/2}\overline{S}$ inequality (7.39) holds. For such $\boldsymbol{a}_0$, $\mathbb{E}[I_\Omega\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/R_*] \le 8\varphi^{-2}\mathbb{E}[\langle \boldsymbol{a}_0, \boldsymbol{h}\rangle^2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|^2] \le 8\varphi^{-2}C^*/a_p \to 0$ by (B.2) for the first inequality and (7.39) for the second. □

APPENDIX C: STRICT KKT CONDITIONS WITH PROBABILITY ONE FOR THE GROUP LASSO

LEMMA C.1. *Consider a design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and a response vector $\boldsymbol{y} \in \mathbb{R}^n$ for which the joint distribution of $(\boldsymbol{X}, \boldsymbol{y})$ admits a density with respect to the Lebesgue measure. Consider a partition of $\{1,\dots,p\}$ into groups $(G_1,\dots,G_K)$ and any minimizer*

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}\|^2 + \|\boldsymbol{b}\|_{\mathrm{GL}}, \qquad \|\boldsymbol{b}\|_{\mathrm{GL}} \overset{\mathrm{def}}{=} \sum_{k=1,\dots,K} \lambda_k \|\boldsymbol{b}_{G_k}\|_2$$

for some deterministic $\lambda_1,\dots,\lambda_K > 0$. There exists an open set $U \subset \mathbb{R}^{n\times(1+p)}$ such that $\mathbb{P}((\boldsymbol{y},\boldsymbol{X}) \in U) = 1$ and the KKT conditions are strict in $\{(\boldsymbol{y},\boldsymbol{X}) \in U\}$ in the sense that

$$\{(\boldsymbol{y},\boldsymbol{X}) \in U\} \subset \big\{\forall k = 1,\dots,K, \widehat{\boldsymbol{\beta}}_{G_k} = 0 \Rightarrow \|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_2 < n\lambda_k\big\}. \tag{C.1}$$

Finally, $\widehat{B} = \{k \in [K] : \|\widehat{\boldsymbol{\beta}}_{G_k}\| > 0\}$ is constant in a small neighborhood of any point in U.

PROOF OF LEMMA C.1. Consider a fixed $B \subset \{1,\dots,K\}$ and its complementary set B^c, and consider the group Lasso estimator $\widehat{\boldsymbol{\beta}}(B)$ with the additional constraint $\boldsymbol{b}_{G_k} = 0$ for every $k \in B^c$. Now consider a group $k \in B^c$. Since the joint distribution of $(\boldsymbol{X},\boldsymbol{y})$ has a density with respect to the Lebesgue measure, the conditional distribution of $\boldsymbol{X}_{G_k}$ given $(\boldsymbol{y}, (\boldsymbol{X}\boldsymbol{e}_j)_{j\notin G_k})$ also admits a density with respect to the Lebesgue measure. Conditionally, on $(\boldsymbol{y}, (\boldsymbol{X}\boldsymbol{e}_j)_{j\notin G_k})$, two cases may appear:

(i) If $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B) = 0$, the KKT condition for group G_k hold strictly since $\lambda_k \ne 0$.

(ii) If $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B) \ne 0$, the distribution of $\boldsymbol{X}_{G_k}$ given $(\boldsymbol{y}, (\boldsymbol{X}\boldsymbol{e}_j)_{j\notin G_k})$ and the distribution of $\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))$ given $(\boldsymbol{y}, (\boldsymbol{X}\boldsymbol{e}_j)_{j\notin G_k})$ both have a density with respect to the Lebesgue measure. The sphere of radius $n\lambda_k$ has measure 0 for any continuous distribution, hence

$$\mathbb{P}\big(\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2 \ne n\lambda_k | \boldsymbol{y}, (\boldsymbol{X}\boldsymbol{e}_j)_{j\notin G_k}\big) = 1.$$

Finally, the unconditional probability $\mathbb{P}(\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2 \neq n\lambda_k)$ is also one. Let $U = \bigcap_{B\subset\{1,\dots,K\}}\bigcap_{k\notin B}\{(\boldsymbol{y},\boldsymbol{X}) : \|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2 \neq n\lambda_k\}$. Then $\mathbb{P}((\boldsymbol{y},\boldsymbol{X}) \in U) = 1$ as a finite intersection of events of probability one and (C.1) holds. The set U is open as a finite intersection of open sets, since $\{(\boldsymbol{y},\boldsymbol{X}) : \|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2 \neq n\lambda_k\}$ is open by continuity of $(\boldsymbol{y},\boldsymbol{X}) \mapsto \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))$ by the claim following (7.2).

Next, to show that $\widehat{B}$ is constant in a neighborhood of every point in U, set $U_\delta = \bigcap_{B\subset\{1,\dots,K\}}\bigcap_{k\notin B}\{(\boldsymbol{y},\boldsymbol{X}) : |\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2/(n\lambda_k) - 1| > \delta\}$ for all $\delta > 0$. We have $U = \bigcup_{\delta>0} U_\delta$ and the set U_δ is open by continuity of $(\boldsymbol{y},\boldsymbol{X}) \mapsto |\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))\|_2/(n\lambda_k) - 1|$, which follows from the continuity of $(\boldsymbol{y},\boldsymbol{X}) \mapsto \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(B))$ by the claim following (7.2). For any $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}}) \in U$, there exists some $\delta > 0$ with $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}}) \in U_\delta$. Let $\overline{B} = \{k \in [K] : \|\overline{\boldsymbol{\beta}}_{G_k}\| > 0\}$. By continuity of $(\boldsymbol{y},\boldsymbol{X}) \mapsto \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ thanks to the claim following (7.2), there exists a neighborhood $\mathcal{N}$ of $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$ with $\mathcal{N} \subset U_\delta$ such that for all $(\boldsymbol{y},\boldsymbol{X}) \in \mathcal{N}$, $\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|/(n\lambda_k) < 1 - \delta/2$ for $k \notin \overline{B}$ and $\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|/(n\lambda_k) > 1 - \delta/2$ for $k \in \overline{B}$. Since $\mathcal{N} \subset U_\delta$, $\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|/(n\lambda_k) > 1 - \delta/2$ implies $\|\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|/(n\lambda_k) = 1$, so that $\widehat{B} = \overline{B}$ in $\mathcal{N}$. $\square$

PROPOSITION 4.2. *The following holds for almost every* $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}}) \in \mathbb{R}^{n\times(1+p)}$. *The set* $\overline{B} = \{k \in [K] : \|\overline{\boldsymbol{\beta}}_{G_k}\| > 0\}$ *of active groups is the same for all minimizers* $\overline{\boldsymbol{\beta}}$ *of* (3.33) *at* $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$ *and* $\widehat{B} = \overline{B}$ *for all* $(\boldsymbol{y},\boldsymbol{X})$ *in a sufficiently small neighborhood of* $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$. *If additionally* $\overline{\boldsymbol{X}}_{\overline{S}}^\top\overline{\boldsymbol{X}}_{\overline{S}}$ *is invertible where* $\overline{S} = \bigcup_{k\in\overline{B}} G_k$ *then the map* $(\boldsymbol{y},\boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ *is Lipschitz in a sufficiently small neighborhood of* $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$. *In this neighborhood, we have*

$$[\nabla\widehat{\boldsymbol{\beta}}(z_0)]_{\widehat{S}^c} = 0, \qquad [\nabla\widehat{\boldsymbol{\beta}}(z_0)]_{\widehat{S}}^\top = (\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}[(\boldsymbol{a}_0)_{\widehat{S}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top - \langle \boldsymbol{a}_0, \boldsymbol{h}\rangle \boldsymbol{X}_{\widehat{S}}^\top],$$

$\widehat{\boldsymbol{H}} = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}\boldsymbol{X}_{\widehat{S}}^\top$ *and* (3.9) *holds with* $\boldsymbol{w}_0 = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + \boldsymbol{M})^{-1}(\boldsymbol{a}_0)_{\widehat{S}}$.

PROOF OF PROPOSITION 4.2. By Lemma C.1, $\widehat{B}$ and $\widehat{S}$ are constant in a sufficiently small neighborhood of almost every $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$. The additional assumption that $\overline{\boldsymbol{X}}_{\overline{S}}^\top\overline{\boldsymbol{X}}_{\overline{S}}$ is invertible provides that $\boldsymbol{X}_{\overline{S}}^\top\boldsymbol{X}_{\overline{S}}$ is invertible by continuity of the smallest eigenvalue in a small enough compact neighborhood of $(\overline{\boldsymbol{y}},\overline{\boldsymbol{X}})$, and in this neighborhood $(\boldsymbol{y},\boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}$ is Lipschitz by the sentence following (7.4), and thus almost everywhere differentiable by Rademacher's theorem. The formulae for $\nabla\widehat{\boldsymbol{\beta}}(z_0)$, $\widehat{\boldsymbol{H}}$ and $\boldsymbol{w}_0$ involving the matrix $\boldsymbol{M}$ in (4.5) are then obtained by differentiating the KKT conditions restricted to $\widehat{S}$ in this neighborhood, that is, $\boldsymbol{X}_{G_k}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = n\lambda_k\widehat{\boldsymbol{\beta}}_{G_k}/\|\widehat{\boldsymbol{\beta}}_{G_k}\|$ for all $k \in \widehat{B}$. $\square$

APPENDIX D: PROOF OF THEOREM 2.3

PROOF OF THEOREM 2.3. With $\mathbb{E}[\|\overline{\boldsymbol{\mu}} + \overline{\boldsymbol{A}}^\top\boldsymbol{z}\|^2] = \|\overline{\boldsymbol{\mu}}\|^2 + \|\overline{\boldsymbol{A}}\|_F^2$ in mind, consider

$$\begin{aligned}\widehat{\operatorname{Var}[\xi]} &= \|f(\boldsymbol{z}) - (\overline{\boldsymbol{\mu}} + \overline{\boldsymbol{A}}^\top\boldsymbol{z})\|^2 + \operatorname{trace}[\{\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\}^2] \\ &\quad + 2(f(\boldsymbol{z}) - \overline{\boldsymbol{\mu}} - \overline{\boldsymbol{A}}^\top\boldsymbol{z})^\top(\overline{\boldsymbol{\mu}} + \overline{\boldsymbol{A}}^\top\boldsymbol{z}) + 2\operatorname{trace}[\{\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\}\overline{\boldsymbol{A}}] \\ &\quad + (\|\overline{\boldsymbol{\mu}} + \overline{\boldsymbol{A}}^\top\boldsymbol{z}\|^2 - \|\overline{\boldsymbol{\mu}}\|^2 - \|\overline{\boldsymbol{A}}\|_F^2) + \|\overline{\boldsymbol{\mu}}\|^2 + \|\overline{\boldsymbol{A}}\|_F^2 + \operatorname{trace}[\overline{\boldsymbol{A}}^2].\end{aligned}$$

By the triangle and Cauchy–Schwarz inequalities,

$$\begin{aligned}&\mathbb{E}[|\|f(\boldsymbol{z})\|^2 + \operatorname{trace}[\{\nabla f(\boldsymbol{z})\}^2] - \operatorname{Var}[\xi]|] \\ &\quad\le \mathbb{E}[\|f(\boldsymbol{z}) - (\overline{\boldsymbol{\mu}} + \overline{\boldsymbol{A}}^\top\boldsymbol{z})\|^2] + \mathbb{E}[\|\nabla f(\boldsymbol{z}) - \overline{\boldsymbol{A}}\|_F^2]\end{aligned}$$

$$+2\{\mathbb{E}[\|f(z)-(\overline{\boldsymbol{\mu}}+\overline{\boldsymbol{A}}^\top z)\|^2]+\mathbb{E}[\|\nabla f(z)-\overline{\boldsymbol{A}}\|_F^2]\}^{1/2}\{\|\overline{\boldsymbol{\mu}}\|^2+2\|\overline{\boldsymbol{A}}\|_F^2\}^{1/2}$$
$$+\mathbb{E}[|\|\overline{\boldsymbol{\mu}}+\overline{\boldsymbol{A}}^\top z\|^2-\|\overline{\boldsymbol{\mu}}\|^2-\|\overline{\boldsymbol{A}}\|_F^2|]+|\|\overline{\boldsymbol{\mu}}\|^2+\|\overline{\boldsymbol{A}}\|_F^2+\operatorname{trace}(\overline{\boldsymbol{A}}^2)-\operatorname{Var}[\xi]|.$$

We have $\mathbb{E}[\|f(z)-(\overline{\boldsymbol{\mu}}+\overline{\boldsymbol{A}}^\top z)\|^2]\le\mathbb{E}[\|\nabla f(z)-\overline{\boldsymbol{A}}\|_F^2]\le\overline{\overline{\epsilon}}_{1,2}^2\operatorname{Var}[\xi]/2$ by the Gaussian Poincaré inequality, $\mathbb{E}[|\|\overline{\boldsymbol{\mu}}+\overline{\boldsymbol{A}}^\top z\|^2-\|\overline{\boldsymbol{\mu}}\|^2-\|\overline{\boldsymbol{A}}\|_F^2|^2]=\|\overline{\boldsymbol{A}}\overline{\boldsymbol{\mu}}\|^2+2\operatorname{trace}\{(\overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top)^2\}\le\|\overline{\boldsymbol{A}}\|_{\mathrm{op}}^2C_0^2\operatorname{Var}[\xi]$, and $0\le 1-\{\|\overline{\boldsymbol{\mu}}\|^2+\|\overline{\boldsymbol{A}}\|_F^2+\operatorname{trace}(\overline{\boldsymbol{A}}^2)\}/\operatorname{Var}[\xi]\le\overline{\overline{\epsilon}}_{1,2}^2$ as in (2.11). Thus, (2.14) holds and the conclusions follow. □

Funding. P.C. Bellec's Research was partially supported by the NSF Grants DMS-1811976 and DMS-1945428.

C.-H. Zhang's research was partially supported by the NSF Grants DMS-1721495, IIS-1741390, CCF-1934924, DMS-2052949 and DMS-2210850.

REFERENCES

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

[2] BELLEC, P. and TSYBAKOV, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern Problems of Stochastic Analysis and Statistics. Springer Proc. Math. Stat.* **208** 315–333. Springer, Cham. MR3747672 https://doi.org/10.1007/978-3-319-65313-6_13

[3] BELLEC, P. C. (2018). The noise barrier and the large signal bias of the lasso and other convex estimators. Available at: arXiv:1804.01230, https://arxiv.org/pdf/1804.01230.pdf.

[4] BELLEC, P. C. (2020). Out-of-sample error estimate for robust m-estimators with convex penalty. arXiv preprint. Available at arXiv:2008.11840.

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

[6] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Conference on Learning Theory* 1912–1947. PMLR.

[7] BELLEC, P. C. and ZHANG, C.-H. (2021). Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *Ann. Statist.* **49** 1864–1903. MR4319234 https://doi.org/10.1214/20-aos2005

[8] BELLEC, P. C. and ZHANG, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli* **28** 713–743. MR4389062 https://doi.org/10.3150/21-BEJ1348

[9] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect.* **28** 29–50.

[10] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

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

[12] BLANCHARD, J. D., CARTIS, C. and TANNER, J. (2011). Compressed sensing: How sharp is the restricted isometry property? *SIAM Rev.* **53** 105–125. MR2785881 https://doi.org/10.1137/090748160

[13] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726 https://doi.org/10.1016/j.acha.2009.04.002

[14] BOGACHEV, V. I. (1998). *Gaussian Measures. Mathematical Surveys and Monographs* **62**. Amer. Math. Soc., Providence, RI. MR1642391 https://doi.org/10.1090/surv/062

[15] BU, Z., KLUSOWSKI, J., RUSH, C. and SU, W. (2019). Algorithmic analysis and statistical estimation of slope via approximate message passing. In *Advances in Neural Information Processing Systems* 9361–9371.

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

[17] CELENTANO, M. and MONTANARI, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Statist.* **50** 170–196. MR4382013 https://doi.org/10.1214/21-aos2100

[18] CELENTANO, M., MONTANARI, A. and WEI, Y. (2020). The lasso with general gaussian designs with applications to hypothesis testing. arXiv preprint. Available at arXiv:2007.13716.

[19] CHATTERJEE, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields* **143** 1–40. MR2449121 https://doi.org/10.1007/s00440-007-0118-6

[20] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, *Vol. I* 317–366. North-Holland, Amsterdam. MR1863696 https://doi.org/10.1016/S1874-5849(01)80010-3

[21] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[22] EDELMAN, A. (1988). Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.* **9** 543–560. MR0964668 https://doi.org/10.1137/0609045

[23] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[24] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[25] FENG, L. and ZHANG, C.-H. (2019). Sorted concave penalized regression. *Ann. Statist.* **47** 3069–3098. MR4025735 https://doi.org/10.1214/18-AOS1759

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

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

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

[29] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

[30] LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.* **46** 611–641. MR3782379 https://doi.org/10.1214/17-AOS1562

[31] LEI, L., BICKEL, P. J. and EL KAROUI, N. (2018). Asymptotics for high dimensional regression M-estimates: Fixed design results. *Probab. Theory Related Fields* **172** 983–1079. MR3877551 https://doi.org/10.1007/s00440-017-0824-7

[32] LOUREIRO, B., GERBELOT, C., CUI, H., GOLDT, S., KRZAKALA, F., MEZARD, M. and ZDEBOROVÁ, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Adv. Neural Inf. Process. Syst.* **34** 18137–18151.

[33] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[34] MOHAMED, N. (2020). Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. arXiv preprint. Available at arXiv:2008.12236.

[35] NICULESCU, C. P. and PERSSON, L.-E. (2006). *Convex Functions and Their Applications. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC* **23**. Springer, New York. MR2178902 https://doi.org/10.1007/0-387-31077-0

[36] SILVERSTEIN, J. W. (1985). The smallest eigenvalue of a large-dimensional Wishart matrix. *Ann. Probab.* **13** 1364–1368. MR0806232

[37] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[38] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166 https://doi.org/10.1093/biomet/ass043

[39] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466

[40] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[41] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses. Volume I: Basic Examples. Ergebnisse der Mathematik und Ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics*] **54**. Springer, Berlin. MR2731561 https://doi.org/10.1007/978-3-642-15202-3

[42] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems* 3420–3428.

[43] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720

[44] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375 https://doi.org/10.1214/13-EJS815

[45] VAITER, S., DELEDALLE, C., PEYRÉ, G., FADILI, J. and DOSSAL, C. (2012). The degrees of freedom of the group lasso. arXiv preprint. Available at arXiv:1205.1481.

[46] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[47] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[48] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

[49] ZHANG, C.-H. (2011). Statistical inference for high-dimensional data. *Math. Forsch. Oberwolfach* **48** 28–31.

[50] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

[51] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026