DE-BIASING CONVEX REGULARIZED ESTIMATORS AND INTERVAL ESTIMATION IN LINEAR MODELS

BY PIERRE C BELLÉC AND CUN-HUI ZHANG

New upper bounds are developed for the $L_2$ distance between $\xi/\sqrt{\text{Var}[\xi]}^{1/2}$ and linear and quadratic functions of $z \sim N(0, I_n)$ for random variables of the form $\xi = z^T f(z) + \text{div} f(z)$. 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 de-biased estimators in linear regression with correlated design matrices $X$ that the "de-biased" estimate can be used for inference about $\beta_0$. By rotational invariance, the results vastly broaden the scope of applicability of de-biasing methodologies to obtain confidence intervals in high-dimensions. In the absence of strong convexity for $p > n$, asymptotic normality of the de-biased 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

$$y = X\beta + \varepsilon$$

with an unknown coefficient vector $\beta \in \mathbb{R}^p$, a Gaussian noise vector $\varepsilon \sim N(0, \sigma^2 I_n)$, and a Gaussian design matrix $X \in \mathbb{R}^{n \times p}$ with iid $N(0, \Sigma)$ rows independent of $\varepsilon$. We assume throughout the sequel that $\Sigma$ is invertible. The paper develops confidence intervals for $\theta = \langle a_0, \beta \rangle$ from a given regularized initial estimator $\hat{\beta} \in \mathbb{R}^p$, using a technique referred to as de-biasing: a correction to the initial estimate $\langle a_0, \beta \rangle$ in the direction $a_0$ is constructed so that the "de-biased" estimate can be used for inference about $\theta = \langle a_0, \beta \rangle$.

1.1. Regularization induces bias. If $X^TX$ is invertible, the unregulated least-squares estimate $\hat{\beta}^{ls} = (X^TX)^{-1}X^Ty$ is unbiased, that is, $\mathbb{E}[\hat{\beta}^{ls} - \beta | X] = 0$. On the other hand, if the square loss is regularized with an additive penalty,

$$\tilde{\beta} = \arg\min_{b \in \mathbb{R}^p} \| y - Xb \|^2/(2n) + g(b)$$

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

For ridge regression $\beta = (X^TX + n\mu I_p)^{-1}X^Ty$, this bias can be quantified explicitly when $\Sigma = I_p$ as a shrinkage to the origin. Let $\sum_{i=1}^{r} u_i s_i v_i^\top$ be the SVD of $X$ with $s_i > 0$ and $r = \min(n, p)$. By rotational invariance, $v_i$ is independent of $s_i$ and uniformly distributed in the unit sphere in $\mathbb{R}^p$. Thus, with $G_\gamma$ being the Marchenko-Pastur law,

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E} \left[ \sum_{i=1}^{r} \frac{s_i^2 v_i^\top v_i}{s_i^2 + n\mu} \beta \right] \approx \beta \int \frac{(r/p)x}{x + (r/p)}G_\gamma(dx) \text{ as } \frac{p}{n} \to 0.$$
The Lasso penalty $g(b) = \lambda \|b\|_1$ also introduces bias. For example, for deterministic orthonormal designs, the Lasso estimator of the coefficient $\beta_j$ is the soft-thresholding of $N(\beta_j, \sigma^2/n)$ which is again biased toward the origin. For Gaussian designs with $\Sigma = I_p$ and in an average sense, the Lasso is approximately the soft-thresholding of $N(\beta_j, \tau^2_n/n)$ with certain $\tau_n \geq \sigma$ under proper conditions [1]. Thus, with $s_1 = \# \{ j : |\beta_j| > \lambda \}$, the squared bias of the Lasso, $\| \hat{\beta} - E[\beta] \|_2^2$, is expected to have no smaller order than the lower bound $s_1 \lambda^2$ for its $\ell_2$ risk [3, Theorem 3.1]. Alternative approaches were proposed to remove or reduce the bias of the Lasso for strong signals, e.g., by using concave penalty functions (e.g., SCAD [25], MCP [49]) or iterated hard thresholding algorithms [14]. These approaches yield an error term of the order $(|\beta|_0 - s_1')^2 + s_1' \sigma^2/n$ where $s_1' = \{ j = 1, ..., p : |\beta_j| > c \lambda \}$ for some constant $c > 0$ [26, 35], alleviating the bias of the Lasso for large coefficients at typical penalty levels $\lambda > \sigma/n^{1/2}$.

De-biasing the Lasso, asymptotic normality and confidence intervals. If the goal is the estimation of a single scalar parameter $\theta = \langle \alpha_0, \beta \rangle$ in a predetermined direction $\alpha_0$ instead of the full vector $\beta \in \mathbb{R}^p$, it is possible to correct the bias of the Lasso and to construct confidence intervals for $\theta$: there is already a vast literature on asymptotic normality of de-biased estimates in sparse linear regression for the Lasso [52, 47, 27, 28, 10, 29, 34, 8, among others]. In this literature $\alpha_0$ is usually the $j$-th canonical basis vector and $\beta_j$ the scalar parameter of interest. Given the Lasso $\hat{\beta}$ as an initial estimator of $\beta$, the idea is to add a de-biasing term to achieve asymptotic normality which then yields confidence intervals for $\theta = \langle \alpha_0, \beta \rangle$. If $s_0 = |\beta|_0$ in (1.1), several de-biased estimators have been proposed and their asymptotic normality hold under certain rate conditions on $s_0, n, p$. The earliest works on this topic [52, 47, 27, 10] 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 de-biasing constructions in these papers are all first order equivalent to each other, and under normalization $\|\Sigma^{-1/2}a_0\|_2 = 1$ to

\[
\begin{align*}
\hat{\theta} &= \langle \alpha_0, \hat{\beta} \rangle + \|z_0\|_2^{-2}z_0^\top(y - Xu) + \sqrt{n}(|\hat{\theta} - \theta| - \epsilon) + O_p(R_n),
\end{align*}
\]

where $u_0 = \Sigma^{-1}a_0/\langle \alpha_0, \Sigma^{-1}a_0 \rangle$ and $z_0 = Xu_0 \sim N(0, I_n)$. While these works do not assume $\Sigma$ known and construct an estimated score vector $\hat{z}$ for $z_0$, the impact of using $\hat{z}$ can be absorbed into the remainder in (1.3) with $R_n = \sigma s_0 \log(p)/\sqrt{n}$. The direction $u_0$ and the de-biasing correction in (1.3) have a natural semi-parametric interpretation [50]. Viewing $\theta : \mathbb{R}^p \to \mathbb{R}$ as the function $\theta(\beta) = \langle \alpha_0, \beta \rangle$, the Fischer information for the estimation of $\theta(\beta)$ in (1.1) is $F_\theta = 1/(\sigma^2 \langle \alpha_0, \Sigma^{-1}a_0 \rangle)$, and the direction $u_0$ above is the only $u \in \mathbb{R}^p$ with

\[
\langle \nabla \theta(\beta), u \rangle = \langle a_0, u \rangle = 1
\]

such that $F_\theta$ is also the Fischer information in the one-dimensional submodel $\{ \beta + tu, t \in \mathbb{R} \}$. For this reason the line $\{ \beta + tu_0, t \in \mathbb{R} \}$ is referred to as the least-favorable one-dimensional submodel for the estimation of $\theta$. The normalization (1.4) ensures that $\theta(\beta + tu) = \theta(\beta) + t$ and $\hat{\theta} = \theta(\hat{\beta} + \hat{t}u_0)$ with $\hat{t} = \|z_0\|_2^{-2}z_0^\top(y - Xu)$, so that (1.3) replaces the initial $\hat{\beta}$ with its one-step correction $\hat{\beta} + \hat{t}u_0$, where $\hat{t}$ maximizes the likelihood in the least-favorable submodel. We refer to [11] for a systematic study of this semi-parametric perspective.

If $s_0 \log(p)/\sqrt{n} \to +\infty$ and $\Sigma$ is unknown with bounded spectrum, the minimax estimation error of the form $\sqrt{n}(\hat{\theta} - \theta)$ diverges for any estimator $\hat{\theta}$ [17]. 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 $\Sigma$. However, if $z_0$ 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 $s_0 \leq n^{2/3}$ up to logarithmic factors, cf. [8, Corollary 3.3]. Similarly, [29, Theorem 3.8] provides (1.3) with $a_0 = e_j \in \mathbb{R}^p$ a canonical basis vector and $R'_n = \log(p) \sqrt{s_0/n} \max_j |\Sigma^{-1}e_j|_1$. 

imsart-aos ver. 2020/05/28 file: asymptotic-normality.tex date: October 18, 2022
Already in the regime $\sqrt{n} \ll s_0 \ll n^{2/3}$, the arguments of [29, 8] differ significantly from the $\ell_1-\ell_\infty$ Hölder inequality argument of [52, 47, 27, 10]: while these earlier works prove asymptotic normality with a remainder term of order $O(p \log(p)/\sqrt{n})$, [29, 8] analyze explicitly the smaller order terms hidden in this $O(p \log(p)/\sqrt{n})$ remainder.

For $s_0 \gg n^{2/3}$, the de-biasing correction in (1.3) needs to be modified:

\begin{equation}
(1.5) \quad \tilde{\theta} = (a_0, \beta) + (n - |\tilde{S}|)^{-1} z_0^\top (y - X \beta), \quad \sqrt{n}(\tilde{\theta} - \theta) = \sqrt{n}||z_0||^2 z_0^\top \epsilon + O_p(R'_{n})
\end{equation}

with $\tilde{S} = \{j \in [p]: \tilde{\beta}_j \neq 0\}$ and $R'_{n} = \sigma(s_0 \log(p/s_0)/n)^{1/2}$, cf. [8, Theorem 3.1]. For $\|\Sigma^{-1/2}a_0\| = 1$ the difference from (1.3) is the replacement of $||z_0||^2$ by $\approx n^{-1}$ in the de-biasing correction with $(n - |\tilde{S}|)^{-1}$ to amplify it by a factor $(1 - \tilde{S}/n)^{-1}$. This modification is required as soon as $s_0 \gg n^{2/3}$ up to logarithmic factors [8, Section 3]. These asymptotic results for $s_0 \gg \sqrt{n}$ are amenable to the lack of knowledge of $\Sigma$; in this case estimation of $z_0$ is possible when $\Sigma^{-1}a_0$ is sufficiently sparse, see [29] if the direction of interest $a_0$ is canonical basis vector and [8, Section 2.2] for arbitrary direction $a_0$. These results [29, 8] for $s_0 \gg \sqrt{n}$ and correlated $\Sigma$ are so far restricted to random Gaussian designs.

**Inflated asymptotic variance for non-vanishing prediction error.** In the results discussed so far for the Lasso, $s_0 \log(p/s_0)/n \rightarrow 0$ or stronger conditions are required for asymptotic normality, and the asymptotic variance of $\sqrt{n}(\tilde{\theta} - \theta)$ is $\sigma^2$. The condition $s_0 \log(p/s_0)/n \rightarrow 0$ implies the consistency of the Lasso in prediction and estimation thanks to error bounds of the form $\|\Sigma^{1/2}(\beta - \beta)\|_2^2 \lesssim s_0 \log(p/s_0)/n$ [51, 39, 9, 8]. It turns out that the asymptotic variance of $\sqrt{n}(\tilde{\theta} - \theta)$ is larger than $\sigma^2$ if $\|\Sigma^{1/2}(\beta - \beta)\|_2^2$ does not vanish; this is the situation studied in the present work. The literature on asymptotic normality of de-biased estimates in the regime

\begin{equation}
(1.6) \quad p/n \rightarrow \gamma \in (0, +\infty), \quad s_0/n \rightarrow \kappa \in (0, 1)
\end{equation}

for constants $\gamma, \kappa > 0$ is more scarce. In this regime where $p, n$ and $s_0$ are all of the same order, [28, 34] provide asymptotic normality results for the de-biased Lasso (1.5) in the estimation of $\beta_j$ (canonical $a_0 = e_j$) in the isotropic Gaussian design. In these works, the asymptotic variance of $\sqrt{n}(\tilde{\theta} - \theta)$ equals a constant $\tau^2$ satisfying the system of two nonlinear equations in [1] and [34, Proposition 3.1,Theorem 3.1]. The constant $\tau^2$ is related to the residual sum of squares [34, Corollary 4.1] and out-of-sample error [34, Theorem 3.2] as in

\begin{equation}
(1 - |\tilde{S}|/n)^{-2}\|y - X \beta\|_2^2/n \rightarrow^P \tau^2, \quad \sigma^2 + \|\Sigma^{1/2}(\beta - \beta)\|_2^2 \rightarrow^P \tau^2
\end{equation}

where $\rightarrow^P$ denotes convergence in probability. These results for $\Sigma = I_p$ highlight that the asymptotic variance is strictly larger than $\sigma^2$ 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 $\sigma^2$ for all permutation-invariant penalty functions [18, Proposition 4.3].

In this regime where $n$ and $p$ are of the same order, [24, 22] proved asymptotic normality and characterized the variance for unregularized $M$-estimators. For $M$-estimators a de-biasing 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 [24, Lemma 1]. However, this rotational invariance is lost in the presence of a penalty such as the $\ell_1$-norm. New techniques are called for to analyse the asymptotic behavior, in the regime (1.6) and under correlated designs, of estimators that are not rotational invariant. More recently, the Approximate Message Passing techniques used in [28, 22] were used to obtain similar results in logistic regression [41]; 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 de-biased estimates in the regime (1.6) for correlated designs in the presence of a penalty not depending on $\Sigma$ (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, 34, 18, 16] and that the empirical distribution of the true $\{\sqrt{n}\beta_j, j \leq p\}$ converges to some prior distribution. A second goal of the present paper is to show that asymptotic normality of de-biased 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 $\{\sqrt{n}\beta_j, j \leq p\}$.

1.2. A general construction of de-biased estimators. This section describes a general approach to systematically construct de-biased estimates in the linear model (1.1) where $X$ has iid $N(0, \Sigma)$ rows. Our goal is to construct confidence intervals for the one-dimensional parameter $\theta = \langle a_0, \beta \rangle$. Consider an initial estimator $\hat{\beta}$, viewed as a function of $(y, X)$, i.e., $\hat{\beta} : \mathbb{R}^{n \times (1 + p)} \to \mathbb{R}^p$ and assume that this function $\hat{\beta}$ is Fréchet\(^1\) differentiable. For a given observed data $(y, X)$ from the linear model (1.1) and a $\hat{\beta}$ Fréchet differentiable at $(y, X)$, there exist uniquely matrices $\hat{H} \in \mathbb{R}^{n \times n}$ and $\hat{G} \in \mathbb{R}^{n \times p}$ such that

\[
(1.7) \quad X\hat{\beta}(y + \eta, X) - X\hat{\beta}(y, X) = \hat{H}^\top \eta + o(\|\eta\|), \quad \hat{\beta}(y, X) = \eta a_0^\top - \hat{\beta}(y, X) = \hat{G}^\top \eta + o(\|\eta\|),
\]

for all $\eta \in \mathbb{R}^n$. With $X = (x_{ij})_{i \in [n], j \in [p]}$, if the partial derivatives of $\hat{\beta}(y, X)$ at the observed data $(y, X)$ are $(\partial / \partial x_{ij})\hat{\beta}(y, X)$ and $(\partial / \partial y_i)\hat{\beta}(y, X)$ then (1.7) implies $\hat{H}^\top e_i = X(\partial / \partial y_i)\hat{\beta}(y, X)$ and $\hat{G}^\top e_i = \sum_{j=1}^p (a_0, e_j)(\partial / \partial x_{ij})\hat{\beta}(y, X)$ for canonical basis vectors $e_i \in \mathbb{R}^n$ and $e_j \in \mathbb{R}^p$. The derivatives of $\hat{\beta}$ and the matrices $\hat{H}$ and $\hat{G}$ can be computed by only looking at the observed data $(y, X)$, for instance by finite difference schemes.

Next, consider the function $\phi$ defined as

\[
\phi : \mathbb{R}^{n \times (1 + p)} \to \mathbb{R}^n, \quad (y, X) \mapsto \phi(y, X) = X\hat{\beta}(y, X) - y.
\]

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

\[
(1.8) \quad \phi(y + \eta, X) - \phi(y, X) = \left[(\hat{H} - I_n)^\top \eta + o(\|\eta\|)\right],
\]

\[
\phi(y, X + \eta a_0^\top) - \phi(y, X) = \left[(a_0, \hat{\beta})I_n + \hat{G}X^\top\right]^\top \eta + o(\|\eta\|).
\]

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

\[
\sum_{j=1}^p (a_0, e_j)(\partial / \partial x_{ij})\phi(y, X) = \left[(a_0, \hat{\beta})I_n + \hat{G}X^\top\right]^\top e_i
\]

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

Observe that the arguments $(y, X)$ of $\phi$ are centered and jointly normal random variables and their correlations are computed explicitly, e.g., $\mathbb{E}[x_{ij} y_l] = e_j^\top \Sigma \beta I_{i(l)}$, with basis vectors $e_j \in \mathbb{R}^p$. One version of Stein’s formula, also known as Gaussian integration by parts, is $\mathbb{E}[Gh(Z_1, ..., Z_q)] = \sum_{k=1}^q \mathbb{E}[GZ_k]\mathbb{E}[\partial h(Z_1, ..., Z_q)]$ provided that the function $h(z_1, ..., z_q)$ is differentiable and that $G, Z_1, ..., Z_q$ are centered jointly normal random variables, provided the existence of the expectations [42, Appendix A.4]. We leverage this version of Stein’s formula to obtain an unbiased estimating equation involving only one unknown parameter, the scalar $\theta = \langle a_0, \beta \rangle$ of interest. For $G_i = e_i^\top X \Sigma^{-1} a_0$ we find

\[\text{\textsuperscript{1}Although the Fréchet derivative is the usual definition of derivative in finite dimension, we write Fréchet to emphasize that the derivative is linear. Linearity may fail for weaker notions such as Gateaux differentiability.}\]
\( \mathbb{E}[G_i y_k] = \mathbb{E}[G_i x_{kj}] = 0 \) if \( i \neq k \) while \( \mathbb{E}[G_i x_{ij}] = (a_0, e_j) \) and \( \mathbb{E}[G_i y_i] = (a_0, \beta) \) so that by reading the partial derivatives in (1.8),

\[
\mathbb{E} \left[ G_i \phi_i(y, X) \right] = (a_0, \beta) \mathbb{E} \left[ \frac{\partial \phi_i}{\partial y_i}(y, X) \right] + \sum_{j=1}^p (a_0, e_j) \mathbb{E} \left[ \frac{\partial \phi_i}{\partial x_{ij}}(y, X) \right]
\]

(1.9)

\[
= \mathbb{E} \left[ (a_0, \beta)(\mathbb{H} - 1) \right] + \mathbb{E} \left[ (a_0, \beta) + e_i^T \mathbb{G} X^T e_i \right].
\]

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

\[
\mathbb{E} \left[ X \Sigma^{-1} a_0, X \hat{\beta} - y \right] = \mathbb{E} \left[ -\langle a_0, \beta \rangle \text{trace}[I_n - \mathbb{H}] + (a_0, \hat{\beta})n + \text{trace}[X^T \mathbb{G}] \right].
\]

To transform this equation into a form representative of the results of the paper, define the scalars \( \hat{df} \) and \( \hat{A} \) by

\[
\hat{df} = \text{trace}[\mathbb{H}], \quad \hat{A} = \text{trace}[X^T \mathbb{G}] + (a_0, \hat{\beta})\hat{df}.
\]

(1.10)

The notation \( \hat{df} \) underlines that \( \text{trace}[\mathbb{H}] \) has the interpretation of degrees-of-freedom of the estimator \( \beta \) in Stein’s Unbiased Risk Estimate (SURE) [38]: regarding \( \hat{\mu} = X \hat{\beta} \) as an estimate of \( \mu = X \beta \) in the Gaussian sequence model with observation \( y = \mu + \epsilon \), the quantity \( \text{SURE} = \|y - \hat{\mu}\|^2 + 2\sigma^2 \hat{df} - \sigma^2 n \) is an unbiased estimate of the in-sample error \( \|\hat{\mu} - \mu\|^2 = \|X(\beta - \beta')\|^2 \). With this notation, we obtain the unbiased estimating equation

\[
0 = \mathbb{E} \left[ \langle X \Sigma^{-1} a_0, y - X \beta \rangle \right] + (n - \hat{df})(\langle a_0, \hat{\beta} \rangle - \theta) + \hat{A}
\]

(1.11)

where the only unobserved quantity inside the expectation is \( \theta = \langle a_0, \beta \rangle \), the scalar parameter we wish to estimate. In the above application of Stein’s formula, \( G_i = e_i^T X \Sigma^{-1} a_0 \) was chosen on purpose so that \( \beta \) appears in (1.9) only through \( \langle a_0, \beta \rangle \) thanks to \( \mathbb{E}[G_i y_i] = (a_0, \beta) \). Note that replacing \( G_i \) in (1.9) by \( e_i^T X u \) for any \( u \in \mathbb{R}^p \) not proportional to \( \Sigma^{-1} a_0 \) brings a scalar projection of \( \beta \) different from \( (a_0, \beta) \): this shows the unique role of the random vector \( X \Sigma^{-1} a_0 \) to derive an unbiased estimating equation for \( \theta = \langle a_0, \beta \rangle \).

It is notable that the direction \( \Sigma^{-1} a_0 \) coincides with the least-favorable direction described around (1.4). Equation (1.11) is obtained for an arbitrary initial estimator \( \beta \) provided that its derivatives with respect to \( y, X \) exist and the integrability conditions hold to ensure existence of the expectations involved. From (1.11), the method of moments suggests to estimate \( \theta \) with \( \hat{\theta} = (a_0, \hat{\beta}) + (n - \hat{df})(\langle X \Sigma^{-1} a_0, y - X \beta \rangle + \hat{A}) \) which resembles (1.5) for the Lasso if \( \hat{df} = |\hat{S}| \) and \( X \Sigma^{-1} a_0 = 0 \) under the normalization \( (a_0, \Sigma^{-1} a_0) = 1 \).

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 \( g(b) = \mu \|b\|^2 \) for some \( \mu > 0 \), \( \beta(y, X) = (X^T X + \mu I_p)^{-1} X^T y \) and

\[
\hat{H}^T = X \Sigma^{-1} a_0, y - X \beta \]

(1.12)

\[
\hat{G}^T = (X^T X + \mu I_p)^{-1} [a_0(y - X \beta)^T - X^T (a_0, \beta)].
\]

Indeed, the derivatives of \( \beta(y, X) \) exist as it is the composition of elementary differentiable functions. Differentiation with respect to \( y \) is straightforward as \( \beta \) is linear in \( y \), while in order to compute \( \hat{G} \) we proceed by setting \( b(t) = \beta(y, X(t)) \) with \( X(t) = X + \eta a_0 \). Differentiation of the KKT conditions \( X(t)^T (y - X(t)b(t)) = \eta \mu b(t) \) at \( t = 0 \) provides the directional derivative \( (df/dt)|_{t=0} = \hat{G}^T \eta \). This gives (1.12). It follows from (1.12) that \( \hat{df} = \text{trace}[X(X^T X + \mu I_p)^{-1}]^T \) and \( \hat{A} = (y - X \beta)^T X(X^T X + \mu I_p)^{-1} a_0 \) for the quantities in (1.10) (for \( \hat{A} \), note the fortuitous cancellation of the term \( (a_0, \beta) \hat{df} \)). For the Lasso similar differentiability formulae are derived in [8]. It is however, unclear how to obtain closed form formulae for the derivatives of \( \beta \) 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
\begin{equation}
(1.13) \quad u_0 = \Sigma^{-1}a_0/\langle a_0, \Sigma^{-1}a_0 \rangle, \quad z_0 = Xu_0, \quad Q_0 = I_{p \times p} - u_0a_0^\top.
\end{equation}

The normalizing constant in $u_0$ is such that $\langle a_0, u_0 \rangle = 1$ holds so that the expression (1.13) for $u_0$ coincides with the direction of the least-favorable submodel discussed around (1.4). The vector $z_0$ is independent of $XQ_0$ by construction as $(z_0, XQ_0)$ are jointly normal and uncorrelated. This follows by noting that $X \Sigma^{-1/2}$ has iid $N(0, 1)$ entries and
\begin{equation}
(1.14) \quad X = XQ_0 + z_0a_0^\top \quad \text{with} \quad z_0 \sim N(0, \|\Sigma^{-1/2}a_0\|^2I_n) \quad \text{independent of} \quad XQ_0.
\end{equation}

For brevity, we assume in the sequel and without loss of generality that the direction of interest $a_0$ is normalized such that
\begin{equation}
(1.15) \quad \|\Sigma^{-1/2}a_0\|^2 = \langle a_0, \Sigma^{-1}a_0 \rangle = 1.
\end{equation}

By definition of $u_0$ and $z_0$, the normalization (1.15) gives $z_0 \sim N(0, I_n)$. Conditionally on $(XQ_0, \varepsilon)$, define the function $f(xQ_0, \varepsilon) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by
\begin{equation}
(1.16) \quad f(xQ_0, \varepsilon)(z_0) = X\beta - y.
\end{equation}

By (1.14) and the independence of $\varepsilon$ and $X$, the conditional expectation given $(XQ_0, \varepsilon)$ can be written as integrals against the Gaussian measure of $z_0$, e.g.,
\begin{equation}
\mathbb{E}\left[z_0^\top f(xQ_0, \varepsilon)(z_0) \mid (XQ_0, \varepsilon) \right] = \int \left(z^\top f(xQ_0, \varepsilon)(z)\right) e^{-\|z\|^2/2(2\pi)^{-n/2}} d\varepsilon
\end{equation}

since $z_0 \sim N(0, I_n)$. As we argue conditionally on $(XQ_0, \varepsilon)$, we omit the dependence on $(XQ_0, \varepsilon)$ and write simply $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Since $y = \varepsilon + X\beta$ and $X = XQ_0 + z_0a_0^\top$,
\begin{equation}
f(z_0) = X\beta(\varepsilon + XQ_0\beta + z_0a_0^\top \beta, XQ_0 + z_0a_0^\top) - X\beta - \varepsilon.
\end{equation}

The gradient $\nabla f$ with respect to $z_0$, holding $(XQ_0, \varepsilon)$ fixed, can be computed by the product rule and the chain rule via (1.8):
\begin{equation}
\nabla f(z_0)^\top = I_n \langle a_0, \beta \rangle + [(a_0, \beta)\hat{H}^\top + X\hat{G}^\top].
\end{equation}

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 $\eta$ requires considering difference of an expression at $(\varepsilon, XQ_0, z_0 + t\eta)$ minus the same expression at $(\varepsilon, XQ_0, z_0)$, dividing by $t$ and taking the limit as $t \rightarrow 0$; this is equivalent to considering the difference of an expression at $(\varepsilon, X(t))$ with $X(t) = X + t\eta a_0^\top$ minus the same expression at $(\varepsilon, X)$, dividing by $t$ and taking the limit as $t \rightarrow 0$.

Taking the trace of (1.17) and by definition of $\hat{D}f$ and $\hat{A}$ in (1.10), the identity
\begin{equation}
-\varepsilon_0 \overset{df}{=} \text{div } f(z_0) - z_0^\top f(z_0)
\end{equation}

(1.18) holds where $\text{div } f(z_0) = \text{trace}[\nabla f(z_0)]$. Since $\mathbb{E}[\text{div } f(z_0) - z_0^\top f(z_0) | (XQ_0, \varepsilon)] = 0$ by Stein’s formula [38], this provides the unbiased estimating equation (1.11). Reasoning conditionally on $(\varepsilon, XQ_0)$, using Stein formulæ with respect to $z_0$ involving conditional expectations given $(\varepsilon, XQ_0)$ and gradients of the form $\nabla f(z_0)$ holding $(\varepsilon, XQ_0)$ fixed will
be a recurring theme throughout the paper. In this context, the function \( f \) itself depends on \(( \varepsilon, XQ_0 )\) as in (1.16), although the dependence on \(( \varepsilon, XQ_0 )\) 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 [38, (8.6)] in a different form, the conditional variance of (1.18) given \((\varepsilon, XQ_0)\) is

\[
(1.19) \quad \text{Var}_0 \left[ \alpha \right] = E_0 \left[ \| f(z_0) \|^2 + \text{trace}\{ \nabla f(z_0) \}^2 \right]
\]

\[
= E_0 \left[ V^*(\theta) \right] \quad \text{for} \quad V^*(\theta) = \| y - X\hat{\beta} \|^2 + \text{trace}\{ \nabla f(z_0) \}^2
\]

where \( E_0 = E[\cdot | \varepsilon, XQ_0] \) denotes the conditional expectation with respect to \( \varepsilon \) given \((\varepsilon, XQ_0)\) and \( \text{Var}_0 \) denotes the conditional variance given \((\varepsilon, XQ_0)\). The gradient \( \nabla f(z_0) \) in (1.17) and the unbiased estimate \( V^*(\theta) \) of \( \text{Var}_0 \left[ \alpha \right] \) only depend on the unknown parameter of interest \( \theta \) and observable quantities, and \( V^*(\theta) \) is quadratic in \( \theta \).

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 \((\varepsilon, XQ_0)\), and (ii) \( V^*(\theta) \) is a consistent estimator of (1.19), the conditional variance of the random variable (1.18). Then the set of \( \theta \) for which the inequality

\[
(1.20) \quad \left[ (n - df) \langle \alpha_0, \hat{\beta} \rangle - \theta \rangle + \langle z_0, y - X\hat{\beta} \rangle + \hat{A} \right] - V^*(\theta) z_0^2/2 \leq 0
\]

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

\[
\langle y - X\hat{\beta}, z_0 \rangle, \quad \| y - X\hat{\beta} \|^2, \quad df, \quad \hat{A}, \quad a_{0}^\top \hat{\beta}
\]

and the derivatives of \( \hat{\beta} \). If the coefficient of \( \theta^2 \) in the left hand side of (1.20) is positive, (i.e., if the leading coefficient of (1.20), seen as a polynomial in \( \theta \) with data-driven coefficients, is positive), a \((1 - \alpha)\) confidence interval for \( \theta = a_{0}^\top \hat{\beta} \) is then given by

\[
(1.21) \quad CI = \left[ \Theta_1(z_{\alpha/2}), \Theta_2(z_{\alpha/2}) \right].
\]

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 \( \hat{\beta} \) is a convex penalized estimator. Although a variant of the above construction was briefly presented in [7, Section 6] (there, the function \( z_0 \rightarrow XQ_0(\hat{\beta} - \beta) - \varepsilon \) 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 \( \xi/\text{Var}[\xi]^{1/2} \) and \( N(0,1) \) for random variables of the form \( \xi = z^\top f(z) - \text{div} f(z) \) where \( z \sim N(0, I_n) \). 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 \(a \land b = \min\{a, b\}\), \(a \lor b = \max\{a, b\}\) and \(a_+ = a \lor 0\). Let \(I_d\) be the identity matrix of size \(d \times d\), e.g. \(d = n, p\). For any \(p \geq 1\), let \([p]\) be the set \(\{1, \ldots, p\}\). Let \(\| \cdot \|\) be the Euclidean norm and \(\| \cdot \|_q\) the \(\ell_q\) norm of vectors for any \(q \geq 1\), so that \(\| \cdot \| = \| \cdot \|_2\). Let \(\| \cdot \|_{op}\) be the operator norm of matrices and \(\| \cdot \|_F\) the Frobenius norm. Let \(\phi_{\min}(S)\) be the smallest eigenvalue of a symmetric matrix \(S\). We use the notation \(\langle \cdot, \cdot \rangle\) for the canonical scalar product of vectors in \(\mathbb{R}^n\) or \(\mathbb{R}^p\), i.e., \(\langle a, b \rangle = a^\top b\) for two vectors \(a, b\) of the same dimension. For any event \(\Omega\), denote by \(I_\Omega\) its indicator function. The unit sphere is \(S^{p-1} = \{x \in \mathbb{R}^p : \|x\| = 1\}\). Convergence in distribution is denoted by \(\rightarrow_d\) and convergence in probability by \(\rightarrow_p\). Throughout the paper, \(C_0, C_1, \ldots\) denote positive absolute constants, \(C_k(\gamma)\) positive constants depending on \(\gamma\) only, and \(C_k(\gamma, \mu)\) on \(\{\gamma, \mu\}\) only.

For any vector \(v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p\) and set \(A \subset [p]\), the vector \(v_A \in \mathbb{R}^{|A|}\) is the restriction \((v_j)_{j \in A}\) of \(v\). For any \(n \times p\) matrix \(M\) with columns \((M_1, \ldots, M_p)\) and any subset \(A \subset [p]\), let \(M_A = (M_j, j \in A)\) be the matrix composed of columns of \(M\) indexed by \(A\). If \(M\) is a symmetric matrix of size \(p \times p\) and \(A \subset [p]\), then \(M_{A,A}\) denotes the sub-matrix of \(M\) with rows and columns in \(A\), and \(M_{A,A}^{-1}\) is the inverse of \(M_{A,A}\). For any square matrix \(M\), let \(M^2 = (M + M^\top)/2\) 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, \(\nabla h\) is the transpose of the Jacobian of \(h\) and \(h(x + \eta) = h(x) + \nabla h(x)^\top \eta + o(\|\eta\|)\) if each coordinate \(h_i\) is Fréchet differentiable at \(x\). For deterministic matrices \(A \in \mathbb{R}^{m \times n}\), \(\nabla (Ah) = (\nabla h)A^\top \in \mathbb{R}^{n \times m}\). For \(f\) in (2.1), \(\nabla f(x) \in \mathbb{R}^{n \times n}\) and the divergence is \(\text{div}\, f(x) = \text{trace}\{\nabla f(x)\}\).

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

\[
(2.1) \quad \xi = z^\top f(z) - \text{div} f(z),
\]

for which Stein’s formula [38] states \(\mathbb{E}[\xi] = 0\), where \(z \sim N(0, I_n)\) is standard normal and \(f : \mathbb{R}^n \to \mathbb{R}^n\). We establish \(L_2\) bounds for the linear and quadratic approximations of \(\xi\) and construct consistent variance estimates in the related CLT.

Throughout this paper, the \(i\)-th coordinate \(f_i\) of \(f\) is a function \(f_i : \mathbb{R}^n \to \mathbb{R}\) and its weak gradient is denoted by \(\nabla f_i\). Similarly, the weak derivative of \(g(z)\) is denoted by \(\nabla g\). We refer to [15, Section 1.5] for definitions of weak differentiability. For the application to asymptotic normality of de-biased 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, [38, Eq. (8.6)] [7]] Let \(z \sim N(0, I_n)\) 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, i.e., \(\mathbb{E}[f_i(z)^2] + \mathbb{E}[\|\nabla f_i(z)\|^2] < +\infty\). Then

\[
(2.2) \quad \mathbb{E}[(z^\top f(z) - \text{div} f(z))^2] = \mathbb{E}[\|f(z)\|^2] + \mathbb{E}[\text{trace}\{\nabla f(z)^2\}].
\]

The above result, in the twice differentiable case, was known to Stein [38, Eq. (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 follow. We are looking for linear approximation of the random variable (2.1), of the form $z^\top \mu \sim N(0, \|\mu\|^2)$ for some deterministic $\mu \in \mathbb{R}^n$. We rewrite (2.1) as

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

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

**Theorem 2.2.** Let $z \sim N(0, I_n)$ 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, i.e. $\mathbb{E}[f_i(z)^2] + \mathbb{E}[\|\nabla f_i(z)\|^2] < +\infty$. Then $\xi = z^\top f(z) - \text{div} f(z)$ satisfies

$$\mathbb{E}\left[\left(\xi/\text{Var}[\xi]^{1/2} - Z\right)^2\right] = \epsilon_1^2 + (1 - (1 - \epsilon_1^2)^2) = \epsilon_1^2 + \epsilon_1 \epsilon_1' \tag{2.4}$$

with $Z = z^\top \mathbb{E}[f(z)]/\|\mathbb{E}[f(z)]\| \sim N(0, 1)$, deterministic real $1/4 \leq \epsilon_1 \leq 1$ and

$$\epsilon_1^2 \overset{\text{def}}{=} 1 - \frac{\|\mathbb{E}[f(z)]\|^2}{\text{Var}[\xi]} \leq \epsilon_2^2 \leq \frac{2\mathbb{E}[\|\nabla f(z)\|^2]}{\mathbb{E}[\|f(z)\|^2] + \mathbb{E}[\|\nabla f(z)\|^2]} \tag{2.5},$$

where $\epsilon_1^2 \overset{\text{def}}{=} 2\mathbb{E}[\|\nabla f(z)\|^2]/\|\mathbb{E}[f(z)]\|^2 + 2\mathbb{E}[\|\nabla f(z)\|^2]$. Consequently, \( \sup_{t \in \mathbb{R}} |\mathbb{P}(\xi/\text{Var}[\xi]^{1/2} \leq t) - \mathbb{P}(Z \leq t)| \leq C(\epsilon_1^2 + \epsilon_1 \epsilon_1')^{1/3} \) for $C = 1 + (2\pi)^{-1/2}$.

A direct consequence of Theorem 2.2 is $\epsilon_1^2 \leq (2.4) \leq 2\epsilon_1^2 \leq 2\epsilon_1^2$. Inequality (2.4) provides an upper bound on the 2-Wasserstein distance between $\xi/\text{Var}[\xi]^{1/2}$ and $Z \sim N(0, 1)$. When $\epsilon_1^2 \to 0$, it gives a stronger $L_2$ form of the CLT $\xi/\text{Var}[\xi]^{1/2} \to^d N(0,1)$ 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

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

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

$$\begin{align*}
\mathbb{E}[\xi - \text{Var}[\xi]^{1/2} Z]^2 &= \mathbb{E}[\|f(z) - r \mathbb{E}[f(z)]\|^2 + \text{trace}\{\nabla f(z)\}^2] \\
&= \mathbb{E}[\|f(z) - \mathbb{E}[f(z)]\|^2 + \text{trace}\{\nabla f(z)\}^2] + \text{Var}[\xi]^{1/2} - \|\mathbb{E}[f(z)]\|^2 \\
&= \text{Var}[\xi] - \|\mathbb{E}[f(z)]\|^2 + \{\text{Var}[\xi]^{1/2} - \|\mathbb{E}[f(z)]\|^2\}^2
\end{align*}$$

thanks to $(r - 1)\|\mathbb{E}[f(z)]\|$ = $\text{Var}[\xi]^{1/2} - \|\mathbb{E}[f(z)]\|$. Thus, (2.4) follows from the definition of $\epsilon_1^2$ in (2.5). Moreover, $\mathbb{E}[\|f(z) - \mathbb{E}[f(z)]\|^2] \leq \mathbb{E}[\|\nabla f(z)\|^2]$ by the Gaussian Poincaré inequality and $\|M\|^2_2 + \text{trace}(M^2) = 2\|M^\#\|^2_2$ for $M \in \mathbb{R}^{n \times n}$. Hence, with $a = \|\mathbb{E}[f(z)]\|^2$, $b = \mathbb{E}[\|f(z) - \mathbb{E}[f(z)]\|^2]$, $c = \text{trace}\{\nabla f(z)\}^2$ and $d = \mathbb{E}[\|\nabla f(z)\|^2]$, we have

$$\epsilon_1^2 = \frac{b + c}{a + b + c} \leq \frac{2d}{a + 2d} = \epsilon_2^2 \leq \frac{2\mathbb{E}[\|\nabla f(z)\|^2]}{\mathbb{E}[\|f(z)\|^2] + \|\nabla f(z)\|^2}$$

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

imsart-aos ver. 2020/05/28 file: asymptotic-normality.tex date: October 18, 2022
For any $\delta > 0$, by Markov’s inequality $\Pr(\xi/\text{Var}[\xi]^{1/2} \leq t) - \Pr(Z \leq t) \leq \Pr(|\xi/\text{Var}[\xi]^{1/2} - Z| > \delta) + \Pr(Z \in [t, t + \delta]) \leq (c_1^2 + c_1 \epsilon_1^4)/\delta^2 + \delta(2\pi)^{-1/2}$ 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 $\|\nabla f(z)\|_2^2$ is negligible compared with that of $\|f(z)\|^2$ to obtain $\xi/\text{Var}[\xi]^{1/2} \xrightarrow{d} N(0, 1)$. 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 [20] bounds the total variation distance $d_{TV}$ of $g(z)$ to the Gaussian distribution using the first and second derivatives of $g$: [20, Theorem 2.2] specialized to $W = g(z)$ with $z \sim N(0, I_n)$ states that

\begin{equation}
\label{eq:2.6}
d_{TV}\{W, N(\mu_0, \sigma_0^2)\} \leq (2\sqrt{5}/\sigma_0^2)\text{E}[\|\nabla g(z)\|_2^{1/4}\text{E}[\|\nabla^2 g(z)\|_{op}^{1/4}]
\end{equation}

where $W = g(z)$, $z \sim N(0, I_n)$, $\mu_0 = \text{E}[W]$ and $\sigma_0^2 = \text{Var}[W]$. Above, $\nabla g, \nabla^2 g$ denote the gradient and Hessian matrix of $g$. Inequality (2.6) provides a CLT for $g(z)$ provided that the moments of the derivatives $\text{E}[\|\nabla g(z)\|_2^{1/4}]$ and $\text{E}[\|\nabla^2 g(z)\|_{op}^{1/4}]$ are negligible compared to the variance $\sigma_0^2 = \text{Var}[g(z)]$. 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 [32]. 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 $(X, y)$. In fact, by Proposition 4.1 below, the Lasso is not twice differentiable as $\text{trace}(\partial/\partial y X \partial y X)$ is integer-valued. In Theorem 2.2, while $\xi = z^{-1}f(z) - \text{div} f(z)$ already involves the derivatives of $f$ through the divergence, the ratio $\epsilon_1^2$ that appears in the upper bound (2.5) only involves $f$ and its gradient $\nabla f$; the second derivatives of $f$ need not exist. Section 3 uses Theorem 2.2 to provide asymptotic normality for de-biasing estimators that are only once differentiable.

**Variance estimate.** It follows from Theorem 2.2 that random variables $\xi$ of the form (2.1) are asymptotically normal under the condition $1 - \|\text{E}[f(z)]\|_2^2/\text{Var}[\xi] \to 0$, or under a somewhat stronger but more explicit condition $\epsilon_1^2 \to 0$ as in (2.5). The following theorem provide consistent estimates of $\text{Var}[\xi]$.

**THEOREM 2.3.** Let $f, z, \xi, \epsilon_1^2$ and $c_1 \in [1/4, 1]$ be as in Theorem 2.2. Then,
\begin{equation}
\label{eq:2.7}
\text{E}[\|f(z)\|/\text{Var}[\xi]^{1/2} - 1]^2] \leq \epsilon_1^2 - \text{E}[\text{trace}(\{\nabla f(z)\}^2)]/\text{Var}[\xi] + c_1 \epsilon_1^4
\end{equation}

with $\epsilon_1^2 \overset{\text{def}}{=} 2\text{E}[\|\nabla f(z)\|_2^2]/\{|\text{E}[f(z)]\|_2^2 + 2\text{E}[\|\nabla f(z)\|_2^2]\} \geq \epsilon_1^2$. Consequently,
\begin{equation}
\label{eq:2.8}
\|f(z)\|^2/\text{Var}[\xi] \xrightarrow{d} 1 \quad \text{and} \quad \sqrt[1/2]{\|f(z)\|} \xrightarrow{d} N(0, 1).
\end{equation}

when $\epsilon_1^2 + \epsilon_1^2 I\{\text{E}[\text{trace}(\{\nabla f(z)\}^2)] < 0\} \to 0$.

**PROOF OF THEOREM 2.3.** It follows from the Jensen inequality and (2.4) that
\begin{equation}
\text{E}[\|f(z)\|/\text{Var}[\xi]^{1/2} - 1]^2] \leq \text{E}[\|f(z)\|^2]/\text{Var}[\xi] + 1 - 2\|\overline{\nabla}\|^2/\text{Var}[\xi]^{1/2}
\end{equation}

with $\overline{\nabla} = \text{E}[\nabla f(z)]$ due to $\epsilon_1^2 = 1 - \|\overline{\nabla}\|^2/\text{Var}[\xi]$. For the second inequality in (2.7),
\begin{equation}
\epsilon_1^2 - \text{E}[\text{trace}(\{\nabla f(z)\}^2)]/\text{Var}[\xi] = \text{E}[\|f(z) - \overline{\nabla}\|^2]/\text{Var}[\xi] \leq \text{E}[\|\nabla f(z)\|_2^2]/\text{Var}[\xi],
\end{equation}

thanks to the Gaussian Poincaré inequality, and $(1 - \epsilon_1^2)\epsilon_1^2/(2 - 2\epsilon_1^2)$ equals to the right-hand side above by the definition of $\epsilon_1^2$. \qed
2.2. **Quadratic approximation.** The decomposition (2.3) is especially useful if the linear part $z^\top \mu$ with $\mu = \mathbb{E}[f(z)]$ is a good approximation for $\xi = z^\top f(z) - \text{div } f(z)$. In some cases, e.g., if $f(z) = Az$ for some square deterministic matrix $A$, the decomposition (2.3) is uninformative. It is then natural to look for the best quadratic approximation of $\xi$ in the sense of the $L_2$ orthogonal projection to

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

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

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

$$\mathbb{E}[\{\xi/\text{Var}[\xi^{1/2} - \tilde{\xi}/\text{Var}[\xi^{1/2}])^2\} = 2(1 - \sqrt{1 - \epsilon^2}) = \epsilon^2 + c\epsilon^4$$

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

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

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

$$\mathbb{E}[(\xi - \xi_{\overline{\mu}, \overline{A}})^2] = \mathbb{E}[(\|f(z) - \overline{\mu}\|^2 - \|\overline{A}\|^2) + \mathbb{E}\text{trace}[\{\nabla f(z) - \overline{A}\}^2]] \leq 2\mathbb{E}\text{trace}[\{\nabla f(z) - \overline{A}\}^2].$$

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

For the projection $\xi_{\overline{\mu}, \overline{A}}$ of $\xi$ to $\mathcal{H}_{1,2}, \epsilon_{1,2} \defeq 1 - \mathbb{E}[(\xi_{\overline{\mu}, \overline{A}} - \xi_{\overline{\mu}, 0})^2]/\mathbb{E}[\xi^2] $ satisfies $\epsilon_{1,2} \leq \epsilon_{2,1} \defeq 2\mathbb{E}\text{trace}[\{\nabla f(z) - \overline{A}\}^2]/\|\overline{\mu}\|^2 + 2\mathbb{E}[(\|\nabla f(z)\|^2)/\|\overline{\mu}\|^2]$. and under the condition $\epsilon_{2,1} = o(1)$,

$$\|\overline{A}\|^2/\|\overline{\mu}\|^2 \to 0 \iff \xi/\text{Var}[\xi^{1/2}] \to_d N(0,1).$$

For the projection $\xi_{0, \overline{A}}$ of $\xi$ to $\mathcal{H}_2, \epsilon_{2} = 1 = \mathbb{E}[(\xi_{0, \overline{A}} - \xi_{0, 0})^2]/\mathbb{E}[\xi^2] $ satisfies $\epsilon_{2} \leq \epsilon_{2} \equiv \{\|\overline{\mu}\|^2 + 2\mathbb{E}[(\|\nabla f(z) - \overline{A}\|^2)/\|\overline{\mu}\|^2] \} + 2\mathbb{E}[(\|\nabla f(z)\|^2)/\|\overline{\mu}\|^2]$. and under the condition $\epsilon_{2} = o(1)$,

$$\|\overline{\mu}\|^2/\|\overline{\mu}\|^2 \to 0 \iff \xi/\text{Var}[\xi^{1/2}] \to_d N(0,1).$$

**Proof of Theorem 2.4.** The function $g(z) = f(z) - \mu - A^\top z$ has gradient $\nabla g = \nabla f - A$. Application of the Second Order Stein’s formula in Proposition 2.1 to $g$ yields

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

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

$$\|A\|^2/\|\overline{\mu}\|^2 - 2\mathbb{E}[z^\top A(f(z) - \mu)] = \|A\|^2/\|\overline{\mu}\|^2 - 2\mathbb{E}\text{trace}(\nabla f(z)A^\top)$$

$$= -\|\overline{A}\|^2/\|\overline{\mu}\|^2.$$


We also have $E[|V f(z) - \mathbf{A}|^2] = E[|V f(z)|^2] - E[|\mathbf{A}|^2]$ so that
\[
I = E[|f(z) - \mu|^2 - |V f(z)|^2] + E[|V f(z) - \mathbf{A}|^2] + \|\mathbf{A} - \mathbf{A}\|^2.
\]
For the second term, using that $E[\nabla f(z) - \mathbf{A}] = 0$ we get
\[
II = E[\text{trace}(\nabla f(z) - \mathbf{A})^2] = E[\text{trace}(\nabla f(z) - \mathbf{A})^2] + \text{trace}(\mathbf{A} - \mathbf{A})^2).
\]
Due to $\|\mathbf{M}\|^2 + \text{trace}(\mathbf{M}^2) = 2\|\mathbf{M}\|\|\mathbf{M}\|^2$ for $\mathbf{M} \in \mathbb{R}^{n \times n}$, it follows that
\[
E[(\xi - \xi_0\mathbf{A})^2] = E[|f(z) - \mu|^2 - |\mathbf{A}|^2] + E[|V f(z) - \mathbf{A}|^2] + E[\text{trace}(\mathbf{A} - \mathbf{A})^2] = 2E[|V f(z) - \mathbf{A}|^2] + E[\text{trace}(\mathbf{A} - \mathbf{A})^2] = 2E[|V f(z) - \mathbf{A}|^2]
\]
the optimality of $\mu = \mu$ and $\mathbf{A} = \mathbf{A}$ follows, so that $\xi_0\mathbf{A}$ is the $L_2$ projection of $\xi$ to $\mathcal{M}_1$.

Also, the first line above gives the formula of $E[(\xi - \xi_0\mathbf{A})^2]$ in (2.10), and the second line gives the formulas for the variances of $\xi_{\mathbf{A}, 0}$ and $\xi_{0, \mathbf{A}}$. The upper bound in (2.10) follows from $E[|f(z) - \mu|^2 - |\mathbf{A}|^2] \leq E[|\nabla f(z)|^2] - |\mathbf{A}|^2 = E[|\nabla f(z) - \mathbf{A}|^2]$ thanks to the Gaussian Poincaré inequality. Inequality (2.10) is equivalent to
\[
E[\xi^2] = E[\xi_0^2] + E[(\xi - \xi_0\mathbf{A})^2] \leq \|\mu\|^2 + 2\|\mathbf{A}\|^2 + 2E[|\nabla f(z) - \mathbf{A}|^2]
\]
which provides $c_{1,2}^2 \leq c_{1,2}^2$ by bounding from above the denominator in $c_{1,2}^2 = 1 - E[\xi_0^2]/E[\xi^2]$ and $c_2^2 = 1 - E[\xi_0^2]/E[\xi^2]$. 

For (2.11), we write $\xi_0\mathbf{A} = \sum_{j=1}^{n} (a_j G_j + b_j (G_j^2 - 1))$ with iid $G_j \sim N(0, 1)$, where $a_j = u_j^\top \mu$ and $G_j = u_j^\top z$ with the eigenvalue decomposition $\mathbf{A} = \sum_{j=1}^{n} b_j u_j u_j^\top$. Assume without loss of generality that $\text{Var}(a_j G_j + b_j (G_j^2 - 1)) = a_j^2 + 2b_j^2$ is non-increasing in $j$ and that $\text{Var}(\xi_{\mathbf{A}, 0})$ 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 $k_n \overset{\text{def}}{=} [\|\mathbf{A}\|_{op}] = [(\max_j b_j)^{-1}]$ satisfies $k_n \to +\infty$ and $\sum_{j=1}^{k_n} b_j^2 \ll 1 = \text{Var}(\xi_{\mathbf{A}, 0})$, so that
\[
\xi_0\mathbf{A} = \sum_{j=1}^{k_n} a_j G_j + \sum_{j=k_n+1}^{n} \{a_j G_j + b_j (G_j^2 - 1)\} + o_p(1).
\]
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 $\max_{j \leq k_n} (a_j^2 + 2b_j^2) = a_{k_n+1}^2 + 2b_{k_n+1}^2 \to 0$ so that the second term above is independent of the first and approximately $N(0, 1 - c)$ by the Lyapunov CLT when $\sum_{j=1}^{k_n} a_j^2 \to c \leq 1$. 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 $j_n \leq n$. As $W_{j_n}$ is an independent component of the sum, for any $(a_j, b_j) \to (a, b)$ 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-Lévy theorem and $W_{j_n} \to^d aG + b(G^2 + 1)$ for some $G \sim N(0, 1)$. As $j_n \leq n$ are arbitrary, this gives $\|\mathbf{A}\|_{op} = \max_{j=1, \ldots, n} b_j^2 \to 0$. 

\textbf{Variance estimate: quadratic case.} Theorem 2.4 provides the quadratic normal approximation of $\xi$ under the condition $c_{1,2}^2 \to 0$ with
\[
(2.13) \quad c_{1,2}^2 \geq c_{1,2}^2 = 1 - (\|E[f(z)]\|^2 + 2\|E[\nabla f(z)]\|^2)/\text{Var}[\xi] \geq \|E[f(z)]\|^2 + 2E[\|\nabla f(z)\|^2] \geq \text{Var}[\xi]
\]
where $c_{1,2}^2$ is defined using the upper bound $\|E[f(z)]\|^2 + 2E[\|\nabla f(z)\|^2] \geq \text{Var}[\xi]$.
Theorem 2.5. Let $f, z, \xi, \overline{\mu} = \mathbb{E}[f(z)]$ and $\overline{\Sigma} = \mathbb{E}[\nabla f(z)]$ be as in Theorem 2.4 and $\text{Var}[\xi] = \|f(z)\|^2 + \text{trace}(\{\nabla f(z)\}^2)$. Then,
\[
(2.14) \quad \mathbb{E}\left[\frac{\text{Var}[\xi]}{\text{Var}[\xi]} - 1\right] \leq 2\overline{\epsilon}_{1,2}^2 + 2\overline{\epsilon}_{1,2}C_0 + C_0\|\overline{\Sigma}_{op}\|/\text{Var}[\xi]^{1/2}
\]
with $\overline{\epsilon}_{1,2} \overset{\text{def}}{=} \{2\mathbb{E}[\|\nabla f(z) - \overline{\Sigma}^2_{op}\}/\text{Var}[\xi]\}^{1/2}$ and $C_0 \overset{\text{def}}{=} \{\|\overline{\mu}\|^2 + 2\|\overline{\Sigma}_{op}\|^2\}/\text{Var}[\xi]^{1/2}$. Consequently, under the conditions $\mathbb{E}[\|\nabla f(z) - \overline{\Sigma}^2_{op}\]/\text{Var}[\xi] + \|\overline{\Sigma}_{op}\|/\text{Var}[\xi] = o(1)$ and $\|\overline{\Sigma}^2_{op}\|/\|\overline{\mu}\|^2 + \|\overline{\Sigma}^2_{op}\| = o(1)$ for the quantities in (2.13) and (2.11). The proof is given in Appendix D.

3. De-biasing general convex regularizers. Our main application of the normal approximation in Theorem 2.2 concerns de-biasing regularized estimators of the form
\[
(3.1) \quad \hat{\beta} = \arg\min_{b \in \mathbb{R}^p} \left\{ \|y - Xb\|^2/(2n) + g(b) \right\}
\]
for convex $g : \mathbb{R}^p \rightarrow \mathbb{R}$ in the linear model (1.1). Throughout, let $h = \hat{\beta} - \beta$ be the error vector, $a_0 \in \mathbb{R}^p$ be a direction of interest, $\theta = \langle a_0, \beta \rangle$ be the target of statistical inference, and $u_0, z_0$ and $Q_0$ be as in (1.13) so that (1.14) holds.

3.1. Assumption. We say that $g$ is $\mu$-strongly convex with respect to the norm $b \rightarrow \|\Sigma^{1/2}b\|$ if its symmetric Bregman divergence is bounded from below as
\[
(3.2) \quad (\hat{b} - b)\top ((\partial g)(\hat{b}) - (\partial g)(b)) \geq \mu\|\Sigma^{1/2}(\hat{b} - b)\|^2
\]
for some $\mu > 0$. Here the interpretation of (3.2) is its validity for all choices in the sub-differential $(\partial g)(\hat{b})$ and $(\partial g)(b)$. Condition (3.2) holds for any convex $g$ for $\mu = 0$. If $g$ is twice differentiable, (3.2) holds if and only if $\mu\Sigma$ is a lower bound for the Hessian of $g$. However, (3.2) may also hold for non-differentiable $g$, e.g. the Elastic-Net penalty with $\Sigma = I_p$. Our results require the following assumption.

Assumption 3.1. (i) Let $\gamma > 0, \mu \in [0, 1/2]$ 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 $n, p \rightarrow +\infty$ and $p/n \leq \gamma$. The penalty $g : \mathbb{R}^p \rightarrow \mathbb{R}$ in (3.1) is convex and (3.2) holds. The rows of $X$ are iid $N(0, \Sigma)$ with invertible $\Sigma$ and the noise $\epsilon \sim N(0, \sigma^2I_n)$ is independent of $X$. (ii) $a_0 \in \mathbb{R}^p$ is a sequence of vectors normalized with $\|\Sigma^{−1/2}a_0\| = 1$.

Note that if (3.2) holds for $\mu \geq 0$ 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 $\mathbb{P}(\phi_{\min}(\Sigma^{−1/2}X\top X\Sigma^{−1/2}) > 0) = 1$ (cf. Appendix A) and the objective function of the optimization problem (3.1) is $(\phi_{\min}(\Sigma^{−1/2}X\top X\Sigma^{−1/2}/n) + \mu)$-strongly convex, Assumption 3.1 grants almost surely the existence and uniqueness of the minimizer (3.1).
3.2. Gradient with respect to $y$ and effective degrees-of freedom. Consider a penalized estimator (3.1) viewed as a function $\hat{\beta} = \hat{\beta}(y, X)$. For every $X \in \mathbb{R}^{n \times p}$, the map $y \mapsto X\hat{\beta}(y, X)$ is $1$-Lipschitz (cf. Proposition 7.3). By Rademacher’s theorem, for almost every $y$ there exists a unique matrix $\hat{H} \in \mathbb{R}^{n \times n}$ such that

$$
(X\hat{\beta}(y + \eta, X) = X\hat{\beta}(y, X) + \hat{H}^\top \eta + o(\|\eta\|),
$$

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

Table 1 provides closed-form expressions of $\hat{H}$ for specific penalty functions $g$. The proofs of these closed-form expressions will be given in Section 4. An advantage of defining $\hat{H}$ as the Fréchet gradient (3.3) of any convex penalty $g$, even though for arbitrary penalty $g$ we are unable to provide a closed-form expression for $\hat{H}$. Finally, define the effective degrees-of-freedom $\hat{d}f$ of $\hat{\beta}$ by

$$
\hat{d}f = \text{trace}[\hat{H}]
$$

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

| Penalty | $\hat{H} \in \mathbb{R}^{n \times n}$ | Justification |
|---------|---------------------------------|---------------|
| $g(b) = \lambda \|b\|_1$ (Lasso) | $X\hat{g}(X^\top X\hat{g})^{-1}X^\top \hat{g}$ | [45], Proposition 4.1 |
| $g(b) = \mu \|b\|_2^2$ (Ridge) | $X(X^\top X + n\mu I_p)^{-1}X^\top$ | (1.12), Section 4.1 |
| $g(b) = \lambda \|b\|_1 + \mu \|b\|_2^2$ (Elastic-Net) | $X\hat{g}(X^\top X\hat{g} + n\mu I_{\hat{g}})^{-1}X^\top \hat{g}$ | [45], (28), [7, §3.5.3] |
| $g(b) = \|b\|_{2L} = \sum_{k=1}^K \lambda_k \|b_{G_k}\|_2$ (group Lasso (3.33)) | $X(X^\top X + n\nabla^2 g(\hat{\beta}))^{-1}X^\top$ | [46], Proposition 4.2 |
| $g(b)$ twice continuously differentiable | | Section 4.1 |
| $g(b)$ arbitrary convex function | symmetric with eigenvalues in $[0, 1]$ | Proposition 7.3 |

Table 1

Closed-form expressions $\hat{H}$ from Equation (3.3) for specific convex penalty functions $g : \mathbb{R}^p \to \mathbb{R}$. For the Lasso and Elastic-Net, $\hat{S} = \{ j \in [p] : \beta_j \neq 0 \}$. For the group Lasso, $\hat{S}$ and $M$ are given in Section 4.3.

3.3. Approximation for $\xi_0 = z_0^\top f(z_0) - \text{div} f(z_0)$ and the de-biased vector $\hat{\beta}^{\text{(de-bias)}}$. Consider, for a fixed value of $(XQ_0, \epsilon)$ the function $f(XQ_0, \epsilon) : \mathbb{R}^n \to \mathbb{R}$ given by

$$
f(XQ_0, \epsilon)(z_0) = f(z_0) = X\hat{\beta} - y.
$$

For brevity we will often omit the dependence on $(XQ_0, \epsilon)$ of $f$ as discussed after (1.16). The Fréchet gradient $\nabla f(z_0)$, where it exists, is uniquely defined by

$$
f(XQ_0, \epsilon)(z_0 + \eta) - f(XQ_0, \epsilon)(z_0) = [\nabla f(z_0)]^\top \eta + o(\|\eta\|)
$$

and the divergence by $\text{div} f(z_0) = \text{trace}[\nabla f(z_0)]$. If $\hat{\beta} = \text{arg min}_{b \in \mathbb{R}^p}(\|\epsilon - X(b - \beta)\|^2/(2n) + g(b))$ with $X = X + \eta a_{0\top}$, then (3.6) is equivalent to

$$
(X(\hat{\beta} - \beta) - \epsilon) - (X\hat{\beta} - y) = [\nabla f(z_0)]^\top \eta + o(\|\eta\|).
$$
By Stein’s formula, we have conditionally on \((XQ_0, \varepsilon)\) that almost surely
\[
\mathbb{E}[\xi_0 \mid (XQ_0, \varepsilon)] = 0 \quad \text{for} \quad \xi_0 = z_0^\top f(z_0) - \text{div} f(z_0).
\]
As in (1.18) for the general case discussed in the introduction, (3.8) gives an unbiased estimating equation for \(\theta = \langle a_0, \beta \rangle\). The next lemma provides an expression for \(\nabla f(z_0)\).

**Lemma 3.1.** Let Assumption 3.1(i) be fulfilled, \(a_0 \in \mathbb{R}^p\) and \(\tilde{H}\) be as in (3.3). Then,
\[
\nabla f(z_0)^\top = (I_n - \tilde{H})^\top \langle a_0, h \rangle + w_0(y - X\hat{\beta})^\top
\]
satisfies (3.6) for some random \(w_0 \in \mathbb{R}^n\) almost surely. If additionally \(\|\Sigma^{-1/2}a_0\| = 1\) then
\[
\|w_0\|^2 \leq n^{-1} \min\{(4\mu)^{-1}, \phi_{\min}(\Sigma^{-1/2}X^\top X\Sigma^{-1/2}/n)^{-1}\}.
\]

| Penalty                          | Vector \(w_0 \in \mathbb{R}^n\) in Lemma 3.1 | Justification |
|---------------------------------|---------------------------------------------|---------------|
| \(g(b) = \lambda\|b\|_1\) (Lasso) | \(Xg(Xg^\top Xg)^{-1}(a_0)_g\)            | [8], Proposition 4.1 |
| \(g(b) = \mu\|b\|_2^2\) (Ridge)  | \(X(X^\top X + n\mu I_p)^{-1}a_0\)         | Section 4.1    |
| \(g(b) = \|b\|_{\text{GL}} = \sum_{k=1}^{K} \lambda_k \|b_{\alpha_k}\|_2\) (group Lasso (3.33)) | \(Xg(Xg^\top Xg + M)^{-1}(a_0)_g\) | Proposition 4.2 |
| \(g(b)\) twice continuously differentiable | \(X(X^\top X + n\nabla^2 g(\beta))^{-1}a_0\) | Section 4.1    |

**Table 2**

Closed-form expressions for \(w_0 \in \mathbb{R}^n\) in Lemma 3.1 for specific convex penalties \(g: \mathbb{R}^p \to \mathbb{R}\).

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 \(w_0\) in (3.9) is linear in \(a_0\) in the sense that \(w_0\) can be chosen of the form \(W\Sigma^{-1/2}a_0\) for some matrix \(W \in \mathbb{R}^{n \times p}\). Indeed, the proof of Lemma 3.1 shows that the map \((\varepsilon, X) \mapsto X(\hat{\beta} - \beta) - \varepsilon\) is Fréchet differentiable at almost every point by Rademacher’s theorem. At such a point, with \(a_1, a_2 \in \mathbb{R}^p\), \(t \in \mathbb{R}\), the linear combination \(a_3 = a_1 + ta_2\) and the perturbed design matrix \(\tilde{X} = X + \eta(a_1 + ta_2)^\top\), linearity of the Fréchet derivative implies that
\[
(\tilde{X}(\hat{\beta} - \beta) - \varepsilon) - (X\hat{\beta} - y) = ((a_1 + ta_2, h)(I_n - \tilde{H})^\top + (w_1 + tw_2)(y - X\hat{\beta})^\top) \eta + o(\|\eta\|)
\]
where \(w_1\) and \(w_2\) denote the \(w_0\) from (3.9) for \(a_0 = a_1\) and \(a_0 = a_2\) respectively. Hence with \(w_3 = w_1 + tw_2\), (3.9) holds for \((a_0, w_0) = (a_3, w_3)\). This proves that \(w_0\) is linear in \(a_0\), i.e., it is of the form \(w_0 = W\Sigma^{-1/2}a_0\) for some matrix \(W \in \mathbb{R}^{n \times p}\). One way to construct \(W\) explicitly is the following: define the \(j\)-th column of \(W\) as the vector \(w_0\) corresponding to \(a_0 = \Sigma^{1/2}e_j\) where \(e_j\) is the \(j\)-th canonical basis vector. The linearity proved above for any linear combination \(a_3\) then implies that (3.9) holds for \(w_0 = W\Sigma^{-1/2}a_0\) for any \(a_0 \in \mathbb{R}^p\).

Inequality (3.10) provides an upper bound on the operator norm of \(W\). Linearity of \(w_0\) with respect to \(a_0\) and explicit matrices \(W\) can be seen for some penalty functions in Table 2. Such Fréchet differentiability with respect to \(X\) is used in [4, 5] to develop estimates of \(\|\Sigma^{1/2}(\hat{\beta} - \beta)\|^2\) 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
\[
-\xi_0 = \text{div} f(z_0) - (z_0, f(z_0)) = (n - df)((a_0, \hat{\beta}) - \theta) + (z_0 + w_0, y - X\hat{\beta}) = (n - df)(\theta - \theta)
\]
(3.11)
for
\begin{equation}
\hat{\theta} \equiv \langle a_0, \beta \rangle + (n - \hat{d}f)^{-1} \langle z_0 + w_0, y - X\hat{\beta} \rangle.
\end{equation}

In the present context of the regularized estimator \( \hat{\beta} \) in (3.1) with its effective degrees-of-freedom \( \hat{d}f \) defined in (3.4) and under Assumption 3.1, the quantities \( \xi_0, \hat{d}f, \hat{\theta} \) 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{equation}
0 = \mathbb{E}[-\xi_0] = \mathbb{E}[(n - \hat{d}f)(\hat{\theta} - \theta)]
= \mathbb{E}[(n - \hat{d}f)(\langle a_0, \beta \rangle - \theta) + \langle z_0 + w_0, y - X\hat{\beta} \rangle]
\end{equation}
by taking expectations in (3.11). This provides a first evidence that the correction \( (n - \hat{d}f)^{-1} \langle z_0 + w_0, y - X\beta \rangle \) indeed removes the bias, at least after multiplication of \( (\hat{\theta} - \theta) \) by \( (n - \hat{d}f) \). Since \( z_0 = X\Sigma^{-1}a_0 \) 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 \( w_0 \) and \( \hat{H} \). If \( \hat{\beta} \) is given by (3.1), by identifying the terms in (1.18) and (3.11) we see that the random variable \( \tilde{A} \) defined in (1.10) of the introduction is here \( \tilde{A} = \langle w_0, y - X\hat{\beta} \rangle \) with \( w_0 \) given by Lemma 3.1. The following lemma shows that this term is negligible.

**Lemma 3.2.** Under Assumption 3.1 there exists \( \Omega_n \) with \( \mathbb{P}(\Omega_n^c) \leq C_0(\gamma, \mu)n^{-1/2} \) and
\begin{equation}
\mathbb{E}[I_{\Omega_n} \langle w_0, y - X\hat{\beta} \rangle^2 / \text{Var}_0[\xi_0]] \leq C_1(\gamma, \mu)n^{-1}.
\end{equation}

The proof is given in Section 7.6. Since \( \mathbb{P}(\Omega_n) \to 1 \), inequality (3.14) implies \( \langle w_0, y - X\beta \rangle^2 / \text{Var}_0[\xi_0] \to 0 \). This motivates the definition
\begin{equation}
\hat{\beta}_{\text{de-bias}} \equiv \beta + (n - \hat{d}f)^{-1} \Sigma^{-1}X^\top (y - X\hat{\beta}).
\end{equation}
The de-biased estimate \( \langle a_0, \hat{\beta}_{\text{de-bias}} \rangle \) in direction \( a_0 \) is obtained from \( \hat{\theta} \) in (3.12) by dropping the smaller order term \( (n - \hat{d}f)^{-1}(w_0, y - X\hat{\beta}) \). By Slutsky’s theorem, (3.14) implies
\begin{equation}
\frac{\xi_0}{\text{Var}_0[\xi_0]^{1/2}} \to_F F \quad \text{if and only if} \quad \frac{(n - \hat{d}f)(\langle a_0, \hat{\beta}_{\text{de-bias}} - \beta \rangle)}{\text{Var}_0[\xi_0]^{1/2}} \to d F
\end{equation}
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 \( \hat{\beta} \). 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 \( \xi_0 / \text{Var}_0[\xi_0] \) 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 \( \xi_0 \), the next subsections build estimate of its variance and derive asymptotic normality results by showing that \( F = N(0, 1) \) for most directions \( a_0 \). The next result provides a loose data-driven upper bound on the error \( \langle a_0, \hat{\beta}_{\text{de-bias}} - \beta \rangle \) and

**Theorem 3.3.** Under Assumption 3.1 there exists \( \Omega_n \) with \( \mathbb{P}(\Omega_n^c) \leq C_0(\gamma, \mu)n^{-1/2} \) and
\begin{equation}
\mathbb{E}[I_{\Omega_n} (n - \hat{d}f)^2 \langle a_0, \hat{\beta}_{\text{de-bias}} - \beta \rangle^2 / \|y - X\hat{\beta}\|^2] \leq C_1(\gamma, \mu).
\end{equation}
Furthermore \( \|\langle a_0, \hat{\beta}_{\text{de-bias}} - \beta \rangle\| = O_P(1)\|y - X\hat{\beta}\|/(n - \hat{d}f) = O_P(1)\|y - X\hat{\beta}\|/n. \)

Theorem 3.3 is proved in Section 7.7. If \( \|X(\hat{\beta} - \beta)\|^2/n = O_P(\sigma^2) \) then \( \|y - X\hat{\beta}\|/n = O_P(1)\sigma/\sqrt{n} \) 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 \( \|X(\hat{\beta} - \beta)\|^2/n \), the second term in (3.15) indeed corrects the bias, achieving \( \langle a_0, \hat{\beta}_{\text{de-bias}} - \beta \rangle = O_P(1)\sigma/\sqrt{n}. \)
3.4. Variance estimates. By Proposition 2.1, the conditional variance \( \text{Var}_0[\xi_0] \) can be written as \( \text{Var}_0[\xi_0] = \mathbb{E}_0[V^*(\theta)] \) for

\[
V^*(\theta) \overset{\text{def}}{=} \| y - X \hat{\beta} \|^2 + \text{trace}[\{ \nabla f(z_0) \}^2].
\]

We allow the variance estimate to depend on the unknown \( \theta = (a_0, \beta) \) as the resulting pivotal quantity, \( -V^*(\theta)^{-1/2} \xi_0 = V^*(\theta)^{-1/2}(n - \hat{df})(\hat{\theta} - \theta) \) via (3.11), would depend on \( \theta \) anyway. While \( V^*(\theta) \) itself can be used to estimate \( \text{Var}_0[\xi_0] \), its sign is unclear. The following simplified version of it, obtained by removing the smaller order terms in \( V^*(\theta) \),

\[
\tilde{V}(\theta) \overset{\text{def}}{=} \| y - X \tilde{\beta} \|^2 + \text{trace}[\tilde{H} - I_n]^2(\langle a_0, \tilde{\beta} \rangle - \theta)^2
\]

is non-negative. This follows from Proposition 7.3 since \( I_n - \tilde{H} \) is almost surely positive semi-definite. Lemma 3.4 below shows that the relative bias \( \mathbb{E}_0[\tilde{V}(\theta)]/\text{Var}_0[\xi_0] - 1 \) converges to 0 in probability, i.e., \( \tilde{V}(\theta) \) is asymptotically unbiased for \( \text{Var}_0[\xi_0] \).

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

\[
\mathbb{E} \left[ I_{\Omega_n} \mathbb{E}_0[\tilde{V}(\theta)] - 1 \right] \leq \mathbb{E} \left[ I_{\Omega_n} \mathbb{E}_0[\tilde{V}(\theta)] - \text{Var}_0[\xi_0] \right] \leq C_3(\gamma, \mu)/n.
\]

An alternative variance estimate, that does not depend on the unknown parameter \( \theta \), is given by replacing \( \theta \) in \( \tilde{V}(\theta) \) by the point estimate \( \langle a_0, \hat{\beta}(\text{de-bias}) \rangle \) with \( \hat{\beta}(\text{de-bias}) \) in (3.15):

\[
\tilde{V}(a_0) = \tilde{V}(\langle a_0, \hat{\beta}(\text{de-bias}) \rangle) = \| y - X \hat{\beta} \|^2 + \| I_n - \tilde{H} \|^2 = \langle x_0, y - X \hat{\beta} \rangle^2 / (n - \hat{df})^2.
\]

The next lemma provides \( \tilde{V}(a_0)/\tilde{V}(\theta) \to \mathbb{P} 1 \) and that \( \tilde{V}(a_0) \) is also asymptotically unbiased in the sense \( \mathbb{E}_0[\tilde{V}(a_0)]/\text{Var}_0[\xi_0] \to \mathbb{P} 1 \). Lemmas 3.4 and 3.5 are proved in Section 7.6.

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

\[
\max\left\{ \mathbb{E} \left[ I_{\Omega_n} \left( \tilde{V}(a_0)^{1/2} / \tilde{V}(\theta)^{1/2} - 1 \right)^2 \right], \mathbb{E} \left[ I_{\Omega_n} \left( \mathbb{E}_0[\tilde{V}(a_0)]^{1/2} / \text{Var}_0[\xi_0]^{1/2} - 1 \right)^2 \right] \right\} \leq C_4(\gamma, \mu)/n.
\]

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

\[
\beta^* \overset{\text{def}}{=} \arg\min_{b \in \mathbb{R}^p} \| \Sigma^{1/2}(\beta - b) \|^2/2 + g(b), \quad h^* \overset{\text{def}}{=} \beta^* - \beta, \quad R_\beta \overset{\text{def}}{=} \sigma^2 + \| \Sigma^{1/2}h^* \|^2.
\]

Our first result provides asymptotic normality of the de-biased estimate when the error \( \langle a_0, \hat{\beta} - \beta \rangle \) of \( \hat{\beta} \) in direction \( a_0 \) is negligible compared to \( R_\beta \).

**Theorem 3.6.** Let Assumption 3.1 be fulfilled. Let \( \hat{\beta}(\text{de-bias}) \) be as in (3.15). Then, for any \( a_0 \) with \( \| \Sigma^{-1/2}a_0 \| = 1 \) such that \( \langle a_0, \hat{h} \rangle / R_\beta \to \mathbb{P} 0 \),

\[
\sup_{t \in \mathbb{R}} \left[ \mathbb{P} \left( \frac{\xi_0}{V_0^{1/2}} \leq t \right) - \Phi(t) \right] + \mathbb{P} \left( \frac{\langle a_0, \hat{\beta}(\text{de-bias}) \rangle - \theta}{V_0^{1/2} / (n - \hat{df})} \leq t \right) - \Phi(t) \to 0,
\]

where \( V_0 \) denotes any of the four quantities: \( \text{Var}_0[\xi_0], \| y - X \hat{\beta} \|^2, \tilde{V}(\theta) \) or \( \tilde{V}(a_0) \).
Theorem 3.6 is proved in Section 7.7. The theorem, as well as its variants below, are obtained by applying Theorem 2.2 conditionally on \((\varepsilon, XQ_0)\) to the function \(f(z_0)\) in (3.5). This argument relies on the normality of \(z_0\) conditionally on \((\varepsilon, XQ_0)\) and thus the Gaussian design assumption. Here is an outline. Define

\[
(3.24) \quad \delta_1^2(a_0) \doteq \frac{E_0[\|\nabla f(z_0)\|^2]}{E_0[\|f(z_0)\|^2] + E_0[\|\nabla f(z_0)\|^2]}
\]

where \(E_0\) and \(\text{Var}_0\) are the conditional expectation and conditional variance given \((XQ_0, \varepsilon)\).

Furthermore, with Assumption 3.1 be fulfilled, it is sufficient to show that \(\delta_1^2(a_0) \to^P 0\) in order to prove asymptotic normality of \(\xi_0/\text{Var}_0(\xi_0)^{1/2}\) by Theorem 2.2 and of \(\xi_0/\|f(z_0)\|\) by Theorem 2.3 since \(\delta_1^2 = \delta_2^2(a_0)\) satisfies \(2\delta_1^2 \geq \max(\varepsilon_1^2, \varepsilon_2^2)\) for the \(\varepsilon_1, \varepsilon_2\) in Theorems 2.2 and 2.3. The proof makes rigorous the following informal bound:

\[
\delta_1^2(a_0) = \frac{E_0[\|\nabla f(z_0)\|^2]}{E_0[\|f(z_0)\|^2] + E_0[\|\nabla f(z_0)\|^2]} \leq \frac{E_0[\|I_n - \widehat{H}^2\|_F^2(\langle a_0, h \rangle)^2]}{C_2(\gamma, \mu)nR_s} + o_P(n^{-1/2})
\]

for some constant \(C_2(\gamma, \mu)\), by establishing a lower bound on \(\|f(z_0)\|^2 = \|y - X\tilde{\beta}\|^2\) for the denominator (Lemmas 7.4, 7.6 and 7.7), and by showing that the rank one term \(w_0(y - X\tilde{\beta})^T\) in (3.9) is negligible in the numerator. Finally, \(\|I_n - \widehat{H}\|^2_F \leq n\) always holds by Proposition 7.3 and \(\langle a_0, h \rangle^2/R_s\) is shown to be uniformly integrable, so that the assumption \(\langle a_0, h \rangle^2/R_s \to^P 0\) grants \(E[\delta_2^2(a_0)] \to 0\). The next two results identify directions \(a_0\) such that \(\langle a_0, h \rangle^2/R_s \to^P 0\) holds.

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

\[
(3.25) \quad \mathcal{S} = \left\{ v \in \mathbb{S}^{p-1} : E[(\Sigma^{1/2} v, \hat{\beta})^2/\|\Sigma^{1/2} h\|^2] \leq C^*/a_p \right\}
\]

of the unit sphere \(\mathbb{S}^{p-1}\) in \(\mathbb{R}^p\) has relative volume \(\mathcal{S}/|\mathbb{S}^{p-1}| \geq 1 - 2e^{-p/a_p}\) and

\[
(3.26) \quad \sup_{a_0 \in \Sigma^{1/2} \mathcal{S}} \sup_{t \in \mathbb{R}} \left[ P\left( \left| \frac{\xi_0}{V_0^{1/2}} \right| \leq t \right) - \Phi(t) \right] \to 0
\]

where \(V_0\) denotes any of the four quantities: \(\text{Var}_0(\xi_0), \|y - X\tilde{\beta}\|^2, \hat{V}(\theta)\) or \(\hat{V}(a_0)\). Furthermore, with \(e_j \in \mathbb{R}^p\) the \(j\)-th canonical basis vector and \(\phi_{\text{cond}}(\Sigma) = \|\Sigma^{1/2} \Sigma^{-1} \|_{op}\), the asymptotic normality in (3.26) uniformly holds over at least \((p - \phi_{\text{cond}}(\Sigma))a_p/C^*\) canonical directions in the sense that \(J_p = \{ j \in [p] : e_j/\|\Sigma^{1/2} e_j\| \in \Sigma^{1/2} \mathcal{S} \}\) has cardinality \(|J_p| \geq p - \phi_{\text{cond}}(\Sigma)a_p/C^*\).

Theorem 3.7 is proved in Section 7.7. For a given sequence of directions \(a_0 \in \Sigma^{1/2} \mathbb{S}^{p-1}\), if \(b_p = E[(\langle a_0, h \rangle^2/\|\Sigma^{1/2} h\|^2) \to 0\) then it follows by choosing \(a_p = C^*/b_p\) that \(a_0 \in \Sigma^{1/2} \mathcal{S}\) for the \(\mathcal{S}\) in (3.25) so that (3.26) implies that asymptotic normality holds for this sequence of \(a_0\). In other words, asymptotic normality holds for all \(a_0\) such that \(E[(\langle a_0, h \rangle^2/\|\Sigma^{1/2} h\|^2) \to 0\). Thus a sequence of directions \(a_0\) for which asymptotic normality does not follow from Theorem 3.7 is a sequence such that \(E[(\langle a_0, h \rangle^2/\|\Sigma^{1/2} h\|^2) \neq 0\), i.e., the squared error \(\langle a_0, h \rangle^2\) in direction \(a_0\) carries a constant fraction of the full prediction error \(\|\Sigma^{1/2} h\|^2\). Such direction \(a_0\) must thus be very special, which is embodied by the exponentially small bound on the relative volume \(|\mathcal{S} \setminus \mathbb{S}^{p-1}|/|\mathbb{S}^{p-1}|\).
THEOREM 3.8. Under Assumption 3.1 there exists \(O_n\) with \(\Pr(O_n) \leq C_0(\gamma, \mu)n^{-1/2}\) and
\[
(3.27) \quad \mathbb{E}\left[I_{O_n}(\langle a_0, \hat{\beta} - \beta \rangle + (n - \hat{d}f)^{-1}z_0^\top(y - X\hat{\beta})^2) \right] \leq R_*C_5(\gamma, \mu)/n
\]
If additionally \(g\) is a seminorm then \(|z_0^\top(y - X\hat{\beta})|/n = |a_0^\top\Sigma^{-1}X^\top(y - X\hat{\beta})|/n \leq g(\Sigma^{-1}a_0)\) always holds by properties of the subdifferential of a norm. Consequently, if \(g(\Sigma^{-1}a_0)^2/R_* \to 0\) then \(\langle a_0, h \rangle^2/R_* \to 0\) and the conclusions of Theorem 3.6 hold.

Theorem 3.8 is proved in Section 7.7. The first part of the theorem says that the estimation error \(\langle a_0, h \rangle\) is essentially \(-\langle z_0, y - X\hat{\beta} \rangle/(n - \hat{d}f)\) up to an error term of order \(R_*O(n^{-1/2})\), so that \(\langle a_0, h \rangle^2/R_* \to 0\) if and only if \((n - \hat{d}f)^{-2}\langle z_0, y - X\hat{\beta} \rangle^2/R_* \to 0\). Combined with the fact that \((1 - \hat{d}f/n)\) is bounded away from 0 by Lemma 7.4, this implies that \(\langle a_0, h \rangle^2/R_* \to 0\) if and only if \((n - \hat{d}f)^{-2}\langle z_0, y - X\hat{\beta} \rangle^2/R_* \to 0\). The last part of the theorem relies on the property \(g(\theta) = \sup_{\theta \in \partial g(u)}u^\top s\) for any norm \(g\) where \(\partial g(u)\) is the subdifferential of \(g\) at \(u\). This property also holds if \(g\) is a semi-norm, however it is unclear how to extend the last part of the above theorem if \(g\) is not a semi-norm.

For the Lasso, the penalty function is \(g(b) = \lambda b_1\) and condition \(g(\Sigma^{-1}a_0)^2/R_* \to 0\) becomes \(\lambda^2\|\Sigma^{-1}a_0\|_1^2/R_* \to 0\). Typically, the tuning parameter \(\lambda\) is chosen as \(\lambda \propto \sigma(2\log(p/k)/\tilde{\sigma}^2)^{1/2}\) with \(k = 1\) [12, among others] or \(k = s_0\) [40, 31, 9, 26, 2], where \(s_0 = \|\beta\|_0\). For such choices, the condition \(\lambda^2\|\Sigma^{-1}a_0\|_1^2/R_* \to 0\) can be written as \(\|\Sigma^{-1}a_0\|_1 = (R_*/\sigma^2)^{1/2}o(\sqrt{n}/\log(p/k))\) and since \(R_* \geq \sigma^2\), a sufficient condition is \(\|\Sigma^{-1}a_0\|_1 = o(\sqrt{n}/\log(p/k))\). If \(a_0 = e_j\) is a vector of the canonical basis, the normalization (1.15) gives \((\Sigma^{-1})_{jj} = 1\) and \(\|\Sigma^{-1}e_j\|_1\) is the \(\ell_1\) norm of the \(j\)-th column of \(\Sigma^{-1}\). The condition \(\|\Sigma^{-1}a_0\|_1 = o(\sqrt{n}/\log(p/k))\) allows, for instance, the \(j\)-th column of \(\Sigma^{-1}\) to have \(o(\sqrt{n}/\log(p/k))\) constant entries. This assumption is weaker than that of some previous studies; for instance [29] requires \(\|\Sigma^{-1}a_0\|_1 = O(1)\) for \(a_0 = e_j\). The following example illustrates the benefit of picking a proper penalty level \(\lambda\).

**EXAMPLE 1.** Let \(p/n \to \gamma < 1\) and \(g(b) = \lambda b_1\).

(i) For \(\gamma = 0\), the Lasso and de-biased Lasso are both identical to the least squares estimator and the de-biasing correction proportional to \(z_0^\top(y - X\hat{\beta})\) is 0 since \(X^\top(y - X\hat{\beta}) = 0\), so that \(\hat{\theta} - \theta = (a_0, h) = a_0^\top(X^\top X)^{-1}X^\top e\) in (3.12), \(\hat{d}f = p\), \(\hat{V}(\theta) \approx \|y - X\beta\|^2 \sim \sigma^2X_{n-p}^\alpha \rightarrow N(0, \sigma^2(1 - \gamma))\).

(ii) Suppose \(|\hat{S}|/n + \|Xh\|^2/n + \|\Sigma^{1/2}h\|^2 = o_P(1)\) for suitable \(\lambda \geq \sigma\sqrt{2\log(p/s_0)/n}\) as in [51, 8]. Then, \(\hat{V}(\theta) = (1 + o_P(1))n\sigma^2\) and \(\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \sigma^2)\).

3.6. **Confidence intervals.** Theorems 3.6 to 3.8 are valid for any choice of the variance estimate among \(\|y - X\hat{\beta}\|^2, \hat{V}(\theta)\) in (3.19) and \(\hat{V}(a_0)\) in (3.22) for directions \(a_0\) such that \(\langle a_0, h \rangle^2/R_* \to 0\) holds. For such direction \(a_0\), the choice \(\|y - X\hat{\beta}\|^2\) leads to the narrowest confidence interval for \(\theta\), namely
\[
(3.28) \quad \Pr\left(\theta \in \left[\langle a_0, \hat{\beta}^{(\text{de-bias})} \rangle \pm z_{\alpha/2}(n - \hat{d}f)^{-1}\|y - X\hat{\beta}\| \right] \right) \rightarrow 1 - \alpha
\]
where \([u \pm v]\) denotes the interval \([u - v, u + v]\), \(\Pr(|N(0, 1)| > z_{\alpha/2}) = \alpha\) and \(\hat{\beta}^{(\text{de-bias})}\) is the de-biased estimator in (3.15). The choice \(\hat{V}(a_0)\) leads to intervals with larger multiplicative coefficient for \(z_{\alpha/2}\), namely
\[
(3.29) \quad \theta \in \left[\langle a_0, \hat{\beta}^{(\text{de-bias})} \rangle \pm z_{\alpha/2}\left(\frac{\|y - X\hat{\beta}\|^2}{(n - \hat{d}f)^2} + \frac{I_n - \hat{H}^2}{(n - \hat{d}f)^4}\langle z_0, y - X\hat{\beta} \rangle^2\right)^{1/2}\right]
\]
has probability converging to $1 - \alpha$ for directions $a_0$ satisfying any of the above theorems. For such directions, the choice $\hat{V}(\theta)$ justifies confidence intervals of the form (1.21) as
\begin{equation}
((n - \tilde{d}f)(\langle a_0, \hat{\beta} \rangle - \theta) + \langle z_0, y - X\hat{\beta} \rangle)^2 - \hat{V}(\theta)z_\alpha^2/2 \leq 0
\end{equation}
holds with probability converging to $1 - \alpha$. Given the expression for $\hat{V}(\theta)$ in (3.19), the left-hand side of (3.30) is a quadratic polynomial in $\theta$ with dominant coefficient $(n - \tilde{d}f)^2 - z_\alpha/2 \parallel I_n - \tilde{H} \parallel_F^2$. Since $\parallel I_n - \tilde{H} \parallel_F^2 \leq n - \tilde{d}f$ almost surely by properties of $\tilde{H}$ in Proposition 7.3 and $n - \tilde{d}f \geq C_\ast(\gamma, \mu)n$ for some constant $C_\ast(\gamma, \mu)$ with probability approaching one by Lemma 7.4, in the same event the dominant coefficient is positive. The intersection of events (3.30) and $\{n - \tilde{d}f \geq C_\ast(\gamma, \mu)n\}$ 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 $\theta$.

3.7. Variance spike. One can pick any choice among the three variance estimates in Theorem 3.6 because it assumes $(a_0, h)^2/R_\ast \to^p 0$ and this limit in probability implies both $\hat{V}(\theta)/\parallel y - X\hat{\beta} \parallel^2 \to^p 1$ and $\hat{V}(a_0)/\parallel y - X\hat{\beta} \parallel^2 \to^p 1$. These limits in probability to 1 are made rigorous by (3.22) and by the lower bound $\parallel y - X\hat{\beta} \parallel^2 \geq R_\ast n (C_\ast^2(\gamma, \mu) - O_p(n^{-1/2}))$ obtained from Lemmas 7.4, 7.6 and 7.7 as explained in (7.38) of the proof.

The reason that the estimates $\hat{V}(\theta)$ and $\hat{V}(a_0)$ were introduced is that the simpler estimate $\parallel y - X\hat{\beta} \parallel^2$ is not asymptotically unbiased for $\text{Var}_0[\xi]$ for directions $a_0$ such that $(a_0, h)^2/R_\ast$ does not converge to 0 in probability: While the relative bias of $\hat{V}(\theta)$ and $\hat{V}(a_0)$ provably converges to 0 in Lemma 3.4 and (3.22) for all directions $a_0$, the same cannot be said for the simpler estimate $\parallel y - X\hat{\beta} \parallel^2$.

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

(i) $\parallel y - X\hat{\beta} \parallel^2/\text{Var}_0[\xi] \to^p 1,
(ii) \mathbb{E}_0[\parallel y - X\hat{\beta} \parallel^2/\text{Var}_0[\xi]] \to^p 1,
(iii) \langle a_0, h \rangle^2/R_\ast \to^p 0,
(iv) \langle a_0, h \rangle^2n/\parallel y - X\hat{\beta} \parallel^2 \to^p 0.
(v) \langle z_0, y - X\hat{\beta} \rangle^2/(n\parallel y - X\hat{\beta} \parallel^2) \to^p 0,
(vi) \hat{V}(\theta)/\parallel y - X\hat{\beta} \parallel^2 \to^p 1,
(vii) \hat{V}(a_0)/\parallel y - X\hat{\beta} \parallel^2 \to^p 1.$

Theorem 3.9 is proved in Section 7.6. It shows that for the directions $a_0$ such that $(a_0, h)^2/n/\parallel y - X\hat{\beta} \parallel^2 \to^p 0$ does not hold, e.g., directions such that the error $(a_0, h)^2$ is of the same order as the average squared residual $\parallel y - X\hat{\beta} \parallel^2/n$ (see Lemma 7.2 in Section 7.2), the simpler estimate $\parallel y - X\hat{\beta} \parallel^2$ fails to account for a non-negligible part of the variance $\text{Var}_0[\xi]$ by item (i) above. The goal of the estimates $\hat{V}(\theta)$ and $\hat{V}(a_0)$ is to repair this as $\mathbb{E}_0[\hat{V}(\theta)]/\text{Var}_0[\xi] \to^p 1$ and $\mathbb{E}_0[\hat{V}(a_0)]/\text{Var}_0[\xi] \to^p 1$ hold for all directions by Lemmas 3.4 and 3.5, even for directions $a_0$ such that (i)-(vii) above fail. Note that the quantity $\langle z_0, y - X\hat{\beta} \rangle^2/(n\parallel y - X\hat{\beta} \parallel^2)$ in item (v) is observable (i.e., does not depend on $\beta$), so that (i)-(vii) are expected to hold when this quantity is sufficiently small.

For directions $a_0$ such that $(a_0, h)^2/n/\parallel y - X\hat{\beta} \parallel^2 \to^p 0$ does not hold, we expect a variance spike, i.e., an extra additive term in the variance estimate equal to $\parallel I_n - \tilde{H} \parallel^2_F(a_0, h)^2$ in $\hat{V}(\theta)$ and to $\parallel I_n - \tilde{H} \parallel^2_F(z_0, y - X\hat{\beta})^2/(n - \tilde{d}f)^2$ in $\hat{V}(a_0)$. 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 $a_0$ with large $(a_0, h)^2/\parallel y - X\hat{\beta} \parallel^2$. The wider confidence interval (3.29) is expected to repair this, although for directions $a_0$ such that $(a_0, h)^2/n\parallel y - X\hat{\beta} \parallel^2 \to^p 0$ 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 \( \hat{V}(\hat{\theta}) \) in (3.20) and of \( \hat{V}(a_0) \) in (3.22), combined with the negative result for the simpler variance estimate \( \| y - X\beta \|^2 \).

Theorem 3.9 as discussed above. Figure 1 in Section 5 illustrates the variance spike on simulations for the Lasso and direction \( a_0 \) proportional to the first canonical basis vector.

The second term in the variance estimates (3.19) and (3.21) is necessary for certain assumptions as shown in the following result. Consider either the Lasso

\[
(3.32) \quad \hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \left\{ \| y - Xb \|^2/(2n) + \lambda \| b \|_1 \right\}
\]

for some \( \lambda > 0 \) or the group Lasso norm \( \| \cdot \|_{GL} \) and group Lasso \( \hat{\beta} \) defined as

\[
(3.33) \quad \hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \left\{ \| y - Xb \|^2/(2n) + \| b \|_{GL} \right\}, \quad \| b \|_{GL} = \sum_{k=1}^{K} \lambda_k \| b_{G_k} \|_2
\]

where \( (G_1, ..., G_K) \) is a partition of \( \{1, ..., p\} \) into \( K \) non-overlapping groups and \( \lambda_1, ..., \lambda_K > 0 \) are tuning parameters. If each \( G_k \) is a singleton and \( \lambda_k = \lambda > 0 \) for all \( k = 1, ..., p \), 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 \( \Sigma \). Assume that the group Lasso estimator \( \hat{\beta} \) in (3.33) satisfies

\[
(3.34) \quad \mathbb{P}(\| \hat{\beta} \|_0 \leq \kappa n/2) \to 1.
\]

If \( a_0 \) is such that \( \| \Sigma^{-1/2} a_0 \| = 1 \) and \( \langle a_0, \hat{\beta} - \beta \rangle^2 / R_* \to 0 \) for the \( R_* \) in (3.23) then

\[
(3.35) \quad \sup_{t \in \mathbb{R}} \mathbb{P}(\| y - X\hat{\beta} \|^2/(n - \hat{d}) \langle a_0, \hat{\beta} - \beta \rangle + z_0^\top(y - X\beta) \leq t) - \Phi(t) \to 0.
\]

Furthermore, for any \( a_p \) with \( a_p \to \infty \) and \( \Sigma \) in (3.25), the relative volume given after (3.25) holds, and the asymptotic normality (3.35) holds uniformly over all \( a_0 \) in \( \Sigma^{1/2} \mathcal{S} \) and uniformly over at least \( (p - \phi_{\mathrm{cond}}(\Sigma)a_p/C^*) \) canonical directions in the sense that \( J_p = \{ j \in [p] : e_j/\| \Sigma^{-1/2} e_j \| \in \Sigma^{1/2} \mathcal{S} \} \) has cardinality \( |J_p| \geq p - \phi_{\mathrm{cond}}(\Sigma)a_p/C^* \).
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 $\|\hat{\beta}\|_0 \leq \kappa n/2$ on the number of non-zero coordinates. Surprisingly no conditions are required on the true regression vector $\beta$ or on the tuning parameters, although these quantities affect whether $\mathbb{P}(\|\beta\|_0 \leq \kappa n/2) \to 1$ 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 de-biasing 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, e.g. [24, 22, 28] or more recently [43, 34, 44, 41], it is assumed that $\Sigma = I_p$ 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 $j$-th component $\hat{\beta}_{j}^{\text{(de-bias)}}$ of $\hat{\beta}^{\text{(de-bias)}}$ in (3.15) for $\Sigma = I_p$, that is,

$$\hat{\beta}_{j}^{\text{(de-bias)}} = \beta_j + (Xe_j, y - X\hat{\beta})/(n - \text{df}).$$

Then, the Lasso has the interpretation as its soft thresholded de-biased version,

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

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

$$H_{n,p}(u, t) = p^{-1} \sum_{j=1}^{p} I\{\sqrt{n} \hat{\beta}_{j}^{\text{(de-bias)}} - \sqrt{n} \beta_j \leq u, \sqrt{n} \beta_j \leq t\},$$

converges in distribution and the second moment to the limit $H$ with independent $N(0, \tau_0)$ and $G$ components, where $\tau_0$ is characterized by a system of non-linear equations with 2 or 3 unknowns. These non-linear equations depend on the loss (here, the $\ell_2$ loss), the penalty (here, the $\ell_1$-norm), the distribution of the noise, as well as the prior distribution that governs the empirical distribution of the coefficients of $\beta$. We note that these works typically assume that $X$ has $N(0, 1/n)$ entries, so that their coefficient vector is equivalent to our $\sqrt{n} \beta$.

For instance, [34] characterizes the limit of the empirical distribution of $(\sqrt{n} \hat{\beta}^{\text{(de-bias)}})$ in terms of two parameters, $\{\tau_\gamma(\lambda), \kappa_\gamma(\lambda)\}$, that are defined as solutions of the non-linear equations in [34, Proposition 3.1]; see also [18, 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 $\langle a_0, \beta \rangle$ for a single, fixed direction $a_0$: Theorem 3.10 does not involve the empirical distributions of $\sqrt{n} \beta$, $\sqrt{n} \hat{\beta}$ or its de-biased version. This contrasts with previous literature on the $n/p \to \gamma$ regime where the confidence interval guarantee holds on average over the coefficients $\{1, ..., p\}$ [24, 22, 28, 41]. This improvement is important in practice: if the practitioner is interested in the effect of a specific effect $j_0 \in \{1, ..., p\}$, it is important to construct confidence intervals with strict type I error control for $\beta_{j_0}$, 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 $\beta$ in the limit.

Surprisingly, Theorems 3.6, 3.7 and 3.10 and their proofs completely bypass solving the non-linear 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 $a_0$ (or a fixed coordinate $j \in \{1, \ldots, p\}$ when $a_0 = e_j$), and the correlations in $\Sigma$ 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 [19, 33] to obtain, for $\Sigma \neq I_p$, asymptotic normality results in an averaged sense over $\{1, \ldots, p\}$. 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 $\hat{H}$ and $w_0$ are available. In this section, when computing gradients with respect to $z_0$ in order to find closed-form expressions for $w_0$ in (3.9), we consider $(XQ_0, \varepsilon)$ fixed as in (3.7). Explicitly, $\nabla \hat{\beta}(z_0)^\top$ is uniquely defined as

\begin{equation}
\hat{\beta} - \hat{\beta} = (\nabla \hat{\beta}(z_0))^\top \eta + o(\|\eta\|)
\end{equation}

where $\hat{\beta} = \arg\min_{b \in \mathbb{R}^p} \|X(b - \beta) - \varepsilon\|^2/(2n) + g(b)$ and $\hat{\beta} = \arg\min_{b \in \mathbb{R}^p} \|X + \eta a_0\| (b - \beta) - \varepsilon\|^2/(2n) + g(b)$. When computing gradients with respect to $y$ in order to find closed-form expressions for $\hat{H}$ in (3.3), we view $\hat{\beta}(y, X) = \arg\min_{b \in \mathbb{R}^p} \|Xb - y\|^2/(2n) + g(b)$ as a function of $(y, X)$ and if $y \mapsto \hat{\beta}(y, X)$ is differentiable at $y$ for a fixed $X$ then

\begin{equation}
\hat{\beta}(\bar{y}, X) - \hat{\beta}(y, X) = [(\partial/\partial y)\hat{\beta}(y, X)](\bar{y} - y) + o(\|\bar{y} - y\|)
\end{equation}

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

4.1. Twice continuously differentiable penalty. The simplest example for which closed-form expressions for $\hat{H}, \hat{f}, w_0$ 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 $h = \hat{\beta} - \beta$ with respect to $(\varepsilon, X)$ exist for almost every $(\varepsilon, X)$ by Rademacher’s theorem. At a point $(\varepsilon, X)$ 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 $X^\top(y - X\hat{\beta}) = X^\top(\varepsilon - Xh) = n\nabla g(\hat{\beta})$. Differentiation with respect to $z_0$ for a fixed $(\varepsilon, XQ_0)$ as in (4.1) gives

\begin{equation}
\{X^\top X + n(\nabla^2 g(\hat{\beta}))\} (\nabla \hat{\beta}(z_0))^\top = a_0(y - X\hat{\beta})^\top - X^\top(a_0, h).
\end{equation}

By the product rule, this provides the derivative of $f(z_0) = Xh - \varepsilon$, namely

\begin{equation}
\nabla f(z_0)^\top = X(X^\top X + 2n\nabla g(\hat{\beta}))^{-1} [a_0(y - X\hat{\beta})^\top - (a_0, h)X^\top] + I_n(a_0, h).
\end{equation}

Regarding $\hat{H}$ involving differentiation with respect to $y$, the Lipschitz condition of the map $y \mapsto \hat{\beta}$ 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 $X$ gives $(X^\top X + n\nabla^2 g)\hat{\beta}(y, X) = X^\top$ so that

\begin{equation}
\hat{H} = (X(\partial/\partial y)\hat{\beta}(y, X))^\top = X(X^\top X + n\nabla^2 g(\hat{\beta}))^{-1} X^\top.
\end{equation}

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

\begin{equation}
w_0 = X\{X^\top X + n\nabla g(\hat{\beta})\}^{-1} a_0.
\end{equation}
4.2. Lasso. Consider the Lasso $\widehat{\beta}$ in (3.32). For $(\varepsilon, X)$ with continuous distribution such as Gaussian under consideration here, almost surely $\widehat{\beta}$ is unique and
\[(4.4) \quad X_S^\top(y - X\widehat{\beta})/n = \lambda \text{sgn}(\widehat{\beta}_S), \quad ||X_S^\top(y - X\widehat{\beta})/n||_\infty < \lambda, \quad \text{rank}(X_S) = |\widehat{S}|,\]
for the Lasso as in [49, 45] and [7, Proposition 3.9], so that the Jacobian of the mapping $(z_0, \varepsilon, XQ_0) \to X\beta$ with respect to $z_0$ and $\varepsilon$ can be computed directly by differentiating the KKT condition as in [45, 7, 8]. The following proposition provides closed-form expressions for the gradients for the Lasso estimator which valid almost surely and require no assumption on the sparsity of $\beta$ or the penalty level.

**Proposition 4.1.** Let $\lambda > 0$ and consider the Lasso $\widehat{\beta}$ in (3.32). Let $\widehat{S} = \{j \in [p] : \beta_j \neq 0\}$. For almost every $(\varepsilon, X) \in \mathbb{R}^{n \times (p+1)}$, there exists a neighborhood of $(\varepsilon, X)$ in which the map $(\varepsilon, X) \mapsto \widehat{\beta}$ is constant, $|\widehat{S}| \leq n$, $X_S^\top X_S$ is invertible and the map $(\varepsilon, X) \mapsto \widehat{\beta}$ is Lipschitz. In this neighborhood, almost surely $[\nabla \widehat{\beta}(z_0)^\top]_{\widehat{S}} = 0 \in \mathbb{R}^{n \times |\widehat{S}|}$,
\[\nabla \widehat{\beta}(z_0)^\top = (X_S^\top X_S)^{-1}(a_0)_\widehat{S}(\beta - y) - X_S^\top(a_0, h) \in \mathbb{R}^{n \times |\widehat{S}|},\]
\[\widehat{H} = X_S(X_S^\top X_S)^{-1}X_S^\top, \text{df} = |\widehat{S}| \text{ and (3.9) holds with } w_0 = X_S(X_S^\top X_S)^{-1}(a_0)_\widehat{S}.\]

**Proof of Proposition 4.1.** Proposition 3.9 in [7] proves, for almost every $(\varepsilon, X)$, uniqueness of $\beta$ and (4.4). Let $(\varepsilon, X) \in \mathbb{R}^{n \times (p+1)}$ be a point at which (4.4) holds. It follows from (4.4) and the Holder continuity of $(\varepsilon, X) \mapsto \varepsilon - X\beta$ established after (7.2) that for almost every $(\varepsilon, X)$ there is an open neighborhood in $\mathbb{R}^{n \times (p+1)}$ in which $\widehat{S} = S$, $\text{sgn}(\widehat{\beta}_S) = s_S$ and $\text{rank}(X_S) = |\widehat{S}|$ are constants, so that $\beta$ is locally equal to $\text{arg min}_{\beta \in \mathbb{R}^p}||y - X_S \beta||^2/(2n) + \lambda s_S^\top b_S$ with the linear penalty $\lambda s_S^\top b_S$. In this neighborhood $(\varepsilon, X) \mapsto \beta$ has the analytic expression
\[\widehat{\beta}_S = (X_S^\top X_S)^{-1}(X_S^\top y - n\lambda s_S), \quad \widehat{\beta}_\overline{S} = 0.\]
Differentiating the above immediately yields the formulas for $\widehat{H}$, $\text{df}$ and $[\nabla \widehat{\beta}(z_0)^\top]_{\widehat{S}}$. For $[\nabla \widehat{\beta}(z_0)^\top]_{\widehat{S}}$, differentiating both sides of $X_S^\top(X_S(\beta - \varepsilon) - \varepsilon) = - n\lambda s_S$ yields
\[(a_0)_S(X_S(\beta - \varepsilon)) + X_S^\top a_0^\top h + X_S^\top X_S[\nabla \widehat{\beta}(z_0)^\top]S = 0\]
due to $X = XQ_0 + z_0 a_0^\top$. Finally, the formula for $w_0$ follows from $(\partial/\partial z_0)(X\beta - y) = X(\partial/\partial z_0)\beta + I_n a_0^\top h$ and simple algebra.

4.3. Group Lasso. Consider a partition $(G_1, \ldots, G_K)$ of $\{1, \ldots, p\}$ and the group Lasso estimator in (3.33). Let $\widehat{B} = \{k \in [K] : ||\beta_{G_k}|| \neq 0\}$ be the set of active groups and $\widehat{S} = \cup_{k \in \widehat{B}} G_k$ the union of all active groups. Define the block diagonal matrix $M = \text{diag}((M_{G_k, G_k}, k \in \widehat{B})) \in \mathbb{R}^{|S| \times |\widehat{S}|}$ by
\[(4.5) \quad M_{G_k, G_k} = n\lambda_k ||\beta_{G_k}||^{-1} \left( I_{G_k} - ||\beta_{G_k}||^{-2} \beta_{G_k} \beta_{G_k}^\top \right), \quad M \in \mathbb{R}^{|S| \times |\widehat{S}|}.\]
The following proposition provides closed-form expressions for the gradients for the Group Lasso estimator and related quantities $\widehat{H}$ and $w_0$ in terms of $\widehat{S}$ and $M$. Its proof is given in Appendix C. Note that the formula for $\widehat{H}$ was known [46].
**Proposition 4.2.** The following holds for for almost every \((\hat{y}, \hat{X}) \in \mathbb{R}^{n \times (1+p)}\). The set \(B = \{k \in [K] : \|\beta_{G_k}\| > 0\}\) of active groups is the same for all minimizers \(\hat{\beta}\) of (3.33) at \((\hat{y}, \hat{X})\) and \(\hat{B} = \hat{B}\) for all \((y, X)\) in a sufficiently small neighborhood of \((\hat{y}, \hat{X})\). If additionally \(X_S^\top X_S\) is invertible where \(S = \bigcup_{k \in \hat{B}} G_k\) then the map \((y, X) \mapsto \hat{\beta}\) is Lipschitz in a sufficiently small neighborhood of \((\hat{y}, \hat{X})\). In this neighborhood we have

\[\nabla \hat{\beta}(z_0)|_{S^c} = 0, \quad \nabla \hat{\beta}(z_0)|_{S} = (X_S^\top X_S + M)^{-1}[(a_0)_S(y - X\hat{\beta})^\top - (a_0, h)X_S^\top],\]

\[\hat{H} = X_S(X_S^\top X_S + M)^{-1}X_S^\top\] and (3.9) holds with \(w_0 = X_S(X_S^\top X_S + M)^{-1}(a_0)_S\).

**5. Simulations: Lasso and variance spike.** Figure 1 illustrates the variance spike phenomenon of Section 3.7 for the Lasso and \(a_0\) 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 \(\beta\) is \(s\)-sparse with \(\beta_1 = 20, \beta_j = \pm 1\) for \(j = 2, \ldots, s\) (independent random signs), \(\beta_j = 0\) for \(j > s\); inverse covariance matrix \(\Sigma^{-1} = I_p + 0.9s^{-1/2}(e_1\text{sgn}(\beta)^\top + \text{sgn}(\beta)e_1^\top)\), direction \(a_0 = e_1/(e_1^\top \Sigma^{-1}e_1)^{1/2}\) for \(e_1 \in \mathbb{R}^p\) the first canonical basis vector. 512 repetitions were used and \((\Sigma, \beta)\) are the same across these repetitions. We see that \(V_0 = \|y - X\beta\|^2\) 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 \((\Sigma, \beta)\) is a minor modification of [8, Example 2.1 and Figure 2]: it is constructed so that \(\langle a_0, \beta - \beta \rangle^2\) captures a substantial fraction of the full prediction error \(\|\Sigma^{1/2}(\beta - \beta)\|^2\).

**6. Simulations: Group Lasso.** Figure 2 illustrates Theorem 3.10 for the Group-Lasso (3.33) with standard normal QQ-plots of \(\langle a_0, \hat{\beta}_{\text{de-bias}} - \beta\rangle / ||y - X\beta||\) for \((n, p, \sigma^2) = \ldots\) and with \(V_0 = \hat{V}(\theta) = ||y - X\hat{\beta}||^2 + (n - df)\langle a_0, \beta - \beta \rangle^2\) (third column), prediction error \(\|\Sigma^{1/2}(\beta - \beta)\|^2\) (fourth column) and squared estimation error in direction \(a_0\) (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.

| \(\lambda\) | \(\langle a_0, \hat{\beta}_{\text{de-bias}} - \beta\rangle / ||y - X\beta||\) | \(\|\Sigma^{1/2}(\beta - \beta)\|^2\) | \(\langle a_0, \beta - \beta \rangle^2\) |
|---|---|---|---|
| 0.005 | \(0.001\) | \(3.23 \pm 0.45\) | \(0.32 \pm 0.11\) |
| 0.01 | \(0.002\) | \(2.62 \pm 0.41\) | \(0.40 \pm 0.12\) |
| 0.05 | \(0.003\) | \(4.39 \pm 0.86\) | \(1.81 \pm 0.39\) |
| 0.1 | \(0.004\) | \(11.70 \pm 2.40\) | \(5.6 \pm 1.14\) |

Fig 1: Standard normal QQ-plots of \(\langle a_0, \hat{\beta}_{\text{de-bias}} - \beta\rangle / \sqrt{V_0}\) with \(V_0 = \|y - X\hat{\beta}\|^2\) (second column) and with \(V_0 = \hat{V}(\theta) = ||y - X\beta||^2 + (n - df)\langle a_0, \beta - \beta \rangle^2\) (third column), prediction error \(\|\Sigma^{1/2}(\beta - \beta)\|^2\) (fourth column) and squared estimation error in direction \(a_0\) (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.
(600,900,2) and the group-Lasso (3.33) with 30 non-overlapping groups each of size 30, where all $\lambda_k$ in (3.33) are equal to a single parameter $\lambda$. The unknown coefficient vector $\beta$ 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 sparsity of $\beta$ is 8, and within these 8 groups all coefficients are equal to 1). The design covariance $\Sigma$ is generated once as $\Sigma = W/(5p)$ where $W$ has Wishart distribution with covariance $I_p$ and $5p$ degrees of freedom. This choice of $(\beta, \Sigma)$ is the same across all 256 repetitions. The direction of interest is $a_0 = e_1/||\Sigma^{-1/2} e_1||$ where $e_1 \in \mathbb{R}^p$ is the first canonical basis vector. The first 8 plots above are standard normal QQ-plots across 256 repetitions for 8 different choices of $\lambda$. The ninth plot shows, for each $\lambda$, boxplots of $\hat{\tau}^2 = (1 - \hat{df}/n)^{-2} ||y - X\beta||^2/n$ 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 $\lambda = 0.138$.

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

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 $\hat{\beta}$ be the estimator (3.1), $\hat{H}$ the gradient of $y \rightarrow X\beta$ as in (1.8), $a_0 \in \mathbb{R}^p$ with $||\Sigma^{-1/2} a_0|| = 1$, $z_0$ and $Q_0$ as defined in (1.13),

\[
\theta = \langle a_0, \beta \rangle, \quad f(z_0) = X\beta - y, \quad \xi_0 = z_0^T f(z_0) - \text{div} f(z_0).
\]

Vector $w_0 \in \mathbb{R}^n$ is given by Lemma 3.1. The oracle $\beta^\star$ and its associated noiseless prediction risk $R_\star$ are given by (2.23). Throughout, $E_0$ denotes the conditional expectation given $(\varepsilon, XQ_0)$ and $\text{Var}_0$ the conditional variance given $(\varepsilon, XQ_0)$. 
7.1. Lipschitzness of regularized least-squares and existence of \( w_0 \). By Rademacher’s theorem, a Lipschitz function \( U \rightarrow \mathbb{R} \) 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 \( (\varepsilon, X) \mapsto \beta \) and \( (\varepsilon, X) \mapsto Xh - \varepsilon \) in certain open set \( U \), and consequently differentiability almost everywhere in \( U \).

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

\[
\tilde{\beta} \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{\| \varepsilon + \mathbf{X} (\beta - b) \|_2^2}{2n} + g(b) \right\}, \quad \tilde{\beta} \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{\| \widetilde{\varepsilon} + \widetilde{\mathbf{X}} (\beta - b) \|_2^2}{2n} + g(b) \right\}
\]

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

\[
\begin{align*}
\n & nD_g(\beta, \tilde{\beta}) + \| \bar{f} - \tilde{f} \|^2 = (\bar{h} - \tilde{h})^\top (X - \tilde{X})^\top f + (\varepsilon - \tilde{\varepsilon})^\top (X - \tilde{X})^\top (f - \tilde{f}) \\
& \quad = \text{trace}([X - \tilde{X}]^\top (f\tilde{h}^\top - f\tilde{h}^\top)) + (\varepsilon - \tilde{\varepsilon})^\top (f - \tilde{f}). \\
\end{align*}
\]

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

\[
\begin{align*}
& nD_g(\beta, \tilde{\beta}) + \| X(\beta - \tilde{\beta}) \|_2^2/2 + \| \tilde{X}(\beta - \tilde{\beta}) \|_2^2/2 \\
& \quad = (\tilde{\beta} - \beta)^\top (X^\top \varepsilon - \tilde{X}^\top \widetilde{\varepsilon}) + (\tilde{\beta} - \beta)^\top (X^\top X - \tilde{X}^\top \tilde{X})(h + \tilde{h})/2 \\
& \quad \leq \| \tilde{\beta} - \beta \|_2 \| \varepsilon - \tilde{\varepsilon} \|_2/2 + \| X + \tilde{X} \|_{op}/2 \\
& \quad + \| \tilde{\beta} - \beta \|_2 \| \varepsilon - \tilde{\varepsilon} \|_2/2 + (\| X \|_{op} + \| \tilde{X} \|_{op})(\| \tilde{h} \| + \| h \|)/2. \\
\end{align*}
\]

If either \( g \) is strongly convex or if there exists a constant \( \pi > 0 \) and a bounded neighborhood \( N \) of \( (\varepsilon, X) \) such that \( \| \beta - \tilde{\beta} \| \leq \| X(\beta - \tilde{\beta}) \| \) for all \( \{(\varepsilon, X), (\varepsilon, \tilde{X})\} \in N \) then the map \( (\varepsilon, X) \mapsto \tilde{\beta} \) is Lipschitz in \( N \).

**Proof of Lemma 7.1.** The KKT conditions for \( \tilde{\beta} \) and \( \beta \) provide

\[
n(\tilde{\beta} - \beta)^\top (\partial g)(\beta) = (\bar{h} - \tilde{h})^\top X^\top f, \quad n(\beta - \tilde{\beta}^\top (\partial g)(\beta) = (h - \tilde{h})^\top X^\top \tilde{f}.
\]

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

\[
nD_g(\beta, \tilde{\beta}) + \| f - \tilde{f} \|^2 = \tilde{h}^\top (X - \tilde{X})^\top f + h^\top (X - \tilde{X})^\top \tilde{f} + (\varepsilon - \tilde{\varepsilon})^\top (f - \tilde{f})
\]

so that (7.2) holds.

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

For (7.3) and (7.4), the KKT conditions for \( \beta \) yield

\[
n(\beta - \tilde{\beta})^\top (\partial g)(\beta) + \| X(\beta - \tilde{\beta}) \|_2^2/2 = (\beta - \tilde{\beta})^\top X^\top (\varepsilon - X(h + \tilde{h}))/2.
\]
Summing the above and its counterpart yields the equality (7.3). Writing $\mathbf{X}^\top \varepsilon - \mathbf{X}^\top \bar{\varepsilon} = (\mathbf{X} + \mathbf{X})^\top (\varepsilon - \bar{\varepsilon})/2 + (\mathbf{X} - \mathbf{X})^\top (\varepsilon + \bar{\varepsilon})/2$ and similarly $\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X} = (\mathbf{X} + \mathbf{X})^\top (\mathbf{X} - \mathbf{X})/2 + (\mathbf{X} - \mathbf{X})^\top (\mathbf{X} + \mathbf{X})/2$, inequality (7.4) follows. To prove the Lipschitz condition in $\mathcal{N}$, we note that for a fixed value of $(\varepsilon, \mathbf{X}, h)$, the right-hand side of (7.4) is linear in $\|\mathbf{h}\|$ while the left-hand side is quadratic in $\|\mathbf{h}\|$ thanks to either strong convexity of $g$ or the assumption on $\pi$. This implies that $\|\mathbf{h}\|$ is bounded uniformly for all $(\varepsilon, \mathbf{X})$ in $\mathcal{N}$. Since $\varepsilon, \mathbf{X}, \bar{\varepsilon}, \mathbf{X}, \|\mathbf{h}\|, \|\mathbf{h}\|$ are all bounded in $\mathcal{N}$, (7.4) divided by $\|\bar{\beta} - \tilde{\beta}\|$ provides the desired Lipschitz property. 

**Lemma 3.1.** Let Assumption 3.1(i) be fulfilled, $a_0 \in \mathbb{R}^p$ and $\bar{\mathbf{H}}$ be as in (3.3). Then, 

$$
\nabla f(z_0) = (I_n - \bar{\mathbf{H}})^\top (a_0, h) + w_0 (y - \mathbf{X}\bar{\beta})^\top
$$

satisfies (3.6) for some random $w_0 \in \mathbb{R}^n$ almost surely. If additionally $\|\Sigma^{-1/2}a_0\| = 1$ then 

$$
\|w_0\|^2 \leq n^{-1} \min \left\{ (4\mu)^{-1}, \phi_{\min}(\Sigma^{-1/2} \mathbf{X} \Sigma^{-1/2}/n)^{-1} \right\}.
$$

**Proof of Lemma 3.1.** Under Assumption 3.1, Lemma 7.1 implies that the map $(\varepsilon, \mathbf{X}) \mapsto f = \mathbf{X}h - \varepsilon$ is Lipschitz in an open neighborhood of almost every point, and thus $\bar{\mathbf{H}}$ and $\nabla f(z_0)$ are defined as Fréchet derivatives almost surely in (3.3) and (3.6) respectively. To prove (3.9), i.e. that the range of $\nabla f(z_0) - \langle a_0, h \rangle (I_n - \bar{\mathbf{H}})$ is the linear span of $f$, we study the directional derivative in a direction $\eta \in \mathbb{R}^n$. For two pairs $(\varepsilon, \mathbf{X})$ and $(\bar{\varepsilon}, \mathbf{X})$ with $\mathbf{X} = \mathbf{X} + t\eta a_0^\top = \mathbf{X}Q_0 + (t\eta + z_0)a_0^\top$ and $\bar{\varepsilon} = \varepsilon + t\eta \langle a_0, h \rangle$, consider the solutions $\tilde{\beta}$ and $\tilde{\beta}$ defined in Lemma 7.1 and $\bar{\phi}_t = \mathbf{X}(\tilde{\beta} - \beta) - \bar{\varepsilon}$ with $\phi_0 = \mathbf{X}(\beta - \beta) - \varepsilon = f$. When the map $(\varepsilon, \mathbf{X}) \mapsto f$ is Fréchet differentiable at $(\varepsilon, \mathbf{X})$,

$$
\lim_{t \to 0^+} \langle \phi_t - \phi_0 \rangle / t = (\nabla f(z_0) - \langle a_0, h \rangle (I_n - \bar{\mathbf{H}}))^\top \eta
$$

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

$$
(\mathbf{X} - \mathbf{X})h + \varepsilon - \bar{\varepsilon} = 0.
$$

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

$$
\mu n |\Sigma^{1/2}(\beta - \beta)|^2 + \|\phi_t - \phi_0\|^2 \leq |\langle a_0, h - \bar{\mathbf{h}} \rangle t \eta^\top f|
$$

due to $\mathbf{X} - \mathbf{X} = t\eta a_0^\top$. Consequently $\phi_t - \phi_0 = 0$ when $\eta^\top f = 0$. This and (7.5) give (3.9). Moreover, for $f \neq 0$, $w_0 = \lim_{t \to 0^+} \langle \phi_t - \phi_0 \rangle / t$ for $\eta = -f / \|f\|^2$, so that (3.10) is an upper bound for $\lim_{t \to 0^+} \|\phi_t - \phi_0\| / t$ in the case of $\|\Sigma^{-1/2}a_0\| = 1 = -\eta^\top f$ where 

$$
\mu n |\Sigma^{1/2}(\beta - \beta)|^2 + \|\phi_t - \phi_0\|^2 \leq |t| |\Sigma^{1/2}(h - \bar{h})|
$$

by the previous display. For $\mu > 0$, the above inequality gives $\|\phi_t - \phi_0\|^2 \leq t^2 (4\mu n)^{-1}$ using $uv \leq u^2/4 + v^2$. For $\mu = 0$, 

$$
\phi_{\min}(\bar{\mathbf{W}}) |\Sigma^{1/2}(h - \bar{h})|^2 \leq |\mathbf{X}^\top (h - \bar{h})|^2 = |\phi_t - \phi_0|^2
$$

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

$\square$
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 $\|w_0\|, \langle a_0, h \rangle, \|Xh\|$ and $\|y - X\beta\|$. To this end, define the random variables $F_+$ and $F$ by

\begin{align}
F_+ & \triangleq (\|g\|^2/n) \lor (\|\varepsilon\|^2/(\sigma^2 n)) \lor (\|\varepsilon - Xh\|^2/(nR_*)) \lor 1
\end{align}

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

\begin{align}
F & \triangleq 2/[1 \land \max\{\mu, \phi_{\min}(\Sigma^{-1/2}(X^T X/n)\Sigma^{-1/2})\}].
\end{align}

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

\begin{align}
\mathbb{E}[F_+] & \lor \mathbb{E}[F^{10}] \leq C(\gamma, \mu), \quad \mathbb{E}[(F_+ - 1)^2] \leq 3\text{Var}[^2_2]/n^2 = 6/n.
\end{align}

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

\begin{align}
\langle a_0, h \rangle & \leq \|\Sigma^{1/2}h\|^2 / (n\|Xh\|^2) \leq F_+ F^2 R_* \quad \text{and} \quad \|w_0\|^2 \leq F/(2n),
\end{align}

\begin{align}
\|y - X\beta\|^2/n & \leq 2F_+ + 2F_+ F^2 R_* \leq 4F_+ F^2 R_*,
\end{align}

for the $w_0$ 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 $\|w_0\|, \langle a_0, h \rangle, \|Xh\|$ and $\|y - X\beta\|$ 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 $\beta$ in (3.1) and $\beta^*, h^*, R_*$ be defined in (3.23). Suppose the penalty satisfies $u^T\{(\partial g) (u + \beta^*) - (\partial g) (\beta^*)\} \geq \mu \|\Sigma^{1/2}u\|^2 \forall u \in \mathbb{R}^p$ with $\mu \in [0, 1/2]$. Let $F_+$ be defined in (7.7) and $F$ any random variable satisfying

\begin{align}
\|\Sigma^{1/2}h\|^2 / (n\|Xh\|^2) & \leq F/2 \quad \text{or} \quad \mu^{-1} = F/2,
\end{align}

for instance (7.8). Then,

\begin{align}
\|\Sigma^{1/2}h\|^2 & \leq F^2_{\max} \{(\bar{\sigma}^2, \|\Sigma^{1/2}h^*\|^2) \leq F_+ F^2 R_*,
\|Xh\|^2/n & \leq \max\{2F\bar{\sigma}^2, \bar{\sigma}^2 + F^2_{\max}\|\Sigma^{1/2}h^*\|^2\} \leq F_+ F^2 R_*,
\end{align}

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

**Proof of Lemma 7.2.** The KKT conditions for $\hat{\beta}$, i.e., $n\partial g(\hat{\beta}) = X^T (y - X\hat{\beta})$, yield

\begin{align}
2(\hat{\beta} - \beta^*)^T (\partial g)(\hat{\beta}) = 2(\hat{\beta} - \beta^*)^T X^T (y - X\hat{\beta})/n = (\|y - X\beta^*\|^2 - \|y - X\hat{\beta}\|^2 - \|X(\hat{\beta} - \beta^*)\|^2)/n
\end{align}

\begin{align}
= (\|Xh^*\|^2 - \|Xh\|^2 - \|X(\hat{\beta} - \beta^*)\|^2 + 2\varepsilon^T X(h - h^*)/n
\end{align}

\begin{align}
\leq (\|Xh^*\|^2 - \|Xh\|^2 + \|\varepsilon\|^2)/n.
\end{align}

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

\begin{align}
2(\beta^* - \hat{\beta})^T (\partial g)(\beta^*) + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 \leq \|\Sigma^{1/2}h\|^2 - \|\Sigma^{1/2}h^*\|^2.
\end{align}

Summing the two above displays yields

\begin{align}
(1 + 2\mu)\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 \leq (F_+ - 1)\|\Sigma^{1/2}h^*\|^2 + \|\Sigma^{1/2}h\|^2 - \|Xh\|^2/n + F_+ \sigma^2
\end{align}

\begin{align}
= \bar{\sigma}^2 + \|\Sigma^{1/2}h\|^2 - \|Xh\|^2/n.
\end{align}
For \( \|\Sigma^{1/2}h\| \geq F\|\Sigma^{1/2}h^*\| \), by the triangle inequality

\[
(7.17) \quad \|\Sigma^{1/2}h\|^2(1 - 1/F)^2 \leq \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2
\]

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

\[
(7.18) \quad \sigma^2 \geq \left\{ \begin{aligned}
(1 - 1/F)^2 + 2/F - 1 \right\} \|\Sigma^{1/2}h\|^2, & \text{ if } \|Xh\|^2/n \geq (2/F)\|\Sigma^{1/2}h\|^2, \\
(1 - 1/F)^2(1 + 2\mu) - 1 \right\} \|\Sigma^{1/2}h\|^2, & \text{ if } \|Xh\|^2/n \leq (2/F)\|\Sigma^{1/2}h\|^2,
\end{aligned} \right.
\]

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

\[
\|Xh\|^2/n \leq \sigma^2 + \|\Sigma^{1/2}h\|^2 \{ 1 - (1 - 1/F)^2 \} \leq \sigma^2 + F^2\sigma^2 \{ 1 - (1 - 1/F)^2 \} = 2F\sigma^2,
\]

and for \( \|\Sigma^{1/2}h\| < F\|\Sigma^{1/2}h^*\| \) we have \( \|Xh\|^2/n \leq \sigma^2 + F^2\|\Sigma^{1/2}h^*\|^2 \) and thus (7.14) holds.

\[\square\]

7.3. Existence and properties of \( \hat{H} \) and \( \hat{df} \).

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

(i) \( \|X(\hat{\beta}(y) - \hat{\beta}(\bar{y}))\| \leq \|y - \bar{y}\| \) for all \( y, \bar{y} \in \mathbb{R}^n \), i.e., \( y \mapsto X\hat{\beta}(y) \) is \( 1 \)-Lipschitz. Its gradient \( \hat{H} \) exists almost everywhere by Rademacher’s theorem, that is, for almost every \( y \) there exists \( \hat{H} \in \mathbb{R}^{n \times n} \) with \( \|\hat{H}\|_p \leq 1 \) such that \( X\hat{\beta}(\bar{y}) = X\hat{\beta}(y) + \hat{H}^\top \eta + o(\|\eta\|) \).

(ii) For almost every \( y \), matrix \( \hat{H} \) is symmetric with eigenvalues in \([0, 1]\). Consequently, with \( \hat{df} = \text{trace}(\hat{H}) \) as degrees of freedom, \( (n - \hat{df})(1 - \hat{df}/n) \leq \|I_n - \hat{H}\|^2_F \leq n - \hat{df} \).

**Proof.** A proof of (i) is given in [6]. For completeness, the argument is the following: by (7.3) with \( \bar{X} = X, \bar{\beta} = \bar{y} - X\beta \) and \( \epsilon = y - X\beta \) we have

\[
(7.19) \quad nD_g(\hat{\beta}(\bar{y}), \hat{\beta}(y)) + \|X\hat{\beta}(\bar{y}) - X\hat{\beta}(y)\|^2 \leq (y - \bar{y})^\top X(\hat{\beta}(y) - \hat{\beta}(\bar{y})).
\]

Using \( D_g(\hat{\beta}(\bar{y}), \hat{\beta}(y)) \geq 0 \) by monotonicity of the subdifferential and the Cauchy-Schwarz inequality yields the desired Lipschitz property. For (ii), define

\[
u(y) = (\|y\|^2 - \|y - X\hat{\beta}(y)\|^2)/2 - ng(\hat{\beta}(y))
\]

\[
= \sup_{b \in \mathbb{R}^p} \{ y^\top Xb - \|Xb\|^2/2 - ng(b) \}.
\]

The function \( u : \mathbb{R}^n \to \mathbb{R} \) is convex in \( y \) as a supremum of affine functions, and \( X\hat{\beta}(y) \) is a subgradient of \( u \) at \( y \). Alexandroff’s theorem as stated in [36, Theorem D.2.1] states that any convex \( u : \mathbb{R}^n \to \mathbb{R} \) is twice differentiable at \( y \) for almost every \( y \) in the following sense: \( u \) is Fréchet differentiable at \( y \) with gradient \( \nabla u(y) \) and there exists a symmetric positive semi-definite matrix \( S \) such that for every \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that for all \( \bar{y} \in \mathbb{R}^n \),

\[
\|y - \bar{y}\| \leq \delta \quad \text{implies} \quad \sup_{v \in \partial u(y)} \|v - \nabla u(y) - S(y - \bar{y})\| \leq \varepsilon\|y - \bar{y}\|.
\]

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

\[\square\]
LEMMA 7.4. Let Assumption 3.1 be fulfilled with \( n \geq 2 \). Then there exists an event \( \Omega_0 \) independent of \((z_0, \varepsilon)\) such that
\[
\Omega_0 \subset \left\{ n - \hat{df} \geq \| \mathbf{I}_n - \hat{H} \|_F^2 \geq C_s(\gamma, \mu) n \right\}
\] with
\[
\begin{cases}
\mathbb{P}(\Omega_0^c) = 0 & \text{if } \gamma < 1, \\
\mathbb{P}(\Omega_0^c) \leq e^{-n/2} & \text{if } \gamma \geq 1,
\end{cases}
\]
where \( C_s(\gamma, \mu) \in (0, 1) \) depends on \( \{\gamma, \mu\} \) only.

PROOF OF LEMMA 7.4. If \( \gamma < 1 \), the choice \( C_s(\gamma, \mu) = (1 - \gamma) \) works with probability one because \( \text{rank}(\hat{H}) \leq \text{rank}(X) \leq p \leq \gamma n \) and \( \| \hat{H} \|_{op} \leq 1 \).

If \( \gamma \geq 1 \) then we have \( \mu \in \Omega_0 \) in Assumption 3.1. Let \( \Omega_0 = \{ \| X Q_0 \Sigma^{-1/2} \|_{op} \leq \sqrt{\theta} + 2\sqrt{n} \} \). By [21, Theorem II.13], \( \mathbb{P}(\Omega_0) \geq 1 - e^{-n/2} \) due to \( X Q_0 \Sigma^{-1/2} = X \Sigma^{-1/2}(I_p - \Sigma^{-1/2} a_0)(\Sigma^{-1/2} a_0)^\top \) with \( \| \Sigma^{-1/2} a_0 \| = 1 \). Next, we hold \( X \) fixed and study the derivatives of \( \hat{\beta} \) with respect to \( y \). Let \( \hat{\beta}(y) \) be as in Proposition 7.3. Let \( P = I_n - z_0 z_0^\top / \| z_0 \|^2 \) be the projection onto \( \{ z_0 \}^\perp \) so that \( P X = P X Q_0 \). Let \( \hat{y}, \tilde{y} \) be such that \( z_0 (\hat{y} - \tilde{y}) = 0 \), or equivalently \( P(y - \hat{y}) = y - \tilde{y} \). By (7.19) and (3.2),
\[
n\mu \| \Sigma^{1/2}(\hat{\beta} - \tilde{\beta})(y) \|^2 + \| X(\hat{\beta} - \tilde{\beta})(y) \|^2 \leq \| y - \tilde{y} \|^2 \| X(\hat{\beta} - \tilde{\beta})(y) \|.
\]
On \( \Omega_0 \), \( \mu(\sqrt{\gamma} + 2)^{-2} \| X Q_0 \hat{\beta}(y) - \tilde{\beta}(y) \|^2 \leq n\mu \| \Sigma^{1/2}(\hat{\beta} - \tilde{\beta})(y) \|^2 \). Combined with the above display, this implies \( (1 + \mu(\sqrt{\gamma} + 2)^{-2}) \| P X(\hat{\beta} - \tilde{\beta})(y) \| \leq \| P(y - \tilde{y}) \| \).

With \( \tilde{y} = y + \eta \) and by definition of \( \hat{H} \) we have \( L^{-1} \| \hat{H} \|_{op} \| \eta \| \leq \| P \| \| \eta \| \) for \( L = (1 + \mu(\sqrt{\gamma} + 2)^{-2})^{-1} \), hence \( \| \hat{H} \|_{op} \leq L \). Since \( \text{rank}(P) = n - 1 \), by Cauchy’s interlacing theorem \( I_n - \hat{H} \) has at least \( n - 1 \) eigenvalues no smaller than \( 1 - L > 0 \). Finally, since \( I_n - \hat{H} \) is symmetric with eigenvalues in \([0, 1]\) by Proposition 7.3,
\[
n - \hat{df} = \text{trace}[I_n - \hat{H}] \geq \| I_n - \hat{H} \|_F^2 \geq (n - 1)(1 - L)^2 \geq n C_s
\]
with \( C_s = (1 - L)^2/2 \) thanks to \( n \geq 2 \).

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

LEMMA 7.5. Let Assumption 3.1 be fulfilled. Then \( \mathbb{E}[\xi_0^2] \leq C_0(\gamma, \mu)n R_* \) and
\[
\mathbb{E}[\|(1 - \hat{df}/n)(a_0, h) + n^{-1}(z_0, y - \hat{X} \hat{\beta})^2\|/R_* \leq C_7(\gamma, \mu)n^{-1},
\]
\[
\mathbb{E}[I_{\Omega_0}(a_0, h) + (n - \hat{df})^{-1}(z_0, y - \hat{X} \hat{\beta})^2\|/R_* \leq C_8(\gamma, \mu)n^{-1},
\]
where \( \Omega_0 \) is the event from Lemma 7.4.

PROOF OF LEMMA 7.5. By (3.9) combined with (3.10), (7.11) and (7.9),
\[
\mathbb{E}[\langle w_0, y - \hat{X} \hat{\beta} \rangle^2] \leq \mathbb{E}[\| w_0 \|^2 \| y - \hat{X} \hat{\beta} \|^2] \leq \mathbb{E}[\| f/(2n) \| 4n F_* F^2] \leq C_9(\gamma, \mu) R_*
\]
Similarly, by definition of \( V^*(\theta) \) in (3.18), \( \mathbb{E}[\xi_0^2] = \mathbb{E}[V^*(\theta)] = \mathbb{E}[\| y - \hat{X} \hat{\beta} \|^2 + \| \nabla f(z_0) \|_F^2] \). Using (7.10)-(7.11) we have \( \mathbb{E}[\| y - \hat{X} \hat{\beta} \|^2 \leq 4n R_* \mathbb{E}[F_* F^2] \) and
\[
\mathbb{E}[\| \nabla f(z_0) \|_F^2] \leq \mathbb{E}[\| I_n - \hat{H} \|_F^2 / z_0 a_0, h^2 + 2 \langle w_0, y - \hat{X} \hat{\beta} \rangle^2]
\]
\[
\leq n \mathbb{E}[\| a_0, h^2] + C_{10}(\gamma, \mu) R_*
\]
\[
\leq n R_* C_{11}(\gamma, \mu)
\]
thanks to $\nabla f(z_0)$ in (3.9), and $(a + b)^2 \leq 2(a^2 + b^2)$ for the first inequality, $\|I_n - \hat{H}\|_F^2 \leq n$ by Proposition 7.3 and (7.22) for the second inequality, and (7.10)-(7.9) for the third inequality. This provides $E[\xi_0^2] \leq C_{12}(\gamma, \mu)nR_s$. Next, (7.20) holds due the bound (7.22) and the relationship in (3.11) between $\xi_0, \langle w_0, y - X\hat{\beta} \rangle$ and the integrand in left-hand side of (7.20). Then (7.21) follows from (7.20) and $I_{\Omega_0}(1 - \hat{d}/n)^{-2} \leq C_s(\gamma, \mu)^{-2}$ by Lemma 7.4.

\textbf{LEMMA 7.6.} Let $\hat{\beta}$ be as in (3.1) for convex $g$ and let $\beta^*, h^*, R_s$ be as in (3.23). Then
\begin{align}
(1 - \hat{d}/n)^2/8 & \leq \|y - X\beta^*\|^2/(nR_s) + \Delta^a_n + \Delta^b_n + \Delta^c_n \tag{7.24} \\
& \leq V^*(\theta)/(nR_s) + \Delta^d_n + \Delta^a_n + \Delta^b_n + \Delta^c_n \tag{7.25}
\end{align}
where $V^*(\theta)$ is defined in (3.18) and $\Delta^a_n, \ldots, \Delta^d_n$ are nonnegative terms defined as
\begin{align}
\Delta^a_n & \overset{\text{def}}{=} \sigma^2((1 - \hat{d}/n) - \varepsilon^\top(y - X\beta))/((n^2\sigma^2))^{1/2}/R_s, \\
\Delta^b_n & \overset{\text{def}}{=} (F_+ - 1)\|y - X\beta^*/(nR_s), \\
\Delta^c_n & \overset{\text{def}}{=} |(1 - \hat{d}/n)\langle a_*, h \rangle - g^\top(X\beta^* - y)/n|^2/R_s \\
\Delta^d_n & \overset{\text{def}}{=} n^{-1}|2w^\top_0(I_n - \hat{H})(y - X\beta^*)\langle a_0, h \rangle|/R_s.
\end{align}
where $g = Xh^*/\|\Sigma^{1/2}h^*\|$ and $a_* = \Sigma h^*/\|\Sigma^{1/2}h^*\|$.  

\textbf{PROOF OF LEMMA 7.6.} By the triangle inequality and definitions of $\Delta^a_n, \Delta^c_n$,
\begin{align}
(1 - \hat{d}/n)\sigma & \leq (e/\sigma)^\top(y - X\beta)/n + (\Delta^a_n R_s)^{1/2}, \\
(1 - \hat{d}/n)\langle a_*, h \rangle & \leq (g^\top(X\beta^* - y))/n + (\Delta^c_n R_s)^{1/2}, \\
(1 - \hat{d}/n)(\sigma^2 + h^\top\Sigma h^*) & \leq (e - Xh^*)^\top(y - X\beta)/n + (\Delta^a_n R_s)^{1/2} R_s + (\Delta^c_n R_s)^{1/2} R_s
\end{align}
where the last line follows from the weighted sum $\sigma(7.29) + \|\Sigma^{1/2}h^*\|(7.30)$ and using $\sigma \vee \|\Sigma^{1/2}h^*\| \leq R_s^{1/2}$ for the last two terms. By the KKT conditions of $\beta$ and $\beta^*$,  
\begin{align}
(\beta^* - \beta)^\top\partial g(\beta^*) &= (h - h^*)^\top\Sigma h^*, \\
(\beta - \beta^*)^\top\partial g(\beta) &= (\beta - \beta^*)^\top X^\top(y - X\beta)/n.
\end{align}
Summing these equalities and using the monotonicity of the subdifferential yields
\begin{align}
\|\Sigma^{1/2}h^*\|^2 + \|y - X\beta^*\|^2/n & \leq h^\top\Sigma h^* + (\beta - \beta^*)^\top y - X\beta)/n + \|y - X\beta^*\|^2/n \\
& = h^\top\Sigma h^* + (y - X\beta^*)^\top(y - X\beta)/n.
\end{align}
Combining (7.31) multiplied by $1 - \hat{d}/n$ with the line after (7.30) gives
\begin{align}
& (1 - \hat{d}/n)(R_s + \|y - X\beta^*\|^2/n) \\
& \leq (2 - \hat{d}/n)(y - X\beta^*)^\top(y - X\beta)/n + (\Delta^a_n)^{1/2} R_s + (\Delta^c_n)^{1/2} R_s \\
& \leq 2\|y - X\beta^*\|^2/n + 2\max\{\Delta^a_n, \Delta^c_n\}^{1/2} R_s
\end{align}
using the Cauchy-Schwarz inequality and $(2 - \hat{d}/n) \leq 2$ for the last inequality. Using $(2a + 2b)^2 \leq 8(a^2 + b^2)$ for the right-hand side with $\|y - X\beta^*\|^2\|y - X\beta^*\|^2/(n^2 R_s^2) \leq \|y - X\beta^*\|^2/(nR_s) + \Delta^b_n$ completes the proof of (7.24). The second inequality, (7.25), then follows from (3.9), (3.18) and
\begin{align}
& \text{trace}\left[\left(I_n - \hat{H}\right)(a_0, h) + w_0(y - X\beta)^2\right] \\
& = \|I_n - \hat{H}\|_F^2\langle a_0, h \rangle + (w_0(y - X\beta))^2 + 2w_0^\top(I_n - \hat{H})(y - X\beta)\langle a_0, h \rangle
\end{align}
which implies $V^*(\theta) - \|y - X\beta\|^2 \geq -n\Delta^d_n R_s$. \hfill \Box
Lemma 7.7. Define $\Delta_n \overset{df}{=} \Delta_n^a + \Delta_n^b + \Delta_n^c + \Delta_n^d$ where $\Delta_n^a, \ldots, \Delta_n^d$ are defined in Lemma 7.6. Under Assumption 3.1 we have $E[\Delta_n] \leq C(\gamma, \mu)n^{-1/2}$.

Proof of Lemma 7.7. We bound each of $\Delta_n^a, \Delta_n^b, \Delta_n^c, \Delta_n^d$ separately. We have $\Delta_n^b \leq (F_+ - 1)4F_+F^2$ by (7.11) so that $E[\Delta_n^b] \leq C_{13}(\gamma, \mu)n^{-1/2}$ by virtue of (7.9). For $\Delta_n^a$ we have $\Delta_n^a = n^{-2}\sigma^2(n - \hat{d}f) - \epsilon^T(y - \hat{X}\beta)^2/R_s$. By the Second Order Stein formula (Proposition 2.1) with respect to $\epsilon$ conditioned on $X$,

$$E[\Delta_n^a] = \frac{1}{n^2}E[\|y - \hat{X}\beta\|_2^2/R_s + \sigma^2\text{trace}(\{I_n - \hat{H}\}^2)/R_s] \leq \frac{1}{n^2}E[4F_+F^2 + 1]$$

where we used $\text{trace}(\{I_n - \hat{H}\}^2) \leq n$ from Proposition 7.3 and (7.11) for the inequality. Thanks to (7.9), this shows that $E[\Delta_n^a] \leq n^{-1}C_{14}(\gamma, \mu)$. Similarly for $\Delta_n^d$ in (7.25), $\Delta_n^d \leq 2n^{-1}\|w_0\|\|y - \hat{X}\beta\|\|a_0, h\|/R_s \leq n^{-1/2}(F/2)^{1/2}2F_+F^2$ hence $E[\Delta_n^d] \leq n^{-1}C_{15}(\gamma, \mu)$ by (7.11) and (7.9). For $\Delta_n^c$ we have $g = z_0$ for $a_0 = a_*$ so that $E[\Delta_n^c] \leq C_{16}(\gamma, \mu)n^{-1}$ by (7.20).

7.5. Event $\Omega_n$. With $\Omega_0, C_s(\gamma, \mu)$ in Lemma 7.4 and $\Delta_n$ in Lemma 7.7, let

$$\Omega_n = \Omega_0 \cap \{E_0[\Delta_n] \vee \Delta_n \leq C_s^2(\gamma, \mu)/16\}.$$  

By the union bound, Markov’s inequality and the bound on $E[\Delta_n]$ in Lemma 7.7,

$$P(\Omega_n^c) \leq P(\Omega_0^c) + C_{17}(\gamma, \mu)n^{-1/2} \leq C_0n^{-1/2}$$

thanks to $P(\Omega_0^c) \leq e^{-n^2/2}$ in Lemma 7.4 by (7.24),

$$\Omega_n \subset \{\|y - \hat{X}\beta\|_2^2 \geq R_s nC_s^2(\gamma, \mu)/16\}.$$  

Since $\Omega_0^c$ is independent of $z_0$, taking the condition expectation $E_0$ of (7.25) in $\Omega_0$ gives

$$\Omega_n \subset \{\text{Var}_0[\epsilon_0] \wedge E_0[\|y - \hat{X}\beta\|^2] \geq R_s nC_s^2(\gamma, \mu)/16\}.$$  

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

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

$$E[I_{\Omega_n} \langle w_0, y - \hat{X}\beta \rangle^2/\text{Var}_0[\epsilon_0]] \leq C_{18}(\gamma, \mu)n^{-1}.$$  

Proof of Lemma 3.2. With $\Omega_n$ in (7.32), (3.14) follows from (7.35) and (7.22).

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

$$E[I_{\Omega_n} \left| E_0[\hat{V}(\theta)] - 1 \right| \left| \text{Var}_0[\epsilon_0] \right|] \leq C_{19}(\gamma, \mu)n^{-1}.$$  

Proof of Lemma 3.4. Let $\Omega_n$ be as in (7.32). The first inequality in (3.20) follows from the triangle inequality. By (7.35) we have $E[I_{\Omega_n} E_0[\hat{V}(\theta) - V^*(\theta)]/\text{Var}_0[\epsilon_0]] \leq E[|\hat{V}(\theta) - V^*(\theta)|]16/(nR_s C_s^2(\gamma, \mu))$. With $V^*(\theta), \hat{V}(\theta)$ in (3.18)-(3.19) and $\nabla f(z_0)\top$ in (3.9),

$$V^*(\theta) - \hat{V}(\theta) = \langle w_0, y - \hat{X}\beta \rangle^2 + 2w_0\top(I_n - \hat{H})(y - \hat{X}\beta)(a_0, h).$$

Using $\|I_n - \hat{H}\|_{op} \leq 1$ from Proposition 7.3 and (7.10)-(7.11) we find by the Cauchy-Schwarz inequality $|V^*(\theta) - \hat{V}(\theta)| \leq (2 + 2\sqrt{2})R_s F_+ F^3$. 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 $\Omega_n$ with $\mathbb{P}(\Omega_n^c) \leq C_0(\gamma, \mu)n^{-1/2}$ and

$$\max \left\{ \mathbb{E} \left[ I_{\Omega_n} \left| \frac{V(a_0)}{V(\theta)} \right|_2^2 \right], \mathbb{E} \left[ I_{\Omega_n} \frac{V(a_0)}{\mathbb{E}[V(\theta)]} \right|_2^2 \right\} \leq C_2(\gamma, \mu) \frac{n}{n}. $$

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

$$|V(a_0)| - V(\theta)| \leq \|I_n - \mathcal{F} \|_F (a_0, h) + (n - \bar{d})^{-1} \langle z_0, y - X \bar{\beta} \rangle. $$

Let $\Omega_n$ be as in (7.32). Using $\bar{V}(\theta) \geq \|y - X \bar{\beta}\|_2^2$, the lower bound (7.34) and $\|I_n - \mathcal{F}\|_F^2 \leq n$ by Proposition 7.3,

$$\mathbb{E}[I_{\Omega_n} |V(a_0)|] \leq \mathbb{E}[I_{\Omega_n} (a_0, h) + (n - \bar{d})^{-1} \langle z_0, y - X \bar{\beta} \rangle^2] + 6/(R_\gamma C_2(\gamma, \mu)$$

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 $\mathbb{E}[\cdot]^{1/2}$, we have $\mathbb{E}[V(a_0)]^2 - \mathbb{E}[\bar{V}(\theta)]^2 \leq \mathbb{E}[[\bar{V}(a_0) - V(\theta)]^2]^{1/2}$. The proof is completed by using again (7.36), the lower bound (7.35) on $\mathbb{E}[\|y - X \bar{\beta}\|_2^2]$ in $\Omega_n$ and the same argument as for the first term in the maximum.

\textbf{THEOREM 3.9.} Let Assumption 3.1 be fulfilled. Then the following are equivalent:

(i) $\|y - X \bar{\beta}\|_2^2/\text{Var}_0[\xi_0] \to^p 1$,
(ii) $\text{Var}_0[\|y - X \bar{\beta}\|_2^2/\text{Var}_0[\xi_0]] \to^p 1$,
(iii) $\langle a_0, h \rangle^2/R_s \to^p 0$,
(iv) $\langle a_0, h \rangle^2/n \to^p 0$.
(v) $\langle y, y - X \bar{\beta} \rangle^2/(n\|y - X \bar{\beta}\|_2^2) \to^p 0$.
(vi) $\bar{V}(\theta)/\|y - X \bar{\beta}\|_2^2 \to^p 1$.
(vii) $\bar{V}(a_0)/\|y - X \bar{\beta}\|_2^2 \to^p 1$.

\textbf{PROOF OF THEOREM 3.9.} (v) $\iff$ (iv) due to $C_s(\gamma, \mu)n \leq \|I_n - \mathcal{F}\|_F^2 \leq n$ in $\Omega_n$ by Lemma 7.4 and Proposition 7.3 combined with (3.17).

(iv) $\iff$ (vi) follows from $C_s(\gamma, \mu)n \leq \|I_n - \mathcal{F}\|_F^2 \leq n$ in $\Omega_n$.

(vi) $\iff$ (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 $\|y - X \bar{\beta}\|^2/(nR_s) = O_p(1)$ by (7.11) and (7.9).

(iii) $\Rightarrow$ $\delta^2 \to^p 0$ was shown in the proof of Theorem 3.6, and $\delta^2 \to^p 0$ implies $\mathbb{E}[\|\nabla f(z_0)\|_2^2]/\mathbb{E}[\|y - X \bar{\beta}\|^2] \to^p 0$ and $\mathbb{E}[\text{trace}((\nabla f(z_0))^2)]/\mathbb{E}[\|y - X \bar{\beta}\|^2] \to^p 0$ so that (iii) $\Rightarrow$ (ii) holds.

By Lemma 3.4, (ii) implies that $\mathbb{E}_0[\bar{V}(\theta)]/\mathbb{E}_0[\|y - X \bar{\beta}\|^2] - 1 = \mathbb{E}_0[\|I_n - \mathcal{F}\|^2f(a_0, h)^2]/\mathbb{E}_0[\|y - X \bar{\beta}\|^2]$ converges to 0 in probability. Since $\mathbb{E}_0[\|y - X \bar{\beta}\|^2]/(nR_s) = O_p(1)$ by (7.11) and (7.9), combined with $\|I_n - \mathcal{F}\|_F^2 \geq C_2^2(\gamma, \mu)n$ in $\Omega_0$ by Lemma 7.4, this implies $\mathbb{E}_0[I_{\Omega_0}f(a_0, h)^2]/R_s \to^p 0$ as $\Omega_0$ is independent of $z_0$. Thus (ii) implies (iii) by Markov’s inequality with respect to $\mathbb{E}_0$.

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

$$\text{Var}_0[\|y - X \bar{\beta}\|^2] \leq \mathbb{E}_0[\|\nabla f(z_0)\|_2^2/\mathbb{E}_0[\|y - X \bar{\beta}\|^2] \leq \mathbb{E}_0[\|\nabla f(z_0)\|_2^2/\|y - X \bar{\beta}\|^2].$$

With $\nabla f(z_0) \in (3.9)$ and the bounds (7.10)-(7.11) we have $\|y - X \bar{\beta}\|^2 \leq 4nF^2R_s$ and $\|\nabla f(z_0)\|_2 \leq 2(2F+3) + F^2R_s$ thanks to $\|I_n - \mathcal{F}\|_op \leq 1$ by Proposition 7.3. Combined with the lower bound (7.35) on $\text{Var}_0[\xi_0]$ and the moment upper bounds (7.9) we obtain $\mathbb{E}[I_{\Omega_n} \text{Var}_0[\|y - X \bar{\beta}\|^2]/\text{Var}_0[\xi_0]^2] \leq C_2(\gamma, \mu)n^{-1/2}$ which gives (vii) $\iff$ (i).
7.7. Proofs of Theorem 3.3 and asymptotic normality results.

**THEOREM 3.3.** Under Assumption 3.1 there exists \( \Omega_n \) with \( \mathbb{P}(\Omega_n^c) \leq C_0(\gamma, \mu)n^{-1/2} \) and
\[
\mathbb{E}[\Omega_n(n - \hat{d})^2(a_0, \beta^{(de-bias)} - \beta)^2/\|y - X\beta\|^2] \leq C_{22}(\gamma, \mu).
\]
Furthermore \(|(a_0, \beta^{(de-bias)} - \beta)| = O_P(1)\|y - X\beta\|/(n - \hat{d}) = O_P(1)\|y - X\beta\|/n.

**PROOF OF THEOREM 3.3.** Let \( \Omega_n \) be as in (7.32). Since \( I_{\Omega_n}\|y - X\beta\|^{-2} \leq 16/(C_2^2(\gamma, \mu)nR_*) \) 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 \(|(a_0, \beta^{(de-bias)} - \beta)| = O_P(1)\|y - X\beta\|/(n - \hat{d}) \), and \( I_{\Omega_n}(1 - \hat{d}/n)^{-1} \leq C_*(\gamma, \mu)^{-1} \) for \( \Omega_0 \) in Lemma 7.4 provides \((1 - \hat{d}/n) = O_P(1)\).

**THEOREM 3.6.** Let Assumption 3.1 be fulfilled. Let \( \hat{\beta}^{(de-bias)} \) be as in (3.15). Then, for any \( a_0 \) with \( \|\Sigma^{-1/2}a_0\| = 1 \) such that \( \langle a_0, h \rangle^2/R_* \to^{\mathbb{P}} 0 \),
\[
\sup_{t \in \mathbb{R}} \left[ \mathbb{P}\left( \frac{\xi_0}{V_0^{1/2}} \leq t - \Phi(t) \right) + \mathbb{P}\left( \frac{\langle a_0, \hat{\beta}^{(de-bias)} - \theta \rangle - \delta}{V_0^{1/2}/(n - \hat{d})} \leq t \right) - \Phi(t) \right] \to 0,
\]
where \( V_0 \) denotes any of the four quantities: \( \text{Var}_{\xi_0}[\xi], \|y - X\beta\|^2, \hat{V}(\theta) \) or \( \hat{V}(a_0) \).

**PROOF OF THEOREM 3.6.** Let \( \Omega_n \) be as in (7.32). Let \( \delta_1^2 \) be the quantity in (3.24), omitting the dependence in \( a_0 \) as it is clear from context. Since \( \delta_1^2 \leq 1 \) by definition, \( \mathbb{E}[\delta_1^2] \leq \mathbb{E}[\Omega_n, \delta_1^2] + \mathbb{P}(\Omega_n^c) \). In \( \Omega_n \), (7.35) provides a lower bound on the denominator of \( \delta_1^2 \) so that \( \mathbb{E}[\Omega_n, \delta_1^2] \leq \mathbb{E}[\|\nabla f(z_0)\|^{1/2}]16/(nR_*C_2^2(\gamma, \mu)) \). By (7.23) and the bound (7.33) on \( \mathbb{P}(\Omega_n^c) \), we obtain
\[
(7.37) \quad \mathbb{E}[\delta_1^2] \leq \mathbb{E}[\Omega_n, \delta_1^2] + \mathbb{P}(\Omega_n) \leq C_{23}(\gamma, \mu)\left( \mathbb{E}[\langle a_0, h \rangle^2/R_*] + n^{-1} \right) + C_0(\gamma, \mu)n^{-1/2}
\]
Furthermore, \( \langle a_0, h \rangle^2/R_* \) is bounded in \( L_2 \) thanks to \( \mathbb{E}[\langle a_0, h \rangle^4/R_*^2] \leq \mathbb{E}[F^4_\gamma F^4] \leq C_{24}(\gamma, \mu) \) by (7.10) and (7.9). Since a sequence of random variables uniformly bounded in \( L_2 \) is uniformly integrable, the assumption \( \langle a_0, h \rangle^2/R_* \to^{\mathbb{P}} 0 \) implies \( \mathbb{E}[\langle a_0, h \rangle^2/R_*] \to 0 \) and thus \( \mathbb{E}[\xi_0^{1/2}/R_*] \to 0 \). This completes the proof that \( \mathbb{E}[\delta_1^2] \to 0 \) and that \( \xi_0/\text{Var}_{\xi_0}[\xi_0] \to^{-d} N(0, 1) \) by Theorem 2.2. Next, by (3.16), \( (n - \hat{d})\langle a_0, \beta^{(de-bias)} - \beta \rangle/\text{Var}_{\xi_0}[\xi_0]^{1/2} \to^{d} N(0, 1) \) also holds. It remains to prove \( V_0/\text{Var}_{\xi_0}[\xi_0] \to^{\mathbb{P}} 1 \) for all four possible choices for \( V_0 \). By (2.7), \( \mathbb{E}[\delta_1^2] \to 0 \) implies \( \|y - X\beta\|^2/\text{Var}_{\xi_0}[\xi_0] \to^{\mathbb{P}} 1 \), while
\[
(7.38) \quad 0 \leq \frac{\hat{V}(\theta)}{\|y - X\beta\|^2} - 1 = \frac{\|\hat{H} - I_n\|_F^2 \langle a_0, h \rangle^2/(nR_*)}{\|y - X\beta\|^2/(nR_*)}.
\]
Proposition 7.3 provides \( \|\hat{H} - I_n\|_F^2 \leq n \) so that the numerator converges to 0 in probability thanks to assumption \( \langle a_0, h \rangle^2/R_* \to^{\mathbb{P}} 0 \). The denominator is bounded from below by \( C_2^2(\gamma, \mu)/16 \) in \( \Omega_n \) by (7.34) and \( \mathbb{P}(\Omega_n) \to 1 \). This proves \( \hat{V}(\theta)/\text{Var}_{\xi_0}[\xi_0] \to^{\mathbb{P}} 1 \) and \( \hat{V}(a_0)/\text{Var}_{\xi_0}[\xi_0] \to^{\mathbb{P}} 1 \) follows by Lemma 3.5. Slutsky’s theorem completes the proof as \( V_0/\text{Var}_{\xi_0}[\xi_0] \to^{\mathbb{P}} 1 \) 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, \( \beta^{(de-bias)} \) be as in (3.15). Then for any increasing sequence \( a_p \to +\infty \) (e.g., \( a_p = \log \log p \)), the subset
\[
(3.25) \quad \mathcal{S} = \left\{ v \in S^{p-1} : \mathbb{E}[\|\Sigma^{1/2}v, h\|^2/\|\Sigma^{1/2}h\|^2] \leq C^*/a_p \right\}
\]
of the unit sphere $S^{p-1}$ in $\mathbb{R}^p$ has relative volume $|\mathcal{S}|/|S^{p-1}| \geq 1 - 2e^{-p/v_p}$ and

$$
(3.26) \sup_{a_0 \in \Sigma^{1/2} \mathcal{S}} \sup_{t \in \mathbb{R}} \left[ \left| \mathbb{P} \left( \frac{\xi_0}{V_0^{1/2}} \leq t \right) - \Phi(t) \right| + \left| \mathbb{P} \left( \frac{\langle a_0, \beta^{(\text{de-bias})} \rangle}{V_0^{1/2}/(n - d\bar{f})} \leq t \right) - \Phi(t) \right| \right] \to 0 \tag{3.26}
$$

where $V_0$ denotes any of the four quantities: $\text{Var}_0[\xi_0]$, $\|y - X\hat{\beta}\|^2$, $\tilde{V}(\theta)$ or $\tilde{V}(a_0)$. Furthermore, with $e_j \in \mathbb{R}^p$ the $j$-th canonical basis vector and $\phi_{\text{cond}}(\Sigma) = \|\Sigma\|_{op}\Sigma^{-1/2}\|\Sigma\|_{op}$, the asymptotic normality in (3.26) uniformly holds at least $(p - \phi_{\text{cond}}(\Sigma)a_p/C^*)$ canonical directions in the sense that $J_p = \{ j \in [p] : e_j/\|\Sigma^{-1/2}e_j\| \in \Sigma^{1/2}\mathcal{S} \}$ has cardinality $|J_p| \geq p - \phi_{\text{cond}}(\Sigma)a_p/C^*$.

**Proof of Theorem 3.7.** We construct a subset $\mathcal{S}$ of the sphere such that $\langle a_0, h \rangle^2/R_s \rightarrow^p 0$ uniformly over all $a_0 \in \Sigma^{1/2}\mathcal{S}$. Let $v$ be uniformly distributed on the unit Euclidean sphere $S^{p-1}$, independently of $(X, y)$, and denote by $\nu$ its probability measure. The vector $\sqrt{p}v$ is subgaussian in $\mathbb{R}^p$ [48, Theorem 3.4.6], in the sense that for any non-zero vector $u \in \mathbb{R}^p$, $\int \exp \left( \frac{\langle \sqrt{p}v, u \rangle^2}{C^*\|u\|^2} \right) d\nu(v) \leq 2$ for some absolute constant $C^* > 0$. By Jensen’s inequality and Fubini’s Theorem,

$$
\int \exp \left\{ \frac{\langle v, \Sigma^{1/2}h \rangle^2}{C^*\|\Sigma^{1/2}h\|^2} \right\} d\nu(v) \leq \int \exp \left\{ \frac{\langle \sqrt{p}v, h \rangle^2}{C^*\|\Sigma^{1/2}h\|^2} \right\} d\nu(v) \leq 2.
$$

Hence by Markov’s inequality, for any positive $x$, $\nu(\{ v \in S^{p-1} : \mathbb{E}[\langle v, \Sigma^{1/2}h \rangle^2/\|\Sigma^{1/2}h\|^2 > C^*x/p] \}) \leq 2e^{-x}$. Setting $x = p/a_p$, we obtain that the subset $\mathcal{S} \subset S^{p-1}$ defined by (3.25) has relative volume at least $|\mathcal{S}|/|S^{p-1}| \geq 1 - 2e^{-p/v_p}$, and for all $a_0 \in \Sigma^{1/2}\mathcal{S}$,

$$
(7.39) \mathbb{E}[\langle a_0, h \rangle^2/\|\Sigma^{1/2}h\|^2] \leq C^*/a_p.
$$

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

$$
\sum_{j=1}^p \frac{1}{\|\Sigma^{-1/2}e_j\|^2} \mathbb{E}[\langle e_j, h \rangle^2/\|\Sigma^{1/2}h\|^2] \leq \|\Sigma\|_{op}\mathbb{E}[\|h\|^2/\|\Sigma^{1/2}h\|^2] \leq \phi_{\text{cond}}(\Sigma).
$$

To show that $\sup_{a_0 \in \Sigma^{1/2}\mathcal{S}} \mathbb{E}[\delta^2_0(a_0)] \to 0$, thanks to (7.37) it is enough to prove that $\mathbb{E}[\langle a_0, h \rangle^2/R_s] \to 0$ uniformly over $a_0 \in \Sigma^{1/2}\mathcal{S}$. By the Cauchy-Schwarz inequality,

$$
\mathbb{E}[\langle a_0, h \rangle^2/R_s] = \mathbb{E}[\langle a_0, h \rangle\|\Sigma^{1/2}h\|/R_s \langle a_0, h \rangle/\|\Sigma^{1/2}h\|] \leq \mathbb{E}[\|\Sigma^{1/2}h\|^4/R_s^2(\Phi_{cap})^{1/2}]
$$

for any $a_0 \in \Sigma^{1/2}\mathcal{S}$ thanks to (7.39), while $\mathbb{E}[\|\Sigma^{1/2}h\|^4/R_s^2] \leq \mathbb{E}[F^2F^4] \leq C_{25}(\gamma, \mu)$ by (7.10) and (7.9). This implies $\sup_{a_0 \in \Sigma^{1/2}\mathcal{S}} \mathbb{E}[\langle a_0, h \rangle^2/R_s] \to 0$ and $\sup_{a_0 \in \Sigma^{1/2}\mathcal{S}} \mathbb{E}[\delta^2_0(a_0)] \to 0$ hold.

By Theorem 2.2 this shows that $\xi_0/\text{Var}_0[\xi_0]^{1/2} \rightarrow^d N(0, 1)$ uniformly over $a_0 \in \Sigma^{1/2}\mathcal{S}$. Since the bounds (3.14), (7.38) are all uniform over all $a_0$ with $\|\Sigma^{-1/2}a_0\| = 1$, Slutzky’s theorem implies $V_0/\text{Var}_0[\xi_0] \rightarrow^p 1$, $\xi_0/V_0^{1/2} \rightarrow^d N(0, 1)$ and $(n - d\bar{f})(a_0, h) + z_0^2(y - \hat{X}\hat{\beta})$ has Kolmogorov distance follows from convergence in distribution.


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

\begin{equation}
\mathbb{E}[I_{\Omega_n}(a_0, \beta - \beta) + (n - d\tilde{f})^{-1} z_0^T(y - X\beta)^2] \leq R_c C_{26}(\gamma, \mu) / n
\end{equation}

If additionally $g$ is a seminorm then $|z_0^T(y - X\beta)|/n = |a_0^T \Sigma^{-1} X^T(y - X\beta)|/n \leq g(\Sigma^{-1}a_0)$ always holds by properties of the subdifferential of a norm. Consequently, if $g(\Sigma^{-1}a_0)^2/R_c \to 0$ then $(a_0, h)^2/R_c \to 0$ 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 $\beta$, $|z_0^T(y - X\beta)| = n |(\Sigma^{-1}a_0)^T \partial g(\beta)| \leq ng(\Sigma^{-1}a_0)$ since for a norm $g(u) = \sup_{v \in g(u)} (u, v)$.

Appendix A: Integrability of $\phi_{\min}^{-1}(X \Sigma^{-1/2}/\sqrt{n})$ When $p/n \to \gamma \in (0, 1)$

In our regression model with Gaussian covariates, the matrix $X \Sigma^{-1/2}$ has iid $N(0, 1)$ 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 $G$ be a matrix with $n$ rows, $p$ columns and iid $N(0, 1)$ entries. Then $G^T G$ 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 $n, p$, 

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

Proof. Throughout the proof, $p = p_n$ is an implicit function of $n$; we omit the subscript for brevity. Since $S_n = \phi_{\min}(G^T G/n) \to (1 - \sqrt{\gamma})^2$ almost surely (cf. [37]), it is enough to show that the sequence of random variables $(S_n^{-1})_{n \geq n_0}$ is uniformly integrable for some $n_0 > 0$, i.e., that $\sup_{n \geq n_0} \mathbb{E}[S_n^{-k} I_{|S_n| \leq \epsilon}] \to 0$ as $\epsilon \to 0$. For uniform integrability, we use the following argument from [23, Section 5]. The matrix $G^T G$ is a Wishart matrix and the density of $L = \phi_{\min}(G^T G)$ satisfies for $\lambda \geq 0$, 

\[ f_L(\lambda) = \frac{\sqrt{\pi}}{\Gamma\left(\frac{n+1}{2}\right)} \left(\frac{\lambda^{n-p-1/2}}{\Gamma\left(\frac{n-p+1}{2}\right)}\right)^{n-p+1/2} e^{-\lambda/2} \] 

cf. [23, Section 5]. The density of $S_n = L/n = \phi_{\min}(G^T G/n)$ that we are interested in, is given by $f_{S_n}(x) = n f_L(nx)$ for $x \geq 0$. Hence if $0 < \epsilon < (1 - \gamma)/2$, 

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

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

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

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

\[ \sup_{n \geq n_2} \frac{\log(\Lambda_n)}{(n - p + 1)/2} \leq C_2(\gamma) \]
because the main terms (coming from $x^2$ in Stirling’s formula) cancel each other. Then for any $n \geq n_1 \lor n_2$,
\[
\mathbb{E}[S_n^{-k} I_{(S_n < \epsilon)}] \leq (\exp(C_2(\gamma)) e^{(n-p+1)/2} e^{-k} e^{-ne/2} (\exp(C_2(\gamma))) e^{(n-p+1)/2-k} e^{kC_2(\gamma)-ne/2}.
\] (A.2)

For $n \geq n_1$, (A.1) holds and if $\epsilon < (\exp(C_2(\gamma))^{-1}$ we have
\[
\sup_{n \geq n_1 \lor n_2} \mathbb{E}[S_n^{-k} I_{(S_n < \epsilon)}] \leq (\exp(C_2(\gamma)) e^{(n_1(1-\gamma)/4} e^{kC_2(\gamma)}
\]
which converges to 0 as $\epsilon \to 0$. This shows uniform integrability of the sequence and proves the claim. \hfill \Box

\section*{Appendix B: Proof: $p > n$ Without Strong Convexity}

\subsection*{Lemma B.1.}

Let $\beta \in \mathbb{R}^p$ and assume that $p/n \leq \gamma$. Then for any $\kappa < 1$,
\[
\mathbb{P} \left( \inf_{u \in \Sigma : \|u\| = 1} \|G u\|^2/n < \epsilon \right) \leq \exp(C_2(\kappa') \log(e)(n-p+1)/2-ne/2).
\]

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

Lemma B.1 and its proof are straightforward extensions of [13, Proposition 2.10] which treats the case $\Sigma = I_p, \beta = 0$.

\subsection*{Proof.}

If $V \subset \mathbb{R}^p$ is a subspace of dimension $d = \lfloor \kappa n \rfloor + 1$ and $G = X \Sigma^{-1/2}$ then by (A.2) with $k = 0$, $\epsilon \in (0, (1-d/n)/2)$ and $n$ large enough,
\[
\mathbb{P} \left( \inf_{u \in \Sigma : \|u\| = 1} \|G u\|^2/n < \epsilon \right) \leq \exp(C_2(\kappa') \log(e)(n-p+1)/2-ne/2).
\]

for constant $\kappa' = (\kappa + 1)/2$ thanks to $1 > \kappa' \geq d/n$. Applying this bound to the subspace $V_B = \{u - t\beta, (u, t) \in \mathbb{R}^{p+1} : u_B = 0\}$ for $B \subset [p]$ with $|B| \leq \kappa n$ and using the union bound,
\[
\mathbb{P} \left( \inf_{u \in \Sigma : \|u\| = 1} \|X(u - t\beta)\|^2/n \leq \epsilon \right) \leq \left( \frac{p}{\kappa n} \right) \exp(C_2(\kappa')(\log(e)(n-d+1)/2-ne/2
\]
\[
\leq e^{n \log(e) + C_2(\kappa')(\log(e)(n-d+1)/2-ne/2)}
\]

using $\binom{p}{q} \leq q \log(q)/q \leq n \log(n/p)$ with $q = \lfloor \kappa n \rfloor \leq n$ and $p/n \leq \gamma$. Since $d \leq \kappa n + 1$,

choosing $\epsilon = 1 \land \exp(C_2(\kappa')^{-1}(1-\kappa^{-1}2 \log(e\gamma))$ the right-hand side of the previous display is bounded from above by $e^{-ne/2}$. This value of $\epsilon$ provides $\varphi(\gamma, \kappa)$. \hfill \Box

\subsection*{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 $\Sigma$. Assume that the group Lasso estimator $\widehat{\beta}$ in (3.33) satisfies
\[
\mathbb{P}(\|\widehat{\beta}\|_0 \leq \kappa n/2) \to 1.
\]

If $a_0$ is such that $\|\Sigma^{-1/2} a_0\| = 1$ and $\langle a_0, \widehat{\beta} - \beta \rangle^2 / R_* \to 0$ for the $R_*$ in (3.23) then
\[
\sup_{t \in \mathbb{R}} \mathbb{P}(\|y - X\widehat{\beta}\|^{-1}(n - \delta)\langle a_0, \widehat{\beta} - \beta \rangle + z_0^T (y - X\widehat{\beta})\rangle \leq t) - \Phi(t) \to 0.
\]

Furthermore, for any $a_p$ with $a_p \to \infty$ and $\Sigma$ in (3.25), the relative volume bound given after (3.25) holds, and the asymptotic normality (3.35) holds uniformly over all $a_0 \in \Sigma^{1/2} \Sigma$ and uniformly over at least $(p - \phi_{\mathrm{cond}}(\Sigma) a_p / C^*)$ canonical directions in the sense that $J_p = \{j \in [p] : e_j / \|\Sigma^{-1/2} e_j\| \in \Sigma^{1/2} \Sigma\}$ has cardinality $|J_p| \geq p - \phi_{\mathrm{cond}}(\Sigma) a_p / C^*$.
PROOF OF THEOREM 3.10. As in the rest of the paper, \( f(z_0) = y - X\hat{\beta} \) and we wish to apply Theorem 2.2 to \( z_0 \) conditionally on \((\varepsilon, XQ_0)\). Instead of applying Theorem 2.2 to \( f \), and in order to avoid certain events of small probability where the sparse eigenvalues of \( X \) are not well behaved, we will apply it to a different function. Consider \( F_+ \) in (7.7) and the events

\[
\Omega_L = \{ \|\hat{\beta}\|_0 \leq \kappa n/2 \}, \quad \Omega_\chi = \{ F_+ < 2, \quad (F_+ - 1)_+ < 4\sqrt{\log(n)/n} \}, \\
\Omega_E = \min_{\xi \in \mathbb{R}, u \in \mathbb{R}^p: |u|_0 \leq \kappa} \left\{ \frac{\|X(u - t\hat{\beta})\|}{\sqrt{\|\Sigma^{1/2}(u - t\hat{\beta})\|}} > \varphi\sqrt{n}, \quad \|X\Sigma^{-1/2}\|_{op} < \sqrt{n}(2 + \sqrt{\gamma}) \right\}
\]

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

\[
\Omega \triangleq \Omega_L \cap \Omega_E \cap \Omega_{KKT} \cap \Omega_\chi.
\]

We have \( \mathbb{P}(\Omega_L) \to 1 \) by (3.34) and standard concentration bounds for \( \chi_n^2 \) random variables [30, Lemma 1] give \( \mathbb{P}(\Omega_\chi) \to 1 \). Lemma B.1 and [21, Theorem II.13] provide \( \mathbb{P}(\Omega_E) \to 1 \) and (C.1) gives \( \mathbb{P}(\Omega_{KKT}) = 1 \). These bounds imply \( \mathbb{P}(\Omega) \to 1 \) by the union bound.

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

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

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

For \( F = 2 \max\{1, \|\Sigma^{1/2}h\|^2/(n\|Xh\|^2)\} \) in Lemma 7.2, (7.12) is satisfied so that (7.13)-(7.14) hold. In \( \Omega \), we thus have \( \|\Sigma^{1/2}h\|^2 \leq (\|Xh\|^2)/(2n) \leq 8\varphi^{-2}R_s \) and \( \|y - X\hat{\beta}\|/\sqrt{n} \leq F_{+}^{1/2}\sigma + \sqrt{8}\varphi^{-1}R_s^{1/2} \leq 3\sqrt{2}\varphi^{-1}R_s \). Furthermore \( \|w_0\|^2 \leq \varphi^{-1}/n \) in \( \Omega_E \) thanks to (7.3) \( |\hat{S}| \leq \kappa n/2 \) and the compact expression for \( w_0 \) in Proposition 4.2. In summary we have in \( \Omega \)

(B.2) \( \|\Sigma^{1/2}h\|^2 \leq 8\varphi^{-2}R_s, \|y - X\hat{\beta}\|^2 \leq 18\varphi^{-2}R_s, \|w_0\|^2 \leq \varphi^{-2}/n \)

which replace (7.10)-(7.11) in the present context. By the deterministic inequality (7.29), in \( \Omega \) we have \( \delta \leq |\hat{S}| \leq \kappa n/2 \) since \( \tilde{H} \) is rank at most \( |\hat{S}| \) with operator norm at most one, so that

(B.3) \( I_\Omega(1 - \kappa/2)^2/8 \leq \|y - X\hat{\beta}\|^2/(nR_s) + \Delta^a + \Delta^b_n + \Delta^c_n \).

Let \((\varepsilon, X), (\varepsilon, \tilde{X})\) both in \( \Omega \), let \( \tilde{\varepsilon} = \varepsilon \), and let \( h, \tilde{h}, f, \tilde{f} \) be as in Lemma 7.1. Thanks to event \( \Omega_E \) and the fact that \( |\hat{S}| \leq \kappa n/2 \) and similarly for \( \tilde{\beta} \) we have \( \varphi^2\|\Sigma^{1/2}(h - \tilde{h})\|^2 \leq \|X(h - \tilde{h})\|^2/(2n) + \|X(h - \tilde{h})\|^2/(2n) \). Thus by (7.3),

\[
n\varphi^2\|\Sigma^{1/2}(h - \tilde{h})\|^2 \leq \|h - \tilde{h}\|^2 + \|h - \tilde{h}\|^2 + (h - \tilde{h})^\top(X - \tilde{X})^\top \varepsilon + (h - \tilde{h})^\top(X - \tilde{X})^\top \tilde{X}(h + \tilde{h})/2.
\]
Summing this inequality with the first line in (7.2) we find
\begin{equation}
\begin{split}
n_{\varphi^2} & \mathbb{E}^2 \langle (\mathbf{h} - \mathbf{\bar{h}}), (\mathbf{h} - \mathbf{\bar{h}}) \rangle + \mathbb{E}^2 \| f - \tilde{f} \|^2 \\
\leq (\mathbf{h} - \mathbf{\bar{h}})^\top (\mathbf{X} - \mathbf{\bar{X}}) \mathbf{\varepsilon} + (\mathbf{h} - \mathbf{\bar{h}})^\top (\mathbf{X} - \mathbf{\bar{X}})^\top (\mathbf{h} + \mathbf{\bar{h}})/2 \\
& + (\mathbf{h} - \mathbf{\bar{h}})^\top (\mathbf{X} - \mathbf{\bar{X}})^\top (f - \tilde{f})
\end{split}
\end{equation}
(B.4)

Thanks to the bounds in (B.2), this implies \( \| f - \tilde{f} \| \leq L \| (\mathbf{X} - \mathbf{\bar{X}}) \Sigma^{-1/2} \|_{op} \) if \( \{ (\mathbf{e}, \mathbf{X}), (\mathbf{e}, \mathbf{\bar{X}}) \} \subset \Omega \), where \( L = C_{27}(\gamma, \kappa)R_s^{1/2} \).

For a given \( (\mathbf{e}, \mathbf{X}Q_0) \), we define \( U_0 = \{ z_0 \in \mathbb{R}^n : (\mathbf{e}, \mathbf{X}Q_0 + z_0a_0^\top) \in \Omega \} \). In \( U_0 \), the function \( f(z_0) = \mathbf{X}(\mathbf{\beta} - \mathbf{\bar{\beta}}) - \mathbf{e} \) is \( L \)-Lipschitz. By Kirszenbaum’s theorem, there exists a function \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) that is an extension of \( f \), i.e., \( F(z_0) = f(z_0) \) 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 \( (\mathbf{e}, \mathbf{X}Q_0) \). Since \( \Omega \) is open, \( U_0 \) is also open, and thus conditionally on \( (\mathbf{X}Q_0, \mathbf{e}) \),
\begin{equation}
\nabla f(z_0) = \nabla F(z_0), \quad \text{for all } z_0 \in U_0.
\end{equation}
(B.5)

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

Since \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is such that \( F(z_0) = f(z_0) \) in \( \Omega \), by (B.3)
\begin{equation}
\begin{split}
(1 - \kappa/2)/2 \mathbb{I}_\Omega & \leq \mathbb{E} \| \mathbf{Y} - \mathbf{X} \mathbf{\hat{\beta}} \|/(nR_s) + \mathbb{E} \mathbb{I}_\Omega \| F(z_0) \|/(nR_s) + \mathbb{E} \mathbb{I}_\Omega \| \Delta_n^b + \Delta_n^c \|_n/nL^2
\end{split}
\end{equation}
(B.6)

Taking conditional expectations and multiplying both sides by \( \delta_n^2 \) \( \mathbb{E} = \mathbb{E}_0[\|\nabla F(z_0)\|^2_F]/[\mathbb{E}_0[\|\nabla F(z_0)\|^2_F] + \mathbb{E}_0[\| F(z_0) \|]^2] \) we find
\begin{equation}
\mathbb{E}(\delta_n^2(1 - \kappa/2)/2 \mathbb{I}_\Omega \| \nabla F(z_0) \|^2_F/(nR_s)) \leq \mathbb{E}_0[\| \nabla F(z_0) \|^2_F/(nR_s)] + \mathbb{E}_0[\mathbb{I}_\Omega \| \Delta_n^b + \Delta_n^c \|_n]
\end{equation}
due to \( \delta_n^2 \mathbb{E}_0[\| F(z_0) \|] \leq \mathbb{E}_0[\| \nabla F(z_0) \|^2_F] \) for the first term and \( \delta_n^2 \leq 1 \) for the second. Using \( \delta_n^2 \leq 1 \) and \( 1 = \mathbb{I}_\Omega + I_{\mathbb{I}^c} \),
\begin{equation}
\mathbb{E}(\delta_n^2(1 - \kappa/2)/2 \mathbb{I}_\Omega \| \nabla F(z_0) \|^2_F/(nR_s)) \leq \mathbb{E}_0[\| \nabla F(z_0) \|^2_F/(nR_s)] + \mathbb{E}_0[\mathbb{I}_\Omega \| \Delta_n^b + \Delta_n^c \|_n] + \mathbb{P}(\mathbb{I}^c)(1 + L^2/R_s)
\end{equation}

where we used that \( \| \nabla F(z_0) \|^2_F \leq n \| \nabla F(z_0) \|^2_{op} \leq nL^2 \) in \( \mathbb{I}^c \) since \( F \) is \( L \)-Lipschitz. We now prove that the three terms on the right-hand side of converge to 0. For the third term, \( L^2/R_s \leq C_{28}(\gamma, \kappa) \) and \( \mathbb{P}(\mathbb{I}^c) \rightarrow 0 \) as \( \Omega \) has probability approaching one. For the first term, since \( F \) is \( L \)-Lipschitz, \( \| \nabla F(z_0) \|^2_F \leq nL^2 \) almost surely so that the sequence of random variables \( I_\Omega \| \nabla F(z_0) \|^2_F/(R_s n) \) is uniformly integrable. Thanks to uniform integrability, if we can prove \( \| \nabla F(z_0) \|^2_F/(R_s n) \rightarrow^p 0 \) then \( \mathbb{E}_0[\| \nabla F(z_0) \|^2_F/(R_s n)] \rightarrow 0 \) holds. We use that \( I_\Omega \| \nabla F(z_0) \| = I_\Omega \| \nabla F(z_0) \| (\mathbf{X}Q_0, \mathbf{e}) \) by (B.5), and that in \( \Omega \) the gradients of \( f \) with respect to \( z_0 \) are given in Proposition 4.2 so that by (B.2)
\begin{equation}
\begin{split}
I_\Omega \| \nabla F(z_0) \|/(R_s n) & \leq I_\Omega \| \nabla F(z_0) \|/(R_s n) \leq I_\Omega \| \nabla F(z_0) \|/(R_s n) \leq C_{29}(\gamma, \kappa)(n^{-1} + (a_0, h)/R_s^{1/2})
\end{split}
\end{equation}

which converges to 0 in probability thanks to assumption \( (a_0, \mathbf{\beta} - \mathbf{\bar{\beta}}) \rightarrow^p 0 \). Thanks to uniform integrability, this proves \( \mathbb{E}(I_\Omega \| \nabla F(z_0) \|^2_F/(\sigma^2 n)) \rightarrow 0 \). It remains to show \( \mathbb{E}[I_\Omega \| \Delta_n^b + \Delta_n^c \|_n] \rightarrow 0 \). By definition of \( \Delta_n^b \) in (7.27), thanks to \( \Omega_\chi \) and (B.2) we have
\begin{equation}
I_\Omega \Delta_n^b \leq 18\varphi^{-2}(F_+ - 1) \leq C_{30}(\gamma, \kappa)\sqrt{\log(\gamma)/n}. \quad \text{For } \Delta_n^c \text{ in (7.26), let } \Pi : \mathbb{R}^n \rightarrow \mathbb{R}^n \text{ be the convex projection onto the Euclidean ball of radius } \sqrt{18\varphi^{-2}R_s}, \text{ then } \Pi(y - X\mathbf{\bar{\beta}}) = y - X\mathbf{\bar{\beta}} \text{ in } \Omega \text{ by (B.2) so that}
\end{equation}
\begin{equation}
\mathbb{E}[I_\Omega \mathbb{I}_\Omega] = \mathbb{E}[I_\Omega \mathbb{I}_\Omega \{ (1 - \tilde{d}f)/n - \epsilon^\top \Pi(y - X\mathbf{\bar{\beta}})/(\sigma^2 n)^2 \}] \leq C_{30}(\gamma, \kappa)\sqrt{\log(\gamma)/n}.
\[
\begin{align*}
    \Delta_n & \leq \mathbb{E}[\|\Pi(y - X\hat{\beta})\|^2]/(n^2R_* + \sigma^2/(nR_*) \\
    & \leq 18\varphi^{-2}/n + 1/n
\end{align*}
\] (B.7)

by applying Proposition 2.1 to the function $v \mapsto \Pi(y - X\hat{\beta})$ which is 1-Lipschitz as the composition of two 1-Lipschitz functions (cf. Proposition 7.3(i)). For $\Delta_n$ in (7.28), let $g, a_s$ be as in Lemma 7.6 and set $u_s = \Sigma^{-1}a_s$, $Q_s = I_p - u_s a_s^\top$ and note that $(g, u_s, Q_s) = (x_0, u_0, Q_0)$ for $a_0 = a_s$. Let also $w_s$ be the $w_0$ from Proposition 4.2 for $a_0 = a_s$. As above for $a_0$, for a fixed $(\varepsilon, XQ_s)$ the function $g \mapsto y - X\beta$ is $L$-Lipschitz in $U_s = \{g \in \mathbb{R}^n : (\varepsilon, XQ_s + ga_s^\top) \in \Omega\}$ by (B.4) for the value of $L$ given after (B.4). Furthermore $\|y - X\beta\|^2 \leq 18n\varphi^{-2}R_*$ in $\Omega$. By Kirszbraun’s theorem, there exists an extension $F_s : \mathbb{R}^n \to \mathbb{R}^n$ implicitly depending on $(\varepsilon, XQ_s)$ such that $F_s(g) = y - X\beta$ in $\Omega$ and $\|F_s(g)\|^2 \leq 18n\varphi^{-2}R_*$ 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 $g$ conditionally on $(\varepsilon, XQ_s)$

\[
\mathbb{E}\left[\mathbb{I}_\Omega((n - \bar{d}f)(a_s, h) + (w_s + g, y - X\hat{\beta}))^2\right] = \mathbb{E}\left[\mathbb{I}_\Omega(\text{div} F_s(g) + (g, F_s(g)))^2\right] \\
\leq 18n\varphi^{-2}R_* + nL^2.
\]

For the value of $L$ given after (B.4) and using the bound (B.2) to control $\langle w_s, y - X\hat{\beta} \rangle$ in $\Omega$, this gives $\mathbb{E}[\mathbb{I}_\Omega\Delta_n^2] \leq C_{\mathbb{I}_\Omega}(\gamma, \kappa)n^{-1}$.

This proves $(1 - \kappa/2)/2\mathbb{E}[\hat{S}_n^2] \to 0$. Consequently $\Xi_0 = z_0^\top F(z_0) - \text{div} F(z_0)$ satisfies $\sup_{t, \xi_0} \mathbb{P}((\Xi_0/\|F(z_0)\|) = t - \Phi(t)) \to 0$ by (2.8). Since $\xi_0 = z_0^\top f(z_0) - \text{div} f(z_0)$ is equal to $\Xi_0$ on the event $\Omega$ because $F$ is an extension of $f$, we have $\mathbb{P}((\xi_0/\|F(z_0)\|) \leq t \to 0 \to 0$ so that $\sup_{t, \xi_0} \mathbb{P}((\xi_0/\|F(z_0)\|) \leq t - \Phi(t)) \to 0$ as well. The conclusion (3.35) is obtained by controlling the term $w_0^2(y - X\hat{\beta})^2/\|y - X\hat{\beta}\| \to 0$ which is bounded as in (B.2) in $\Omega$.

It remains to show that (3.35) holds uniformly over all $a_0 \in \Sigma^{1/2}\Sigma$ and to derive the properties of $\Sigma$. The proof of the relative volume bound on $\Sigma$ and the lower bound on the cardinality of $\{j \in [p] : e_j/\|\Sigma^{1/2}e_j\| \in \Sigma^{1/2}\Sigma\}$ is the same as in the proof of Theorem 3.7 given around (7.39), and for $a_0 \in \Sigma^{1/2}\Sigma$ inequality (7.39) holds. For such $a_0$, $\mathbb{E}[\mathbb{I}_\Omega(a_0, h)^2/\|a_0\|^2] \leq 8\varphi^{-2}E[(a_0, h)^2/\|\Sigma^{1/2}h\|^2] \leq 8\varphi^{-2}C^*/a_p \to 0$ 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 $X \in \mathbb{R}^{n \times p}$ and a response vector $y \in \mathbb{R}^n$ for which the joint distribution of $(X, y)$ admits a density with respect to the Lebesgue measure. Consider a partition of $\{1, \ldots, p\}$ into groups $(G_1, \ldots, G_K)$ and any minimizer

$$
\hat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \frac{1}{2n} ||Xb - y||^2 + \|b\|_{GL}, \quad \|b\|_{GL} \overset{df}{=} \sum_{k=1}^K \lambda_k \|b_{G_k}\|_2
$$

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

$$
\left\{(y, X) \in U \right\} \subset \left\{ \forall k = 1, \ldots, K, \quad \hat{\beta}_{G_k} = 0 \quad \Rightarrow \quad \|X_{G_k}^\top (y - X\hat{\beta})\|_2 < n\lambda_k \right\}.
$$

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

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

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

(ii) If $y - X\hat{\beta}(B) \neq 0$, the distribution of $X_{G_k}$ given $(y, (Xe_j)_{j \notin G_k})$ and the distribution of $X_{G_k}^\top (y - X\hat{\beta}(B))$ given $(y, (Xe_j)_{j \notin G_k})$ 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

$$\mathbb{P}\left(\|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2 \neq n\lambda_k \mid y, (Xe_j)_{j \notin G_k}\right) = 1.$$ 

Finally, the unconditional probability $\mathbb{P}(\|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2 \neq n\lambda_k)$ is also one. Let $U = \cap_{B \subseteq \{1, \ldots, K\}} \cap_{k \notin B} \{(y, X) : \|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2 \neq n\lambda_k\}$. Then $\mathbb{P}((y, X) \in U) = 1$ 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 $(y, X) : \|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2 \neq n\lambda_k$ is open by continuity of $(y, X) \mapsto X_{G_k}^\top (y - X\hat{\beta}(B))$ by the claim following (7.2).

Next, to show that $\hat{\beta}$ is constant in a neighborhood of every point in $U$, set $U_{\delta} = \cap_{B \subseteq \{1, \ldots, K\}} \cap_{k \notin B} \{(y, X) : \|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2/(n\lambda_k) - 1 > \delta\}$ for all $\delta > 0$. We have $U = \cup_{\delta > 0} U_{\delta}$ and the set $U_{\delta}$ is open by continuity of $(y, X) \mapsto \|X_{G_k}^\top (y - X\hat{\beta}(B))\|_2/(n\lambda_k) - 1$, which follows from the continuity of $(y, X) \mapsto X_{G_k}^\top (y - X\hat{\beta}(B))$ by the claim following (7.2). For any $(\tilde{y}, \tilde{X}) \in U$, there exists some $\delta > 0$ with $(\tilde{y}, \tilde{X}) \in U_{\delta}$. Let $\tilde{B} = \{k \in [K] : \|\tilde{X}_{G_k}\| > 0\}$.

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

$$[\nabla \hat{\beta}(z_0)]_{\tilde{S}} = 0, \quad [\nabla \hat{\beta}(z_0)]_{\tilde{S}}^\top = (X_{\tilde{S}}^\top X_{\tilde{S}} + M)^{-1}([a_0]_{\tilde{S}} (y - X\hat{\beta})^\top - (a_0, h) X_{\tilde{S}}^\top],$$

$$\tilde{H} = X_{\tilde{S}} (X_{\tilde{S}}^\top X_{\tilde{S}} + M)^{-1}X_{\tilde{S}}^\top$$

and (3.9) holds with $w_0 = X_{\tilde{S}} (X_{\tilde{S}}^\top X_{\tilde{S}} + M)^{-1}(a_0)_{\tilde{S}}$.

Proof of Proposition 4.2. By Lemma C.1, $\hat{B}$ and $\tilde{S}$ are constant in a sufficiently small neighborhood of almost every $(\tilde{y}, \tilde{X})$. The additional assumption that $X_{\tilde{S}}^\top X_{\tilde{S}}$ is invertible provides that $X_{\tilde{S}}^\top X_{\tilde{S}}$ is invertible by continuity of the smallest eigenvalue in a small enough compact neighborhood of $(\tilde{y}, \tilde{X})$, and in this neighborhood $(y, X) \mapsto \hat{\beta}$ is Lipschitz by the sentence following (7.4) and thus almost everywhere differentiable by Rademacher’s theorem. The formulae for $\nabla \hat{\beta}(z_0)$, $\tilde{H}$ and $w_0$ involving the matrix $M$ in (4.5) are then obtained by differentiating the KKT conditions restricted to $\tilde{S}$ in this neighborhood, that is, $X_{G_k}^\top (y - X\hat{\beta}) = n\lambda_k \hat{\beta}_{G_k}/\|\hat{\beta}_{G_k}\|$ for all $k \in \tilde{B}$. □
APPENDIX D: PROOF OF THEOREM 2.3

Proof of Theorem 2.3. With \( \mathbb{E}[\|\mathbf{\overline{p}} + \mathbf{A}^\top z\|^2] = \|\mathbf{\overline{p}}\|^2 + \|\mathbf{A}\|_{\text{F}}^2 \) in mind, consider
\[
\mathbb{V} \mathbb{a} \mathbb{r}[\xi] = \|f(z) - (\mathbf{\overline{p}} + \mathbf{A}^\top z)\|^2 + \text{trace}\{[\nabla f(z) - \mathbf{A}]^2\}
\]
\[
+ 2(f(z) - \mathbf{\overline{p}} - \mathbf{A}^\top z)^\top (\mathbf{\overline{p}} + \mathbf{A}^\top z) + 2 \text{trace}\{[\nabla f(z) - \mathbf{A}]\mathbf{A}\}
\]
\[
+ \left(\|\mathbf{\overline{p}} + \mathbf{A}^\top z\|^2 - \|\mathbf{\overline{p}}\|^2 - \|\mathbf{A}\|_{\text{F}}^2\right) + \|\mathbf{\overline{p}}\|^2 + \|\mathbf{A}\|_{\text{F}}^2 + \text{trace}(\mathbf{A}^2).
\]

By the triangle and Cauchy-Schwarz inequality inequlities,
\[
\mathbb{E}[\|f(z)\|^2 + \text{trace}\{[\nabla f(z)]^2\} - \mathbb{V} \mathbb{a} \mathbb{r}[\xi]]
\]
\[
\leq \mathbb{E}[\|f(z) - (\mathbf{\overline{p}} + \mathbf{A}^\top z)\|^2 + \text{trace}(\nabla f(z) - \mathbf{A})^2]
\]
\[
+ 2\left\{\mathbb{E}[\|f(z) - (\mathbf{\overline{p}} + \mathbf{A}^\top z)\|^2] + \mathbb{E}[\|\nabla f(z) - \mathbf{A}\|_{\text{F}}^2]\right\}^{1/2}\left\{\|\mathbf{\overline{p}}\|^2 + 2\|\mathbf{A}\|_{\text{F}}^2\right\}^{1/2}
\]
\[
+ \mathbb{E}[\|\mathbf{\overline{p}} + \mathbf{A}^\top z\|^2 - \|\mathbf{\overline{p}}\|^2 - \|\mathbf{A}\|_{\text{F}}^2] + \|\mathbf{\overline{p}}\|^2 + \|\mathbf{A}\|_{\text{F}}^2 + \text{trace}(\mathbf{A}^2) - \mathbb{V} \mathbb{a} \mathbb{r}[\xi].
\]

We have \( \mathbb{E}[\|f(z) - (\mathbf{\overline{p}} + \mathbf{A}^\top z)\|^2] \leq \mathbb{E}[\|\nabla f(z) - \mathbf{A}\|_{\text{F}}^2] \leq \epsilon_1^2 \mathbb{V} \mathbb{a} \mathbb{r}[\xi]/2 \) by the Gaussian Poincaré inequality, \( \mathbb{E}[\|\mathbf{\overline{p}} + \mathbf{A}^\top z\|^2 - \|\mathbf{\overline{p}}\|^2 - \|\mathbf{A}\|_{\text{F}}^2] = \|\mathbf{A}\|_{\text{op}}^2 \mathbb{V} \mathbb{a} \mathbb{r}[\xi] \), and \( 0 \leq 1 - \left\{\|\mathbf{\overline{p}}\|^2 + \|\mathbf{A}\|_{\text{F}}^2 + \text{trace}(\mathbf{A}^2)\right\}/\mathbb{V} \mathbb{a} \mathbb{r}[\xi] \leq \epsilon_2^2 \) as in (2.11). Thus (2.14) holds and the conclusions follow. \( \square \)

REFERENCES

[1] Mohsen Bayati and Andrea Montanari.
The lasso risk for gaussian matrices.
IEEE Transactions on Information Theory, 58(4):1997–2017, 2012.
[2] Pierre C Bellec.
The noise barrier and the large signal bias of the lasso and other convex estimators.
\textit{arXiv}:1804.01230, 2018.
URL https://arxiv.org/pdf/1804.01230.pdf.
[3] Pierre C Bellec.
The noise barrier and the large signal bias of the lasso and other convex estimators.
\textit{arXiv preprint arXiv}:1804.01230, 2018.
[4] Pierre C Bellec.
Out-of-sample error estimate for robust m-estimators with convex penalty.
\textit{arXiv preprint arXiv}:2008.11840, 2020.
[5] Pierre C Bellec and Yiwei Shen.
Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning.
In Conference on Learning Theory, pages 1912–1947. PMLR, 2022.
[6] Pierre C Bellec and Alexandre B Tsybakov.
Bounds on the prediction error of penalized least squares estimators with convex penalty.
In Vladimir Panov, editor, Modern Problems of Stochastic Analysis and Statistics, Selected Contributions In Honor of Valentin Konakov, Springer, 2017.
URL https://arxiv.org/pdf/1609.06675.pdf.
[7] Pierre C Bellec and Cun-Hui Zhang.
Second order stein: Sure for sure and other applications in high-dimensional inference.
\textit{arXiv preprint arXiv}:1811.04121, 2018.
[8] Pierre C Bellec and Cun-Hui Zhang.
De-biasing the lasso with degrees-of-freedom adjustment.
\textit{preprint}, 2019.
[9] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov.
Slope meets lasso: Improved oracle bounds and optimality.
\textit{Ann. Statist.}, 46(6B):3603–3642, 2018.
ISSN 0090-5364.
URL https://arxiv.org/pdf/1605.08651.pdf.
[10] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. 
High-dimensional methods and inference on structural and treatment effects. 
*Journal of Economic Perspectives*, 28(2):29–50, 2014.

[11] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and YA’Acov Ritov. 
Efficient and adaptive estimation for semiparametric models. 
Johns Hopkins University Press Baltimore, 1993.

[12] Peter J. Bickel, Yaacov Ritov, and Alexandre B. Tsybakov. 
Simultaneous analysis of lasso and dantzig selector. 
*Ann. Statist.*, 37(4):1705–1732, 08 2009.

URL: http://dx.doi.org/10.1214/08-AOS620.

[13] Jeffrey D Blanchard, Coralia Cartis, and Jared Tanner. 
Compressed sensing: How sharp is the restricted isometry property? 
*SIAM review*, 53(1):105–125, 2011.

[14] Thomas Blumensath and Mike E Davies. 
Iterative hard thresholding for compressed sensing. 
*Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[15] Vladimir Igorevich Bogachev. 
*Gaussian measures*. 
Number 62. American Mathematical Soc., 1998.

[16] Zhiqi Bu, Jason Klusowski, Cynthia Rush, and Weijie Su. 
Algorithmic analysis and statistical estimation of slope via approximate message passing. 
In *Advances in Neural Information Processing Systems*, pages 9361–9371, 2019.

[17] T Tony Cai and Zijian Guo. 
Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. 
*The Annals of statistics*, 45(2):615–646, 2017.

[18] Michael Celentano and Andrea Montanari. 
Fundamental barriers to high-dimensional regression with convex penalties. 
*arXiv preprint arXiv:1903.10603*, 2019.

[19] Michael Celentano, Andrea Montanari, and Yuting Wei. 
The lasso with general gaussian designs with applications to hypothesis testing. 
*arXiv preprint arXiv:2007.13716*, 2020.

[20] Sourav Chatterjee. 
Fluctuations of eigenvalues and second order poincaré inequalities. 
*Probability Theory and Related Fields*, 143(1-2):1–40, 2009.

[21] Kenneth R Davidson and Stanislaw J Szarek. 
Local operator theory, random matrices and banach spaces. 
*Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[22] David Donoho and Andrea Montanari. 
High dimensional robust m-estimation: Asymptotic variance via approximate message passing. 
*Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[23] Alan Edelman. 
Eigenvalues and condition numbers of random matrices. 
*SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[24] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghay Lim, and Bin Yu. 
On robust regression with high-dimensional predictors. 
*Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[25] Jianqing Fan and Runze Li. 
Variable selection via nonconcave penalized likelihood and its oracle properties. 
*Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[26] Long Feng and Cun-Hui Zhang. 
Sorted concave penalized regression. 
*Annals of Statistics*, 47(6):3069–3098, 2019.

[27] Adel Javanmard and Andrea Montanari. 
Confidence intervals and hypothesis testing for high-dimensional regression. 
*The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[28] Adel Javanmard and Andrea Montanari. 
Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. 
*IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.
ASYMPTOTIC NORMALITY WHEN $P/N \to \gamma$

[29] Adel Javanmard and Andrea Montanari.
Debiasing the lasso: Optimal sample size for gaussian designs.
The Annals of Statistics, 46(6A):2593–2622, 2018.

[30] B. Laurent and P. Massart.
Adaptive estimation of a quadratic functional by model selection.
Ann. Statist., 28(5):1302–1338, 10 2000.

URL http://dx.doi.org/10.1214/aos/1015957395.

[31] Guillaume Lecuè and Shahar Mendelson.
Regularization and the small-ball method i: Sparse recovery.
Ann. Statist., 46(2):611–641, 04 2018.

URL https://doi.org/10.1214/17-AOS1562.

[32] Lihua Lei, Peter J Bickel, and Noureddine El Karoui.
Asymptotics for high dimensional regression m-estimates: fixed design results.
Probability Theory and Related Fields, 172(3-4):983–1079, 2018.

[33] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová.
Learning curves of generic features maps for realistic datasets with a teacher-student model.
Advances in Neural Information Processing Systems, 34:18137–18151, 2021.

[34] Léo Mikolaj and Andrea Montanari.
The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning.
arXiv preprint arXiv:1811.01212, 2018.

[35] Mohamed Ndaoud.
Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach.
arXiv preprint arXiv:2008.12236, 2020.

[36] Constantin Niculescu and Lars-Erik Persson.
Convex functions and their applications.
Springer, 2006.

[37] Jack W Silverstein.
The smallest eigenvalue of a large dimensional wishart matrix.
The Annals of Probability, 13(4):1364–1368, 1985.

[38] Charles M Stein.
estimation of the mean of a multivariate normal distribution.
The annals of Statistics, pages 1135–1151, 1981.

[39] Tingni Sun and Cun-Hui Zhang.
Scaled sparse linear regression.
Biometrika, 99(4):879–898, 2012.

[40] Tingni Sun and Cun-Hui Zhang.
Sparse matrix inversion with scaled lasso.
The Journal of Machine Learning Research, 14(1):3385–3418, 2013.

[41] Pragya Sur and Emmanuel J Candès.
A modern maximum-likelihood theory for high-dimensional logistic regression.
arXiv preprint arXiv:1803.06964, 2018.

[42] Michel Talagrand.
Mean field models for spin glasses: Volume I: Basic examples, volume 54.
Springer, 2010.

[43] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi.
Lasso with non-linear measurements is equivalent to one with linear measurements.
In Advances in Neural Information Processing Systems, pages 3420–3428, 2015.

[44] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi.
Precise error analysis of regularized m-estimators in high dimensions.
IEEE Transactions on Information Theory, 64(8):5592–5628, 2018.

[45] Ryan J Tibshirani.
The lasso problem and uniqueness.
Electronic Journal of Statistics, 7:1456–1490, 2013.

[46] Samuel Vaiter, Charles Deledalle, Gabriel Peyré, Jalal Fadili, and Charles Dossal.
The degrees of freedom of the group lasso.
arXiv preprint arXiv:1205.1481, 2012.
[47] Sara Van de Geer, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure.
On asymptotically optimal confidence regions and tests for high-dimensional models.
The Annals of Statistics, 42(3):1166–1202, 2014.

[48] Roman Vershynin.
High-dimensional probability: An introduction with applications in data science, volume 47.
Cambridge University Press, 2018.

[49] Cun-Hui Zhang.
Nearly unbiased variable selection under minimax concave penalty.
The Annals of statistics, 38(2):894–942, 2010.

[50] Cun-Hui Zhang.
Statistical inference for high-dimensional data.
Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report, (48):28–31, 2011.

[51] Cun-Hui Zhang and Jian Huang.
The sparsity and bias of the lasso selection in high-dimensional linear regression.
Ann. Statist., 36(4):1567–1594, 08 2008.
URL https://doi.org/10.1214/07-AOS520.

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