The Annals of Statistics
2021, Vol. 49, No. 4, 1864–1903
https://doi.org/10.1214/20-AOS2005
© Institute of Mathematical Statistics, 2021

SECOND-ORDER STEIN: SURE FOR SURE AND OTHER APPLICATIONS IN HIGH-DIMENSIONAL INFERENCE

BY PIERRE C. BELLEC* AND CUN-HUI ZHANG†

Department of Statistics, Rutgers University, *pierre.bellec@rutgers.edu; †czhang@stat.rutgers.edu

Stein's formula states that a random variable of the form $z^\top f(z) - \operatorname{div} f(z)$ is mean-zero for all functions f with integrable gradient. Here, $\operatorname{div} f$ is the divergence of the function f and z is a standard normal vector. This paper aims to propose a second-order Stein formula to characterize the variance of such random variables for all functions $f(z)$ with square integrable gradient, and to demonstrate the usefulness of this second-order Stein formula in various applications.

In the Gaussian sequence model, a remarkable consequence of Stein's formula is Stein's Unbiased Risk Estimate (SURE), an unbiased estimate of the mean squared risk for almost any given estimator $\widehat{\boldsymbol{\mu}}$ of the unknown mean vector. A first application of the second-order Stein formula is an Unbiased Risk Estimate for SURE itself (SURE for SURE): an unbiased estimate providing information about the squared distance between SURE and the squared estimation error of $\widehat{\boldsymbol{\mu}}$. SURE for SURE has a simple form as a function of the data and is applicable to all $\widehat{\boldsymbol{\mu}}$ with square integrable gradient, for example, the Lasso and the Elastic Net.

In addition to SURE for SURE, the following statistical applications are developed: (1) upper bounds on the risk of SURE when the estimation target is the mean squared error; (2) confidence regions based on SURE and using the second-order Stein formula; (3) oracle inequalities satisfied by SURE-tuned estimates under a mild Lipschtiz assumption; (4) an upper bound on the variance of the size of the model selected by the Lasso, and more generally an upper bound on the variance of the empirical degrees-of-freedom of convex penalized estimators; (5) explicit expressions of SURE for SURE for the Lasso and the Elastic Net; (6) in the linear model, a general semiparametric scheme to de-bias a differentiable initial estimator for the statistical inference of a low-dimensional projection of the unknown regression coefficient vector, with a characterization of the variance after debiasing; and (7) an accuracy analysis of a Gaussian Monte Carlo scheme to approximate the divergence of functions $f : \mathbb{R}^n \to \mathbb{R}^n$.

1. Introduction.

1.1. *Stein's formula.* The univariate version of Stein's formula reads $\mathbb{E}[Zg(Z)] = \mathbb{E}[g'(Z)]$ where $Z \sim N(0, 1)$ and g is absolutely continuous with finite $\mathbb{E}[|g'(Z)|]$.[1] In [45], Stein observed that the set of equations $\mathbb{E}[Zg(Z) - g'(Z)] = 0$ for all absolutely continuous g with integrable gradient characterize the standard normal distribution. Stein went on to show that if $\mathbb{E}[Wg(W) - g'(W)]$ is small for a large class of functions g, then W is close in distribution to $N(0, 1)$; see, for instance, [18], Proposition 1.1, for a precise statement. Since then, this powerful technique has been used to obtain central limit theorems under dependence; we

Received July 2019; revised July 2020.
MSC2020 subject classifications. Primary 62H10, 62H12, 62J07; secondary 62G15, 62F35.
Key words and phrases. Stein's formula, variance estimate, risk estimate, SURE, SURE for SURE, Lasso, Elastic Net, model selection, variance of model size, debiased estimation, regression.

[1]Existence of $\mathbb{E}|Zg(Z)| < +\infty$ follows from $\mathbb{E}|g'(Z)| < +\infty$. Indeed by Fubini's theorem $\mathbb{E}|\max(Z,0)\{g(Z) - g(0)\}| \le \int_0^\infty |g'(u)|\mathbb{E}[ZI_{\{Z\ge u\}}]\,du = \mathbb{E}[I_{Z>0}|g'(Z)|]$ and similarly for $Z < 0$.

refer the reader to the book [19] for a recent survey on Stein's formula and its application to normal approximation.

The multivariate version of Stein's formula [46] can be described as follows. Let $\boldsymbol{z} = (z_1, \dots, z_n)^\top \sim N(\mathbf{0}, \boldsymbol{I}_n)$ be a standard normal random vector. Let $f_1, \dots, f_n$ be functions $\mathbb{R}^n \to \mathbb{R}$ and denote by f the column vector in $\mathbb{R}^n$ with ith component equal to f_i. If each f_i is weakly differentiable with respect to z_i and $\mathbb{E}[|f_i(\boldsymbol{z})| + |(\partial/\partial z_i) f_i(\boldsymbol{z})|] < +\infty$ for each $i \in [n]$, then

$$\mathbb{E}[\boldsymbol{z}^\top f(\boldsymbol{z})] = \mathbb{E}[\operatorname{div} f(\boldsymbol{z})] \tag{1.1}$$

holds, where $\operatorname{div} f = \sum_{i=1}^n (\partial/\partial x_i) f_i$ is the divergence of f. The central object of the paper is the random variable $\boldsymbol{z}^\top f(\boldsymbol{z}) - \operatorname{div} f(\boldsymbol{z})$, which is mean-zero in virtue of (1.1). We provide in the next section an exact identity for the second moment of $\boldsymbol{z}^\top f(\boldsymbol{z}) - \operatorname{div} f(\boldsymbol{z})$ that we termed *second-order Stein formula*.

1.2. *Related work.* Iterated Stein formulae have appeared in several works. If $g : \mathbb{R}^p \to \mathbb{R}$ is smooth with compact support, iterating the univariate Stein's formula directly yields $\mathbb{E}[z_i z_j g(\boldsymbol{z})] = \mathbb{E}[(\partial/\partial x_j)(\partial/\partial x_i) g(\boldsymbol{z})]$ for $i \neq j$ as well as $\mathbb{E}[(z_i^2 - 1) g(\boldsymbol{z})] = \mathbb{E}[(\partial^2/\partial x_i^2) g(\boldsymbol{z})]$. This may be succinctly written as

$$\mathbb{E}[(\boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{I}_n) g(\boldsymbol{z})] = \mathbb{E}[\nabla^2 g(\boldsymbol{z})], \tag{1.2}$$

where $\nabla^2 g$ is the Hessian of g. Formulae that relate $\mathbb{E}[\nabla^2 g(\boldsymbol{z})]$ to expectations involving $g(\boldsymbol{z})$ and the density of $\boldsymbol{z}$ have been obtained for non-Gaussian $\boldsymbol{z}$; see [29], Theorem 1, [25], [57], Theorem 2.2, among others. In these works, (1.2) is used to estimate the Hessian in optimization algorithms [25], to estimate the index in some single-index models [57] or to extract discriminative features for tensor-valued data [29]. If $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_n)$, the second- order iterated Stein formula from these works reduce to (1.2). The fact that iterating Stein's formula introduces higher order derivatives should be expected, because Stein's formula is intrinsically an integration by parts—Stein's formula is sometimes referred to as *Gaussian integration by parts*, for instance in [49], Appendix A.4, where statistical physics results are rigorously derived using Stein's formula and the so-called Smart path method as a central tool. As we will see throughout the paper, the identity for the variance of $\boldsymbol{z}^\top f(\boldsymbol{z}) - \operatorname{div} f(\boldsymbol{z})$, given in (2.3) below, does not involve the second derivatives of f or require their existence, in striking contrast to the iterated formula (1.2). Although both identities can be seen as second-order Stein formulae, they are of a different nature and aim different applications.

1.3. *Motivations.* In the Gaussian sequence model where $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ is observed for an unknown mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and noise $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n)$, Stein's Unbiased Risk Estimate (SURE), proposed in [46], provides an unbiased estimator for the risk $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ of any estimator $\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ with integrable gradient as a consequence of (1.1). Namely, the decomposition

$$\underbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + 2\operatorname{div}\widehat{\boldsymbol{\mu}} - n}_{\widehat{\mathsf{SURE}}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \boldsymbol{\varepsilon}^\top h(\boldsymbol{\varepsilon}) - \operatorname{div} h(\boldsymbol{\varepsilon}),$$

where $h(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon} - 2(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ shows that $\widehat{\mathsf{SURE}}$ defined above is an unbiased estimate of $\mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$ in virtue of Stein's formula (1.1) applied to h. Although the identity $\mathbb{E}[\widehat{\mathsf{SURE}}] = \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$ is well understood and widely used, little is known about the quality of the approximation $\widehat{\mathsf{SURE}} \approx \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ in probability, or equivalently about the variance and concentration properties of the random variable $\boldsymbol{\varepsilon}^\top h(\boldsymbol{\varepsilon}) - \operatorname{div} h(\boldsymbol{\varepsilon})$ above. A first motivation of the present paper is to understand the stochastic behavior of $\boldsymbol{\varepsilon}^\top h(\boldsymbol{\varepsilon}) - \operatorname{div} h(\boldsymbol{\varepsilon})$ (beyond unbiasedness) in order to provide uncertainty quantification for SURE. If multiple estimators $\{\widehat{\boldsymbol{\mu}}^{(1)}, \dots, \widehat{\boldsymbol{\mu}}^{(m)}\}$ are given for estimation of $\boldsymbol{\mu}$, it is natural to use the estimator among those with the smallest SURE. However, statistical guarantees for such SURE-tuned estimate are lacking, except in very special cases, for instance when $\{\widehat{\boldsymbol{\mu}}^{(1)}, \dots, \widehat{\boldsymbol{\mu}}^{(m)}\}$ are linear functions of $\boldsymbol{y}$. When the estimators are linear, $\widehat{\mathsf{SURE}}$ is the same as Mallows' C_p [38] and the SURE-tuned estimate enjoys oracle inequalities with remainder term growing with $\log m$; cf. [5] or Section 3.3 and the references therein. We are not aware of oracle inequalities satisfied by the SURE-tuned estimate for general nonlinear estimators $\{\widehat{\boldsymbol{\mu}}^{(1)}, \dots, \widehat{\boldsymbol{\mu}}^{(m)}\}$; as explained in Section 3.3 below, our bounds on the variance of $\boldsymbol{\varepsilon}^\top h(\boldsymbol{\varepsilon}) - \operatorname{div} h(\boldsymbol{\varepsilon})$ are helpful to fill this gap.

A second motivation of the paper regards the Lasso $\hat{\boldsymbol{\beta}}$ in sparse linear regression with Gaussian noise $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n)$. Despite an extensive body of literature on the Lasso in the last two decades, little is known about the stochastic behavior of the size of the model selected by Lasso; for instance, no previous results are available on the variance of this integer valued random variable. It is well known that the model size of the lasso can be written as $\operatorname{div} f(\boldsymbol{\varepsilon})$ for certain function $f : \mathbb{R}^n \to \mathbb{R}^n$; see Section 3.4 below and the references therein. This integer valued random variable—the size of the model selected by the Lasso—is not differentiable with respect to $\boldsymbol{\varepsilon}$ and the Stein formulae (1.1)–(1.2) cannot be iterated further to understand the variance of this discrete random variable. Another avenue to bound the variance of a random variable of the form $g(z)$, $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$ is the Gaussian Poincaré inequality, which states that $\operatorname{Var}[g(z)] \le \mathbb{E}[\|\nabla g(z)\|^2]$ for differentiable $g : \mathbb{R}^n \to \mathbb{R}$ [14], Theorem 3.20. Again, this is unhelpful if $g(\cdot)$ is discrete-valued and nondifferentiable.

1.4. *Contributions*. The central result of the paper states that the variance of the random variable $z^\top f(z) - \operatorname{div} f(z)$ is exactly given by

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

when both $f : \mathbb{R}^n \to \mathbb{R}^n$ and its derivative are square integrable. Surprisingly, although this is obtained by iterated integration by parts, the second derivatives do not appear. As explained in the next section, the second derivatives of f need not exist for the previous display to hold—in that sense, the previous display is more broadly applicable than (1.2). The above formula appears especially useful for estimators commonly called for in high-dimensional statistics such as the Lasso, the Elastic-Net or the Group-Lasso that are differentiable almost everywhere with respect to the noise, but not twice differentiable. The contributions of the paper are summarized below:

- We derive in Section 2.1 the above second-order Stein formula, an identity for the variance of the random variable $z^\top f(z) - \operatorname{div} f(z)$.
- As a consequence, Section 2.3 provides an identity and an upper bound for $\operatorname{Var}[z^\top f(z) - \operatorname{div} f(z) - g(z)]$ in general, including $\operatorname{Var}[\operatorname{div} f(z)]$.

Statistical applications are then provided in Section 3:

- Section 3.1 leverages the second-order Stein formula to construct an unbiased risk estimate for Stein's Unbiased Risk Estimate (SURE) in the Gaussian sequence model. We shall call this general method SURE for SURE.
- Section 3.2 describes confidence regions based on SURE and the second-order Stein formula.
- Section 3.3 provides oracle inequalities for SURE-tuned estimates.
- Section 3.4 provides new bounds on the variance of the size of the model selected by the Lasso in sparse linear regression.
- Section 3.5 provides SURE for SURE formulas for the Lasso and E-net as well as the consistency of SURE for SURE in the Lasso case.

- Section 3.6 provides a scheme to de-bias a general class of estimators in linear regression where one wishes to estimate a low-dimensional projection of the unknown regression coefficient vector.
- Section 3.7 develops a Monte Carlo scheme to approximate the divergence of a general differentiable estimator when the analytic form of the divergence is unavailable.

The above statistical applications are obtained as consequences of the second-order Stein formula of Section 2.1. The point of the second-order Stein formula is that, in many statistical applications where the random variable $z^\top f(z) - \operatorname{div} f(z)$ appears, the standard deviation $\operatorname{Var}[z^\top f(z) - \operatorname{div} f(z)]^{1/2}$ is of smaller order than either $\operatorname{div} f(z)$ or $z^\top f(z)$ and the approximation $z^\top f(z) \approx \operatorname{div} f(z)$ holds not only in expectation as in (1.2), but also in probability.

As the second-order Stein formulas in Section 2 are explicit identities and sharp inequalities, the results in Section 3 are all nonasymptotic in nature except Corollary 3.5, often stated with explicit constants. In applications to linear regression, the second-order Stein formulas are used in conjunction with results which may require sparsity conditions, but the relatively mild sparsity $\|\boldsymbol{\beta}\|_0 \ll n/\log p$ are sufficient in such cases where $\boldsymbol{\beta}$ is the coefficient vector living in $\mathbb{R}^p$ and n is the sample size.

1.5. *Notation.* Throughout the paper, $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_F$ the Frobenius norm, and $f : \mathbb{R}^n \to \mathbb{R}^m$ is L-Lipschitz if $\|f(\boldsymbol{u}) - f(\boldsymbol{v})\| \le L\|\boldsymbol{u} - \boldsymbol{v}\|$, $\forall \boldsymbol{u}, \boldsymbol{v}$. For $f : \mathbb{R}^n \to \mathbb{R}^n$ with components f_i, denote by $\nabla f_i \in \mathbb{R}^n$ the gradient of each f_i, and by ∇f the matrix in $\mathbb{R}^{n\times n}$ with columns $\nabla f_1, \dots, \nabla f_n$. The random variable X is said to be in L_1 (resp., L_2) if $\mathbb{E}[|X|] < +\infty$ (resp., $\mathbb{E}[X^2] < +\infty$). For two symmetric matrices $\boldsymbol{A}$, $\boldsymbol{B}$ of the same size, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ if and only if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

2. Mathematical results.

2.1. *A second-order Stein formula.* Stein's formula (1.1) states that the random variable

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

is mean-zero. The main result of the current paper is the following *second-order Stein* formula, which provides an identity for the variance of the random variable (2.1) with f in the Sobolev space $W^{1,2}(\gamma_n)$ with respect to the standard Gaussian measure γ_n in $\mathbb{R}^n$. Let $C_0^\infty(\mathbb{R}^n)$ be the space of infinitely differentiable functions $\mathbb{R}^n \to \mathbb{R}$ with compact support. The Sobolev space $W^{1,2}(\gamma_n)$ is defined as the completion of the space $C_0^\infty(\mathbb{R}^n)$ with respect to the norm

$$\|g\|_{1,2} = \mathbb{E}[g(z)^2]^{1/2} + \mathbb{E}[\|\nabla g(z)\|^2]^{1/2}, \quad g \in C_0^\infty(\mathbb{R}^n), \tag{2.2}$$

where $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$. By [13], Proposition 1.5.2, the space $W^{1,2}(\gamma_n)$ corresponds to locally integrable functions $g : \mathbb{R}^n \to \mathbb{R}$ that are weakly differentiable and $\mathbb{E}[g(z)^2] + \mathbb{E}[\|\nabla g(z)\|^2] < +\infty$ for $z \sim N(\mathbf{0}, \boldsymbol{I}_n)$. This definition of the Sobolev space emphasizes the feasibility of approximating the functions in the space by infinitely smooth ones. Alternatively, $W^{1,2}(\gamma_n)$ can be defined as the completion of the space of Lipschitz functions under the same norm. We refer to Section 1.5 in [13] for a complete description of the space $W^{1,2}(\gamma_n)$.

THEOREM 2.1. *Let $z = (z_1, \dots, z_n)$ be a standard normal $N(\mathbf{0}, \boldsymbol{I}_n)$ random vector. Let $f_1, \dots, f_n$ be functions $\mathbb{R}^n \to \mathbb{R}$ and f be the column vector in $\mathbb{R}^n$ with ith component equal to f_i.*

(i) *Assume that each f_i is twice continuously differentiable and that its first- and second-order derivatives have subexponential growth. Then*

$$\mathbb{E}\big[\big(z^\top f(z) - \operatorname{div} f(z)\big)^2\big] = \mathbb{E}\sum_{i=1}^n f_i^2(z) + \mathbb{E}\sum_{i=1}^n\sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(z)\frac{\partial f_j}{\partial x_i}(z). \tag{2.3}$$

An equivalent version of (2.3) *for functions of $\mathbf{y} = N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, in vector notation, is given in* (2.7) *below.*

(ii) *If $f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz for $L < +\infty$, then* (2.3) *holds.*

(iii) *If each component f_i of f belongs to $W^{1,2}(\gamma_n)$, then* (2.3) *holds.*

We note that while (i) and (ii) may have broader appeal due to their simplicity, the assumption in (iii) above is the most general: If f satisfies either the assumption in (i) or (ii) then the components of f are in $W^{1,2}(\gamma_n)$ and the assumption in (iii) holds. The presentation above also highlights the proof strategy: We first derive (2.3) under the smoothness assumption in (i) and (ii)–(iii) then follow by an approximation argument.

Here is a proof of Theorem 2.1(i). Extensions (ii) and (iii) are proved in Appendix A.

PROOF OF THEOREM 2.1(i). When the functions f_i and $(\partial/\partial x_j) f_i$ are treated as random variables, their argument is always z through the proof, so we simply write f_i for $f_i(z)$ and similarly for the partial derivatives. A sum $\sum_i$ or $\sum_k$ always sums over $\{1, \dots, n\}$.

Write the left-hand side in (2.3) as

$$\mathbb{E}\sum_i \left(z_i f_i - \frac{\partial f_i}{\partial x_i}\right)\left(\sum_j z_j f_j - \sum_l \frac{\partial f_l}{\partial x_l}\right).$$

By a first application of Stein's formula for each term z_i below, the identity

$$\mathbb{E}\left[\left(z_i f_i(z) - \frac{\partial f_i}{\partial x_i}(z)\right) g(z)\right] = \mathbb{E}\left[f_i(z)\frac{\partial g}{\partial x_i}(z)\right] \tag{2.4}$$

holds for any continuously differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ such that the partial derivatives $(\partial/\partial x_i) g$ have subexponential growth. Hence the left-hand side in (2.3) equals

$$\mathbb{E}\sum_i f_i^2 + \mathbb{E}\sum_i f_i \sum_j z_j \frac{\partial f_j}{\partial x_i} - \mathbb{E}\sum_i f_i \sum_l \frac{\partial^2 f_l}{\partial x_i \partial x_l}.$$

We again apply Stein's formula to each z_j in the second term above to obtain

$$\mathbb{E}\sum_i f_i^2 + \mathbb{E}\sum_i\sum_j \frac{\partial f_i}{\partial x_j}\frac{\partial f_j}{\partial x_i} + \mathbb{E}\sum_i f_i \sum_j \frac{\partial^2 f_j}{\partial x_j \partial x_i} - \mathbb{E}\sum_i f_i \sum_l \frac{\partial^2 f_l}{\partial x_i \partial x_l}.$$

Since f is twice continuously differentiable, by the Schwarz theorem on the symmetry of the second derivatives, we have $\sum_j (\partial/\partial x_j)(\partial/\partial x_i) f_j = \sum_\ell (\partial/\partial x_i)(\partial/\partial x_\ell) f_\ell$ and the proof of (2.3) is complete. □

The second-order Stein formula (2.3) can then be rewritten as

$$\mathbb{E}\big[\big(z^\top f(z) - \operatorname{div} f(z)\big)^2\big] = \mathbb{E}\big[\|f(z)\|^2 + \operatorname{trace}\big((\nabla f(z))^2\big)\big]. \tag{2.5}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} \mathbb{E}\big[\big(z^\top f(z) - \operatorname{div} f(z)\big)^2\big] &\le \mathbb{E}\sum_{i=1}^n f_i^2(z) + \mathbb{E}\sum_{i=1}^n\sum_{j=1}^n \left(\frac{\partial f_i}{\partial x_j}(z)\right)^2 \\ &= \mathbb{E}\big[\|f(z)\|^2 + \|\nabla f(z)\|_F^2\big]. \end{aligned} \tag{2.6}$$

If $\nabla f(z)$ is almost surely symmetric, then $\mathrm{trace}((\nabla f(z))^2) = \|\nabla f(z)\|_F^2$ and the above inequality is actually an equality. However, the inequality in (2.6) is strict otherwise.

If $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ and $f : \mathbb{R}^n \to \mathbb{R}^n$, then

$$\mathbb{E}\big[\big(\boldsymbol{\varepsilon}^\top f(\boldsymbol{y}) - \sigma^2 \operatorname{div} f(\boldsymbol{y})\big)^2\big] = \sigma^2 \mathbb{E}\big[\|f(\boldsymbol{y})\|^2\big] + \sigma^4 \mathbb{E}\big[\mathrm{trace}\big((\nabla f(\boldsymbol{y}))^2\big)\big] \tag{2.7}$$

is obtained by setting $z = \boldsymbol{\varepsilon}/\sigma$ and applying Theorem 2.1 to $\tilde{f}(\boldsymbol{x}) = \sigma f(\boldsymbol{\mu} + \sigma \boldsymbol{x})$, provided that $\tilde{f}$ satisfies the assumption of Theorem 2.1(iii). Similarly, if $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ then

$$\mathbb{E}\big[\big\{\boldsymbol{\varepsilon}^\top f(\boldsymbol{y}) - \mathrm{trace}\big(\boldsymbol{\Sigma} \nabla f(\boldsymbol{y})\big)\big\}^2\big] = \mathbb{E}\big[\|\boldsymbol{\Sigma}^{1/2} f(\boldsymbol{y})\|^2\big] + \mathbb{E}\,\mathrm{trace}\big[\big\{\boldsymbol{\Sigma} \nabla f(\boldsymbol{y})\big\}^2\big]$$

follows from Theorem 2.1 with $z = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\varepsilon}$ and $\tilde{f}(\boldsymbol{x}) = \boldsymbol{\Sigma}^{1/2} f(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{x})$.

Theorem 2.1 is also applicable under the central limit theorem. Let $\boldsymbol{\varepsilon}_m = m^{-1/2} \sum_{i=1}^m \boldsymbol{x}_i$ with i.i.d. $\boldsymbol{x}_i \in \mathbb{R}^n$, $\mathbb{E}[\boldsymbol{x}_i] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] = \boldsymbol{\Sigma}$. If $\{\|f(\boldsymbol{\varepsilon}_m)\|^2, \|\nabla f(\boldsymbol{\varepsilon}_m)\|_F^2, (\boldsymbol{\varepsilon}_m^\top f(\boldsymbol{\varepsilon}_m))^2, m \geq 1\}$ is uniformly integrable and ∇f is almost everywhere continuous, then $\mathbb{E}[\boldsymbol{\varepsilon}_m^\top f(\boldsymbol{\varepsilon}_m)] = \mathbb{E}[\mathrm{trace}(\boldsymbol{\Sigma} \nabla f(\boldsymbol{\varepsilon}_m))] + o(1)$ and

$$\begin{aligned} &\mathbb{E}\big[\big(\boldsymbol{\varepsilon}_m^\top f(\boldsymbol{\varepsilon}_m) - \mathrm{trace}\big(\boldsymbol{\Sigma} \nabla f(\boldsymbol{\varepsilon}_m)\big)\big)^2\big] \\ &\quad = \mathbb{E}\big[\|\boldsymbol{\Sigma}^{1/2} f(\boldsymbol{\varepsilon}_m)\|^2\big] + \mathbb{E}\big[\mathrm{trace}\big((\boldsymbol{\Sigma} \nabla f(\boldsymbol{\varepsilon}_m))^2\big)\big] + o(1) \end{aligned} \tag{2.8}$$

as $m \to \infty$ for fixed n. If, for instance, $f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz with almost everywhere continuous gradient and $\mathbb{E}[\|\boldsymbol{x}_i\|^4] \leq C$ for some constant $C > 0$ possibly depending on the dimension, then $\{\|\boldsymbol{\varepsilon}_m\|^4, m \geq 1\}$ is uniformly integrable, which implies that $\{\|f(\boldsymbol{\varepsilon}_m)\|^2, \|\nabla f(\boldsymbol{\varepsilon}_m)\|_F^2, (\boldsymbol{\varepsilon}_m^\top f(\boldsymbol{\varepsilon}_m))^2, m \geq 1\}$ is also uniformly integrable, so that (2.8) holds by, for example, [55], Theorem 1.11.3.

2.2. *Inner-product structure.* The two sides of (2.5) are quadratic in f and (2.5) is endowed with an inner-product structure. Indeed, if $f, h : \mathbb{R}^n \to \mathbb{R}^n$ satisfy the assumption of Theorem 2.1(iii), so do $f + h$ and $f - h$. Appliying (2.5) to $f + h$ and $f - h$, taking the difference and dividing by 4 yield

$$\begin{aligned} &\mathbb{E}\big[\big\{z^\top f(z) - \operatorname{div} f(z)\big\}\big\{z^\top h(z) - \operatorname{div} h(z)\big\}\big] \\ &\quad = \mathbb{E}\big[f(z)^\top h(z) + \mathrm{trace}\big(\nabla f(z) \nabla h(z)\big)\big] \end{aligned} \tag{2.9}$$

thanks to $\mathrm{trace}(\nabla f(z) \nabla h(z)) = \mathrm{trace}(\nabla h(z) \nabla f(z))$.

2.3. *Extensions and upper bounds on variance.* The second-order Stein formula (2.3) lets us derive the variance of more general random variables of the form

$$z^\top f(z) - \sigma^2 \operatorname{div} f(z) - g(z) \tag{2.10}$$

and a related upper bound for the variance, where $z \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$ satisfy the smoothness and integrability conditions of respective dimensions as in Theorem 2.1. The formula for the variance of (2.10) can be viewed as an extension of (2.3) from $g = 0$ to general g. As a special case, this provides a formula and an upper bound for the variance of the divergence with $g(\boldsymbol{x}) = \boldsymbol{x}^\top f(\boldsymbol{x})$.

THEOREM 2.2. *Let $z \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, f be as in (2.7) and $g : \mathbb{R}^n \to \mathbb{R}$ such that $\boldsymbol{x} \mapsto g(\sigma \boldsymbol{x})$ belongs to $W^{1,2}(\gamma_n)$. Then the variance of the random variable in (2.10) satisfies*

$$\begin{aligned} &\mathrm{Var}\big(z^\top f(z) - \sigma^2 \operatorname{div} f(z) - g(z)\big) \\ &\quad = \mathbb{E}\big[\sigma^2 \|f(z) - \nabla g(z)\|^2 + \sigma^4 \mathrm{trace}\big((\nabla f(z))^2\big)\big] + \tilde{V}(g) \qquad (2.11) \\ &\quad \leq \mathbb{E}\big[\sigma^2 \|f(z) - \nabla g(z)\|^2 + \sigma^4 \mathrm{trace}\big((\nabla f(z))^2\big)\big] \qquad (2.12) \end{aligned}$$

with $\tilde{V}(g) = \mathrm{Var}(g(z)) - \sigma^2 \mathbb{E}[\|\nabla g(z)\|^2]$. *In particular, with* $g(\boldsymbol{x}) = \boldsymbol{x}^\top f(\boldsymbol{x})$,

$$
\begin{aligned}
\mathrm{Var}[\mathrm{div}\, f(z)] &= \mathbb{E}[\mathrm{trace}((\nabla f(z))^2)] + \mathrm{Var}[z^\top f(z)/\sigma^2] \\
&\quad - \mathbb{E}[\|f(z)/\sigma\|^2] - 2\mathbb{E}[f(z)^\top (\nabla f(z)) z/\sigma^2] \\
&\le \mathbb{E}[\mathrm{trace}((\nabla f(z))^2)] + \mathbb{E}[\|(\nabla f(z)) z/\sigma\|^2].
\end{aligned} \tag{2.13}
$$

PROOF. Assume without loss of generality $\sigma = 1$. By (2.4),

$$
\mathbb{E}[(z^\top f(z) - \mathrm{div}\, f(z)) g(z)] = \mathbb{E}[f(z)^\top \nabla g(z)],
$$

so that (2.11) follows from Theorem 2.1 and some algebra. The Gaussian Poincaré inequality [14], 3.20, applied to g reads $\tilde{V}(g) \le 0$ which gives (2.12). Finally, for $g(\boldsymbol{x}) = \boldsymbol{x}^\top f(\boldsymbol{x})$, we have $(\partial g/\partial x_i)(z) = f_i(z) + \sum_{j=1}^n z_j (\partial f_j/\partial x_i)(z)$ so that $\nabla g(z) = f(z) + (\nabla f(z)) z$. □

As we have mentioned earlier, (2.11) becomes (2.7) when $g = 0$. The extension in Theorem 2.2 is particularly useful in our investigation of the variance of SURE and other quantities of interest. A striking feature of the upper bound in (2.13) is that the variance of the random variable $\mathrm{div}\, f(z)$, defined using the first-order derivatives of f, can be bounded from above using only the first-order partial derivatives of f. In particular, the second partial derivatives of f may be arbitrarily large or may not exist. This feature will be used in the next section to study the variance of the size of the model selected by the Lasso in linear regression, which takes the form $\mathrm{div}\, f$ for a certain function f.

It can be seen from the definition of $\tilde{V}(g)$ that the inequality (2.12) involves a single application of the Gaussian Poincaré inequality to $g(z)$. Thus, it holds with equality if and only if $g(\boldsymbol{x})$ is linear.

Finally, we obtain the following by applying (2.13) of Theorem 2.2 with the function $f(\boldsymbol{x})$ replaced by $f(\boldsymbol{x}) - \mathbb{E}[\nabla f(z)]\boldsymbol{x}$:

$$
\begin{aligned}
&\mathrm{Var}[\mathrm{div}\, f(z)] \\
&\quad \le \mathbb{E}[\mathrm{trace}((\nabla f(z) - \mathbb{E}[\nabla f(z)])^2)] + \mathbb{E}[\|\{\nabla f(z) - \mathbb{E}[\nabla f(z)]\} z/\sigma\|^2].
\end{aligned} \tag{2.14}
$$

This recovers $\mathrm{Var}[\mathrm{div}\, f(z)] = 0$ when $f(z)$ is linear in z and may be useful in other cases where $\nabla f(z) - \mathbb{E}[\nabla f(z)]$ is small.

2.4. *Beyond Gaussian distributions.* Although the next section provides applications of (2.5) and (2.12) in the Gaussian case only, we briefly mention here extensions to non-Gaussian z when z has density $\boldsymbol{x} \to \exp(-\psi(\boldsymbol{x}))$ where $\psi : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable with Hessian $H(\boldsymbol{x}) = \nabla^2 \psi(\boldsymbol{x})$. Throughout, let $h, f : \mathbb{R}^n \to \mathbb{R}^n$ be smooth vector fields with compact support and $g : \mathbb{R}^n \to \mathbb{R}$ be smooth. By integration by parts,

$$
\int \{\nabla \psi(z)^\top h(z) - \mathrm{div}\, h(z)\} g(z) e^{-\psi(z)}\, dz = \int h(z)^\top \nabla g(z) e^{-\psi(z)}\, dz.
$$

Applying this identity twice, first to $h = f$ and $g(\boldsymbol{x}) = \nabla \psi(\boldsymbol{x})^\top f(\boldsymbol{x}) - \mathrm{div}\, f(\boldsymbol{x})$, second to $h(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^\top f(z)$ and $g = 1$ gives

$$
\begin{aligned}
&\mathbb{E}[\{\nabla \psi(z)^\top f(z) - \mathrm{div}\, f(z)\}^2] \\
&\quad = \mathbb{E}[f(z)^\top \{H(z)\} f(z) + \mathrm{trace}\{\nabla f(z)^2\}]
\end{aligned} \tag{2.15}
$$

which extends (2.3) to non-Gaussian distributions. Similar to Theorem 2.2, if ψ is additionally strictly convex then

$$
\begin{aligned}
&\mathrm{Var}(\nabla \psi(z)^\top f(z) - \mathrm{div}\, f(z) - g(z)) \\
&\quad = \mathbb{E}[\|H(z)^{1/2} f(z) - H(z)^{-1/2} \nabla g(z)\}\|^2 + \mathrm{trace}((\nabla f(z))^2)] + \tilde{V}(g),
\end{aligned} \tag{2.16}
$$

where $\tilde{V}(g) = \mathrm{Var}(g(z)) - \mathbb{E}[\|H(z)^{-1/2}\nabla g(z)\|^2]$ and $\tilde{V}(g) \le 0$ follows from the Brasecamp–Lieb inequality [2], Theorem 4.9.1. In the special case where $f = \nabla u$ for some smooth $u : \mathbb{R}^n \to \mathbb{R}$, identity (2.15) is related to the Bochner formula used in the analysis of diffusion operators [2], Section 1.16.1, which gives hope to extend (2.3) and its applications to probability measures defined on non-Euclidean manifolds—although this goal lies outside of the scope of the present paper.

3. Statistical applications.

3.1. *SURE for SURE.* In the Gaussian sequence model, one observes $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ where the noise $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n)$ is standard normal and $\boldsymbol{\mu}$ is an unknown mean. Given an estimator $\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ where $\widehat{\boldsymbol{\mu}} : \mathbb{R}^n \to \mathbb{R}^n$ is some known almost differentiable function with $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ in L_1, SURE provides an unbiased estimate of the mean squared risk $\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ given by

$$\widehat{\mathsf{SURE}} = \|\widehat{\boldsymbol{\mu}} - \boldsymbol{y}\|^2 + 2\,\mathrm{div}\,\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) - n. \tag{3.1}$$

The fact that this quantity is an unbiased estimate of $\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is a consequence of the identity

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \|\widehat{\boldsymbol{\mu}} - \boldsymbol{y}\|^2 + 2\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \|\boldsymbol{\varepsilon}\|^2$$

with $\mathbb{E}[\|\boldsymbol{\varepsilon}\|^2] = n$ and Stein's formula (1.1) which asserts that $\mathbb{E}[\boldsymbol{\varepsilon}^\top\widehat{\boldsymbol{\mu}}(\boldsymbol{y})] = \mathbb{E}[\mathrm{div}\,\widehat{\boldsymbol{\mu}}(\boldsymbol{y})]$ whenever all partial derivatives of $\widehat{\boldsymbol{\mu}}$ are in L_1. The random variable $\mathrm{div}\,\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ can be computed from the observed data since it only involves $\boldsymbol{y}$ as well as the partial derivatives of $\widehat{\boldsymbol{\mu}}$. The quantity $\hat{\mathsf{df}} = \mathrm{div}\,\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ is an estimator sometimes referred to as the empirical degrees of freedom of the estimator $\widehat{\boldsymbol{\mu}}$. In this subsection, we develop second-order Stein methods to evaluate the accuracy of SURE.

3.1.1. *SURE for SURE in general.* We define the mean squared risk of the scalar estimator $\widehat{\mathsf{SURE}}$ by

$$R_{\mathrm{SURE}} = \mathbb{E}\big[\big(\widehat{\mathsf{SURE}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\big)^2\big]. \tag{3.2}$$

This means we treat $\widehat{\mathsf{SURE}}$ as an estimate of the squared prediction error $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ as well. This is reasonable as the actual squared loss $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is often a more relevant target than its expectation. One may also wish to treat $\widehat{\mathsf{SURE}}$ as an estimate of the deterministic $\mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$ and consider the estimation of $\mathrm{Var}(\widehat{\mathsf{SURE}})$. Thanks to Theorem 2.2, we will develop in the next subsection methodologies for the consistent estimation of upper bounds for $\mathrm{Var}(\widehat{\mathsf{SURE}})$ and R_{SURE} under proper conditions on the gradient of $\widehat{\boldsymbol{\mu}}$.

SURE is widely used in practice to estimate $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ or $\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ either because it is of interest to estimate the prediction error of $\widehat{\boldsymbol{\mu}}$, or because several estimators of the mean vector $\boldsymbol{\mu}$ are available and the statistician hopes to use the $\widehat{\mathsf{SURE}}$ to compare them on equal footing. Although $\widehat{\mathsf{SURE}}$ provides an unbiased estimate of the loss $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ and its expectation, such estimate may end up being not so useful, or provide spurious estimates, if the quantity (3.2) is too large. For estimators of interest where $\widehat{\mathsf{SURE}}$ is used in practice, it is important to understand the risk of $\widehat{\mathsf{SURE}}$ given by (3.2) in order to provide some uncertainty quantification about the accuracy of $\widehat{\mathsf{SURE}}$. For instance, one should expect $\widehat{\mathsf{SURE}}$ to be successful if $R_{\mathrm{SURE}}^{1/2}$ is negligible compared to $\widehat{\mathsf{SURE}}$, that is, $R_{\mathrm{SURE}}^{1/2} \lll \widehat{\mathsf{SURE}}$. On the other hand, if $R_{\mathrm{SURE}}^{1/2} \ggg \widehat{\mathsf{SURE}}$ then we would expect that estimates from $\widehat{\mathsf{SURE}}$ would be spurious with strictly positive probability and $\widehat{\mathsf{SURE}}$ should not be trusted. Under the square integrability condition on the first and second partial derivatives of $f(\boldsymbol{y})$, [46] proposed an unbiased estimate of the risk (3.2). However, the twice differentiability condition typically fails to hold for estimators involving less smooth regularizers such as the Lasso. [23] studied the performance of SURE optimized separable threshold estimator (SureShrink), and thus the accuracy of SURE in this special case. [24] derives an identity for the quantity (3.2) in the special case of the Lasso. In a general study of SURE tuned estimators, [51] developed a correction for the excess optimism with the nominal SURE in such schemes. Section 5 in [31] establishes consistency of SURE for the Lasso when the design matrix has i.i.d. $N(0, 1)$ entries and the tuning parameter is large enough. The second-order Stein identity

$$\mathbb{E}\big[\{\|z\|^2 - n + \gamma(z)\}^2\big] = \mathbb{E}\big[2p + 2\Delta\gamma(z) + \gamma(z)^2\big],$$

where $\Delta = \sum_{i=1}^n (\partial/\partial x_i)^2$ is the Laplacian, was used in [32] to prove the inadmissibility of SURE for the estimation of the squared loss of the James–Stein estimator when $n \geq 5$.

The following result, which extends Theorem 3 of [46] to allow application to the Lasso and other estimators only one-time differentiable, computes the expectation of the quantity (3.2) as well as an unbiased estimator of it directly through Theorem 2.1.

THEOREM 3.1. *Let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n)$ and $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$. Let $\widehat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$ with $\widehat{\boldsymbol{\mu}} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying the assumptions of (2.7), and define $\widehat{\mathsf{SURE}}$ by (3.1). Then*

$$\begin{aligned} &\mathbb{E}\big[\big(\widehat{\mathsf{SURE}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\big)^2\big] \\ &\quad = \mathbb{E}\big[4\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}(\boldsymbol{y})\|^2 + 4\,\mathrm{trace}\big((\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y}))^2\big) - 2n\big]. \end{aligned} \tag{3.3}$$

Consequently, SURE for SURE

$$\widehat{R}_{\mathrm{SURE}} = 4\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}(\boldsymbol{y})\|^2 + 4\,\mathrm{trace}\big((\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y}))^2\big) - 2n \tag{3.4}$$

is an unbiased estimate of the risk of $\widehat{\mathsf{SURE}}$ in (3.2).

REMARK 3.1. With the notation $f(\boldsymbol{y}) = \widehat{\boldsymbol{\mu}} - \boldsymbol{y}$ used in [46], $\nabla f(\boldsymbol{y}) = \nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) - \boldsymbol{I}_n$ and SURE for SURE is also given by

$$\widehat{R}_{\mathrm{SURE}} = 2n + 4\|f(\boldsymbol{y})\|^2 + 4\,\mathrm{trace}\big((\nabla f(\boldsymbol{y}))^2\big) + 8\,\mathrm{div}\, f(\boldsymbol{y}). \tag{3.5}$$

PROOF. Write $\mathrm{div}\,\widehat{\boldsymbol{\mu}}$ for $\mathrm{div}\,\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ and $\widehat{\boldsymbol{\mu}}$ for $\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$. By simple algebra,

$$\big(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \widehat{\mathsf{SURE}}\big)^2 = \big(\boldsymbol{\varepsilon}^\top\{2(\widehat{\boldsymbol{\mu}} - \boldsymbol{y}) + \boldsymbol{\varepsilon}\} + (n - 2\,\mathrm{div}\,\widehat{\boldsymbol{\mu}})\big)^2.$$

By (2.7) applied to $\boldsymbol{y} \to 2\{\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) - \boldsymbol{y}\} + (\boldsymbol{y} - \boldsymbol{\mu})$, we obtain in expectation

$$\begin{aligned} R_{\mathrm{SURE}} &= \mathbb{E}\big[\|2\{\widehat{\boldsymbol{\mu}} - \boldsymbol{y}\} + \boldsymbol{\varepsilon}\|^2 + \mathrm{trace}\{(2\nabla\widehat{\boldsymbol{\mu}} - \boldsymbol{I}_n)^2\}\big] \\ &= \mathbb{E}\big[4\|\widehat{\boldsymbol{\mu}} - \boldsymbol{y}\|^2 + 4\,\mathrm{trace}\{(\nabla\widehat{\boldsymbol{\mu}})^2\} - 2n\big] \\ &\quad + \mathbb{E}\big[4\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}} - \boldsymbol{y}) + 4(n - \mathrm{div}\,\widehat{\boldsymbol{\mu}})\big]. \end{aligned} \tag{3.6}$$

The proof of (3.3) is complete as the last line is 0 in virtue of Stein's formula. Equality (3.5) is obtained by observing the following for $f(\boldsymbol{y}) = \widehat{\boldsymbol{\mu}}(\boldsymbol{y}) - \boldsymbol{y}$,

$$\nabla f(\boldsymbol{y}) = \nabla\widehat{\boldsymbol{\mu}} - \boldsymbol{I}_n, \qquad \mathrm{div}\, f(\boldsymbol{y}) = \mathrm{trace}[\nabla\widehat{\boldsymbol{\mu}}] - n,$$

$$4\,\mathrm{trace}\big[(\nabla f(\boldsymbol{y}))^2\big] + 8\,\mathrm{div}\, f(\boldsymbol{y}) = 4\,\mathrm{trace}\big[(\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y}))^2\big] - 4n. \qquad \square$$

REMARK 3.2. In the Gaussian sequence model where the noise $\boldsymbol{\varepsilon}$ has distribution $N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ with $\sigma \neq 1$, the estimator $\widehat{\mathsf{SURE}}$ has the form

$$\widehat{\mathsf{SURE}} = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + 2\sigma^2\,\mathrm{div}\,\widehat{\boldsymbol{\mu}} - \sigma^2 n. \tag{3.7}$$

Theorem 3.1 implies that in this setting, SURE for SURE is

$$\widehat{R}_{\mathrm{SURE}} = 4\sigma^2\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + 4\sigma^4\,\mathrm{trace}\big[(\nabla\widehat{\boldsymbol{\mu}})^2\big] - 2\sigma^4 n, \tag{3.8}$$

as its expectation is identical to $R_{\mathrm{SURE}} = \mathbb{E}[(\widehat{\mathsf{SURE}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2)^2]$.

3.1.2. *Error bounds and consistency.* In the spirit of SURE, SURE for SURE in (3.8) is fundamentally a point estimator for the risk of SURE as defined in (3.2). It could be the starting point for the construction of an interval estimator or simply provide some measure of the performance of SURE when nothing else or better is available. Because of its availability in broad settings, (3.8) is not expected to always yield sensible and theoretically justifiable interval estimates. Still, we provide here consistent estimation of upper bounds for both the risk (3.2) and $\mathrm{Var}(\widehat{\mathsf{SURE}})$ under proper conditions on $\widehat{\boldsymbol{\mu}}$. Furthermore, when the gradient is a random orthogonal projection as in Lasso, isotonic regression and many other cases, the theorem below proves the consistency of SURE for SURE in (3.8). Define

$$\widehat{R}'_{\mathrm{SURE}} = 2\sigma^2\big(\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2 + \widehat{\mathsf{SURE}}\big), \tag{3.9}$$

which satisfies $\widehat{R}'_{\mathrm{SURE}} - \widehat{R}_{\mathrm{SURE}} = 4\sigma^4\{\operatorname{div}\widehat{\boldsymbol{\mu}} - \operatorname{trace}((\nabla\widehat{\boldsymbol{\mu}})^2)\}$ compared with SURE for SURE in (3.8). In particular, $\widehat{R}'_{\mathrm{SURE}} = \widehat{R}_{\mathrm{SURE}}$ if $\nabla\widehat{\boldsymbol{\mu}}$ is a random orthogonal projection.

THEOREM 3.2. *Let* $\widehat{\mathsf{SURE}}$, R_{SURE}, $\widehat{R}_{\mathrm{SURE}}$ *and* $\widehat{R}'_{\mathrm{SURE}}$ *be as in* (3.7), (3.2), (3.8) *and* (3.9), *respectively, and assume* $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$.

(i) *If* $\mathbb{E}\sum_{i=1}^n |(\nabla\widehat{\boldsymbol{\mu}})_{ii}| < +\infty$, *then* $\mathbb{E}[\widehat{R}'_{\mathrm{SURE}}] \geq \sigma^4 n$.

(ii) *If the function* $\boldsymbol{y} \to \widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ *is* 1-*Lipschitz*, *then*

$$\mathbb{E}\big[\big((\widehat{\mathsf{SURE}})_+{}^{1/2} - \mathbb{E}\big[\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2\big]^{1/2}\big)^4\big]^{1/4} \leq R_{\mathrm{SURE}}^{1/4} + 3\sigma. \tag{3.10}$$

(iii) *If the function* $\boldsymbol{y} \to \widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ *is* 1-*Lipschitz and* $\nabla\widehat{\boldsymbol{\mu}}$ *is almost everywhere symmetric positive semidefinite*, *then* $\widehat{R}_{\mathrm{SURE}} \leq \widehat{R}'_{\mathrm{SURE}}$ *and*

$$\mathrm{Var}(\widehat{\mathsf{SURE}}) = \mathbb{E}\big[\big(\widehat{\mathsf{SURE}} - \mathbb{E}\big[\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2\big]\big)^2\big] \leq R_{\mathrm{SURE}} + \sigma^4 n. \tag{3.11}$$

(iv) *If* $\widehat{R}'_{\mathrm{SURE}} \geq \widehat{R}_{\mathrm{SURE}}$ *and* $\max_{\|\boldsymbol{u}\|=1} |\boldsymbol{u}^\top(\boldsymbol{I}_n - \nabla\widehat{\boldsymbol{\mu}})\boldsymbol{u}| \leq 1$ *almost surely*, *then*

$$\mathbb{E}\bigg[\bigg(\frac{\widehat{R}'_{\mathrm{SURE}}}{\mathbb{E}[\widehat{R}'_{\mathrm{SURE}}]} - 1\bigg)^2\bigg] \leq \frac{16\sigma^4}{\mathbb{E}[\widehat{R}'_{\mathrm{SURE}}]} \leq \frac{16}{n}. \tag{3.12}$$

REMARK 3.3. If $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ for some fixed $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and a penalized estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p}\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|^2/2 + g(\boldsymbol{b})\}$ for convex $g: \mathbb{R}^p \to \mathbb{R}$, then $\widehat{\boldsymbol{\mu}}$ is 1-Lipschitz and $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ is almost everywhere symmetric positive semidefinite. This follows from arguments in [3, 9] as explained at the end of Appendix B. Hence for any convex regularized least-squares estimate $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$, $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ satisfies the conditions in (ii)–(iv) above as well as $\widehat{R}_{\mathrm{SURE}} \leq \widehat{R}'_{\mathrm{SURE}}$ almost surely.

If the estimation of the deterministic quantity $\mathbb{E}[\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2]$ is essential, inequality (3.10) asserts that up to an additive absolute constant, R_{SURE} bounds from above the quartic risk of $\widehat{\mathsf{SURE}}^{1/2}$ when the estimation target is $\mathbb{E}[\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2]^{1/2}$. If in addition the gradient is almost everywhere symmetric positive semidefinite as in Remark 3.3, inequality (3.11) asserts that the risk $\mathrm{Var}(\widehat{\mathsf{SURE}})$ is bounded from above by $R_{\mathrm{SURE}} + \sigma^4 n \leq 2\mathbb{E}[\widehat{R}'_{\mathrm{SURE}}]$, with a slight modification $\widehat{R}'_{\mathrm{SURE}}$ of the SURE for SURE. Moreover, under milder conditions in Theorem 3.2(iv) which do not require the symmetry of the random matrix $\nabla\widehat{\boldsymbol{\mu}}$, $\widehat{R}'_{\mathrm{SURE}}$ is a consistent estimator of its expectation.

When the gradient $\nabla\widehat{\boldsymbol{\mu}}$ is a random orthogonal projection, $\widehat{R}'_{\mathrm{SURE}}$ is identical to $\widehat{R}_{\mathrm{SURE}}$, so that SURE for SURE is a consistent estimator of its risk R_{SURE} and $R_{\mathrm{SURE}} + \sigma^4 n \leq 2R_{\mathrm{SURE}}$ are upper bounds for the risk $\mathrm{Var}(\widehat{\mathsf{SURE}})$. In the lasso case, $\nabla\widehat{\boldsymbol{\mu}}$ is an orthogonal projection; cf. Section 3.5.1 below.

The proof of Theorem 3.2 is given in Appendix B. In fact, under the conditions for (3.12), we prove the sharper

$$\operatorname{Var}(\widehat{R}'_{\mathrm{SURE}}) \le 16\sigma^4 \mathbb{E}[\widehat{R}''_{\mathrm{SURE}}] \tag{3.13}$$

with $\widehat{R}''_{\mathrm{SURE}} = (3/4)\widehat{R}'_{\mathrm{SURE}} + (1/4)\widehat{R}_{\mathrm{SURE}} - \sigma^4 \operatorname{div} \widehat{\boldsymbol{\mu}} \le \widehat{R}'_{\mathrm{SURE}}$. This suggests the use of $16\sigma^4 \widehat{R}''_{\mathrm{SURE}}$ or $16\sigma^4 \widehat{R}'_{\mathrm{SURE}}$ as estimated upper bounds for $\operatorname{Var}(\widehat{R}'_{\mathrm{SURE}})$.

3.1.3. *Difference of two estimators.* As we have briefly discussed above, the statement of Theorem 3.1, SURE is often used to optimize among different estimators. Consider for simplicity the comparison between two estimates $\widehat{\boldsymbol{\mu}}^{(1)}$ and $\widehat{\boldsymbol{\mu}}^{(2)}$ of $\boldsymbol{\mu}$. In this setting,

$$R^{(\mathrm{diff})}_{\mathrm{SURE}} = \mathbb{E}\big[\big(\|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{\mu}\|^2 - \widehat{\mathsf{SURE}}^{(\mathrm{diff})}\big)^2\big] \tag{3.14}$$

is the proper risk for SURE, where

$$\begin{aligned} \widehat{\mathsf{SURE}}^{(\mathrm{diff})} &= \widehat{\mathsf{SURE}}^{(1)} - \widehat{\mathsf{SURE}}^{(2)} \\ &= \|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{y}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{y}\|^2 + 2 \operatorname{div}(\widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)}) \end{aligned} \tag{3.15}$$

is the difference in SURE between $\widehat{\boldsymbol{\mu}}^{(1)}$ and $\widehat{\boldsymbol{\mu}}^{(2)}$. When the loss $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is of smaller order than $n^{1/2}$, SURE may produce a spurious estimator due to the estimation of $\|\boldsymbol{\varepsilon}\|^2$ by n in (3.1). However, due to the cancellation of this common chi-square type error, the risk of the estimator (3.15) could be of smaller order than the risk of SURE for both $\widehat{\boldsymbol{\mu}}^{(j)}$. Parallel to Theorem 3.1, the second-order Stein formula leads to the following.

THEOREM 3.3. *Let $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_n)$, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, and $\widehat{\boldsymbol{\mu}}^{(1)}$ and $\widehat{\boldsymbol{\mu}}^{(2)}$ be estimates of $\boldsymbol{\mu}$ based on $\boldsymbol{y}$. Let $\widehat{\mathsf{SURE}}^{(\mathrm{diff})}$ and $R^{(\mathrm{diff})}_{\mathrm{SURE}}$ be as in (3.15) and (3.14), respectively, and $f(\boldsymbol{y}) = \widehat{\boldsymbol{\mu}}^{(1)} - \widehat{\boldsymbol{\mu}}^{(2)}$. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies the assumptions of (2.7). Then*

$$R^{(\mathrm{diff})}_{\mathrm{SURE}} = \mathbb{E}\big[4\|f(\boldsymbol{y})\|^2 + 4 \operatorname{trace}\big((\nabla f(\boldsymbol{y}))^2\big)\big]. \tag{3.16}$$

Consequently, SURE for SURE, given by

$$\widehat{R}^{(\mathrm{diff})}_{\mathrm{SURE}} = 4\|f(\boldsymbol{y})\|^2 + 4 \operatorname{trace}\big((\nabla f(\boldsymbol{y}))^2\big), \tag{3.17}$$

is an unbiased estimate of the risk of $\widehat{\mathsf{SURE}}^{(\mathrm{diff})}$ in (3.14).

PROOF. By algebra,

$$\|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{\mu}\|^2 - \widehat{\mathsf{SURE}}^{(\mathrm{diff})} = 2\boldsymbol{\varepsilon}^\top f(\boldsymbol{y}) - 2 \operatorname{div} f(\boldsymbol{y}).$$

The conclusion follows directly from Theorem 2.1. □

3.2. *Confidence region based on $\widehat{\mathsf{SURE}}$.* While SURE for SURE provides an unbiased point estimator for the (mean) squared difference between $\widehat{\mathsf{SURE}}$ and the squared loss $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$, we may also use the second-order Stein formula to derive interval estimates for $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ and $\mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$ based on $\widehat{\mathsf{SURE}}$. As we are not compelled to directly use the $\widehat{R}_{\mathrm{SURE}}$ in (3.8) to construct such interval estimates, we present the following simpler approach.

THEOREM 3.4. *Let $\boldsymbol{y}$, $\boldsymbol{\mu}$, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ and $\widehat{\mathsf{SURE}}$ be as in (3.7). Then*

$$\begin{aligned} &\mathbb{E}\big[\big(\widehat{\mathsf{SURE}} - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 - \|\boldsymbol{\varepsilon}\|^2 + \sigma^2 n\big)^2\big] \\ &\quad = 4\sigma^2 \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] + 4\sigma^4 \mathbb{E}\big[\operatorname{trace}\big((\nabla \widehat{\boldsymbol{\mu}})^2\big)\big]. \end{aligned} \tag{3.18}$$

If the right-hand side of (3.18) *is bounded by* $\sigma^4 v_0^2 2n\epsilon_n$ *with a constant* v_0, *then*

$$\mathbb{P}\{|\widehat{\mathsf{SURE}} - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2| \le \sigma^2(v_\alpha + v_0)\sqrt{2n}\} \ge 1 - \alpha - \epsilon_n \tag{3.19}$$

for all $\alpha \in (0, 1)$, *where* v_α *is defined by* $\mathbb{P}\{(2n)^{-1/2}|\chi_n^2 - n| > v_\alpha\} = \alpha$, *and*

$$\mathbb{P}\{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 \le \widehat{\mathsf{SURE}} + \sigma^2(v_{-,\alpha} + v_0)\sqrt{2n}\} \ge 1 - \alpha - \epsilon_n, \tag{3.20}$$

where $v_{-,\alpha}$ *is defined by* $\mathbb{P}\{(2n)^{-1/2}(n - \chi_n^2) > v_{-,\alpha}\} = \alpha$.

While the left-hand side of (3.18) is quartic in $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$, the right-hand side is quadratic. Thus, $\widehat{\mathsf{SURE}}$ provides an accurate estimate of $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2$ when the squared error is of greater order than $|\|\boldsymbol{\varepsilon}\|^2 - \sigma^2 n| \approx \sigma^2(2n)^{1/2}|N(0, 1)|$, provided that the second term on the right-hand side of (3.18) is of no greater order than $\max\{\sigma^4 n, \sigma^2\mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]\}$. Specifically, in such scenarios, (3.19) implies that $\widehat{\mathsf{SURE}}$ is within a small fraction of $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2$ when $\sqrt{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2/n}$ is of greater order than $\sigma n^{-1/4}$, and (3.19) and (3.20) provide confidence regions for the entire vector $\boldsymbol{\mu}$. As $\sigma n^{-1/4}$ is known to be a lower bound for the error in the estimation of the average loss in the estimation of $\boldsymbol{\mu}$ [37, 41], (3.19) implies the rate optimality of the upper bound (3.20) for the squared estimation error $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2$, and thus the rate optimality of the resulting confidence region for $\boldsymbol{\mu}$.

COROLLARY 3.5. *If for some sequence* (γ_n) *with* $\gamma_n \to 0$,

$$\mathbb{E}[4\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/(n\sigma^2) + 4\,\mathrm{trace}\{(\nabla\widehat{\boldsymbol{\mu}})^2\})/n] \le \gamma_n, \tag{3.21}$$

then for all fixed $\alpha \in (0, 1)$ *independent of* n,

$$\mathbb{P}\{|\widehat{\mathsf{SURE}} - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2| \le \sigma^2(v_\alpha)\sqrt{2n}\} \to (1 - \alpha),$$
$$\mathbb{P}\{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 \le \widehat{\mathsf{SURE}} + \sigma^2(v_{-,\alpha})\sqrt{2n}\} \to (1 - \alpha),$$

where v_α, $v_{-,\alpha}$ *are defined in Theorem* 3.4.

Similar to Theorem 3.4, Corollary 3.5 provides confidence regions for the entire vector $\boldsymbol{\mu}$ with exact asymptotic quantiles. Under the condition (3.21), $\widehat{\mathsf{SURE}}$ incurs an error characterized by the quantiles of the random variable $\|\boldsymbol{\varepsilon}\|^2 - \sigma^2 n$.

We will verify in Theorems 3.11 and 3.13 that the condition (3.21) holds for the Lasso under commonly imposed regularity conditions in sparse regression theory.

Alternatively to Corollary 3.5, the following result replaces the condition on γ_n in (3.21) by a data-driven surrogate $\hat{\gamma}_n$ and provides nonasymptotic interval estimates for both $\mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$ and $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$.

THEOREM 3.6. *Suppose* $\widehat{\boldsymbol{\mu}}$ *is* 1*-Lipschitz and* $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ *is almost surely symmetric positive semidefinite. Let* $\widehat{\mathsf{SURE}}$ *be as in* (3.7). *Then*

$$\mathrm{Var}(\widehat{\mathsf{SURE}} - \mathbb{E}[\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2] + (n\sigma^2 - \|\boldsymbol{\varepsilon}\|^2)) \le 4\mathbb{E}[\sigma^2\widehat{\mathsf{SURE}} + \sigma^4\hat{\mathsf{df}}], \tag{3.22}$$

where $\hat{\mathsf{df}} = \mathrm{div}\,\widehat{\boldsymbol{\mu}}$. *Moreover, there exist nonnegative random variables* X_n *and* Y_n *with* $\mathbb{E}[X_n^2] = \mathbb{E}[Y_n^2] = 1$ *such that*

$$\begin{aligned} &|\widehat{\mathsf{SURE}} - \mathbb{E}[\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2] + (n\sigma^2 - \|\boldsymbol{\varepsilon}\|^2)|/(\sigma^2\sqrt{n}) \\ &\quad \le X_n[\hat{\gamma}_n^{1/2} + 4Y_n/n^{1/2} + \sqrt{4Y_n}(6/n)^{1/4}], \end{aligned} \tag{3.23}$$

where $\hat{\gamma}_n = 4\{\widehat{\mathsf{SURE}}/(n\sigma^2) + \operatorname{trace}[\nabla\widehat{\boldsymbol{\mu}}]/n\}_+$. *Consequently for positive* α, β_1 *and* β_2 *with* $\alpha + \beta_1 + \beta_2 = \delta < 1$ *and* $\kappa_{n,\beta_2} = 2(6/(\beta_2 n))^{1/4} + 4/(\beta_2 n)^{1/2}$,

$$\mathbb{P}\left\{\frac{|\widehat{\mathsf{SURE}} - \mathbb{E}[\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2]|}{\sigma^2\sqrt{2n}} \le v_\alpha + \frac{\hat{\gamma}_n^{1/2} + \kappa_{n,\beta_2}}{(2\beta_1)^{1/2}}\right\} \ge 1 - \delta. \tag{3.24}$$

Moreover, (3.23) *and* (3.24) *still hold with* $\mathbb{E}[\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2]$ *replaced by* $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2$.

As discussed in Remark 3.3, the conclusions of Theorem 3.6 are applicable to all convex penalized estimators in linear regression. Note that $\hat{\gamma}_n$ is a biased estimate of the left-hand side of (3.21) in general due to $\operatorname{trace}[\{\nabla\widehat{\boldsymbol{\mu}}\}^2] \le \operatorname{trace}\nabla\widehat{\boldsymbol{\mu}}$. The estimate $\hat{\gamma}_n$ involves $\hat{\mathsf{df}} = \operatorname{trace}[\nabla\widehat{\boldsymbol{\mu}}]$ instead of $\operatorname{trace}[\{\nabla\widehat{\boldsymbol{\mu}}\}^2]$ because controlling the variance of $\hat{\mathsf{df}}$ easily follows from Section 2.3, while we are not aware of available tools to bound the variance of $\operatorname{trace}[\{\nabla\widehat{\boldsymbol{\mu}}\}^2]$ except in specific cases where $\nabla\widehat{\boldsymbol{\mu}}$ is a projection. The proofs of Theorem 3.4, Corollary 3.5 and Theorem 3.6 are given in Appendix C.

3.3. *Oracle inequalities for SURE-tuned estimates.* Beyond pairwise comparisons, the following result provides guarantees on the SURE-tuned estimate $\widetilde{\boldsymbol{\mu}}$, which is obtained by selecting the estimator among $\{\widehat{\boldsymbol{\mu}}^{(1)}, \dots, \widehat{\boldsymbol{\mu}}^{(m)}\}$ with the smallest $\widehat{\mathsf{SURE}}$, that is,

$$\widetilde{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}^{(\hat{k})} \quad \text{with } \hat{k} = \arg\min_{j\in[m]} \widehat{\mathsf{SURE}}^{(j)}, \tag{3.25}$$

where $\widehat{\mathsf{SURE}}^{(j)} = \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{y}\|^2 + 2\sigma^2 \operatorname{trace}\nabla\widehat{\boldsymbol{\mu}}^{(j)} - n\sigma^2$ for $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$.

THEOREM 3.7. *Consider the sequence model* $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$. *Let* $\widehat{\boldsymbol{\mu}}^{(1)}(\boldsymbol{y}), \dots, \widehat{\boldsymbol{\mu}}^{(m)}(\boldsymbol{y})$ *be all L-Lipschitz functions of* $\boldsymbol{y}$, $\widetilde{\boldsymbol{\mu}}$ *the SURE tuned estimator in* (3.25), $j_0 = \arg\min_{j=1,\dots,m} \mathbb{E}\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|$, $s^* = \max_{k\in[m]} \mathbb{E}[\operatorname{trace}((\nabla\widehat{\boldsymbol{\mu}}^{(k)} - \nabla\widehat{\boldsymbol{\mu}}^{(j_0)})^2)]$. *Then*

(i) *For any* $\alpha \in (0, 1)$, *with probability at least* $1 - \alpha$,

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\| \le \sigma \max\{(8s^*m/\alpha)^{1/4}, (8m(\sqrt{2}L + 1)/\alpha)^{1/2}\}. \tag{3.26}$$

(ii) *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$\|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\| - \min_{j\in[m]} \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\| \le 2L\sigma\sqrt{2\log(m/\delta)} \tag{3.27}$$

so that the sum of (3.26) *and* (3.27) *provide high probability bounds on* $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{j\in[m]} \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|$.

(iii) *For some absolute constant* $C > 0$, *in expectation*

$$\mathbb{E}\Big[\Big(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{j\in[m]} \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|\Big)^2\Big]^{1/2} \le C\sigma[(s^*m)^{1/4} + (1 + L)m^{1/2}]. \tag{3.28}$$

(iv) *If* $\max_{j\in[m]} \mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2]/n \le L\sigma^2$, *then the squared risk enjoys*

$$\mathbb{E}[\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] - \min_{j\in[m]} \mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2] \le L\sigma^2(32nm)^{1/2}. \tag{3.29}$$

The proof is given in Appendix D. The assumption that the estimators $\widehat{\boldsymbol{\mu}}^{(j)}$ are L-Lipschitz functions of $\boldsymbol{y}$ is mild; cf. Remark 3.3. Under this assumption, $s^* \le 4L^2 n$ and (3.28) implies

$$\frac{\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|}{n^{1/2}} - \min_{j\in[m]} \frac{\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|}{n^{1/2}} \le C\sigma(1 + L)\max\left\{\left(\frac{m}{\alpha n}\right)^{1/4}, \left(\frac{m}{n\alpha}\right)^{1/2}\right\} \tag{3.30}$$

with probability $1-\alpha$ for some absolute $C>0$, where we used the $n^{-1/2}$ scaling to feature the normalized prediction risk $\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2/n$. Theorem 3.7(i) can also be understood in terms of sample size requirement: If $\epsilon>0$ is a fixed precision target and $\alpha\in(0,1)$, then $n\gtrsim\epsilon^{-2}\max\{m/\alpha,(s^*m/\alpha)^{1/2}\}$ samples are sufficient to ensure

$$\mathbb{P}\Big\{n^{-1/2}\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|-\min_{j\in[m]}n^{-1/2}\|\widehat{\boldsymbol{\mu}}^{(j)}-\boldsymbol{\mu}\|\le\sigma\epsilon\Big\}\ge1-\alpha.$$

We are not aware of a previous result of this form that applies with the above level of generality, that is, with no restriction on the nature of the estimators $\{\widehat{\boldsymbol{\mu}}^{(1)},\dots,\widehat{\boldsymbol{\mu}}^{(m)}\}$ beyond the Lipschitz requirement. As shown in the next proposition, the dependence in s^* in the term $(s^*m/\alpha)^{1/4}$ of (3.26) and the dependence in n in the term $(m/(n\alpha))^{1/4}$ of (3.30) are unimprovable without additional assumptions.

PROPOSITION 3.8. *There exist absolute constants $C_1,C_2,C_3>0$ such that for any $n\ge C_1$, there exist $\boldsymbol{\mu}\in\mathbb{R}^n$ and two estimators $\widehat{\boldsymbol{\mu}}^{(1)}(\boldsymbol{y})$, $\widehat{\boldsymbol{\mu}}^{(2)}(\boldsymbol{y})$ that are 1-Lipschitz functions of $\boldsymbol{y}$ such that $s^*=n$ and $\widetilde{\boldsymbol{\mu}}$ in (3.25) satisfies*

$$\mathbb{P}\Big\{\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|-\min_{j=1,2}\|\widehat{\boldsymbol{\mu}}^{(j)}-\boldsymbol{\mu}\|\ge C_2\sigma n^{1/4}\Big\}\ge C_3.$$

Oracle inequalities stronger than (3.30) are available if the estimators $\{\widehat{\boldsymbol{\mu}}^{(j)},j=1,\dots,m\}$ are affine in $\boldsymbol{y}$, that is, of the form $\widehat{\boldsymbol{\mu}}^{(j)}=\boldsymbol{A}_j\boldsymbol{y}+\boldsymbol{b}_j$ for deterministic $\boldsymbol{A}_j\in\mathbb{R}^{n\times n}$ and $\boldsymbol{b}_j\in\mathbb{R}^n$. In this case, $\widehat{\mathsf{SURE}}^{(j)}=\|\widehat{\boldsymbol{\mu}}^{(j)}-\boldsymbol{y}\|^2+2\sigma^2\operatorname{trace}\boldsymbol{A}_j-n\sigma^2$ reduces to Mallows C_p [38] and if $\widetilde{\boldsymbol{\mu}}$ is the estimate among $\{\widehat{\boldsymbol{\mu}}^{(j)},j\in[m]\}$ with the smallest $\widehat{\mathsf{SURE}}$, then

$$\mathbb{E}\Big[\|\widetilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|-\min_{j\in[m]}\|\widehat{\boldsymbol{\mu}}^{(j)}-\boldsymbol{\mu}\|\Big]\le C(1+L)\sigma(\log m)^{1/2} \tag{3.31}$$

for some absolute constant $C>0$, provided that the operator norm of $\boldsymbol{A}_j$ is less than L for all $j\in[m]$; cf. [5], Proposition 3.1, or [1, 51] for related results. If $\boldsymbol{b}_j=0$ for all j and the matrices $\boldsymbol{A}_j$ are symmetric and totally ordered in the sense of positive semidefinite matrices, the right- hand side in (3.31) can even be reduced to $O(\sigma)$ [7, 33]. Equation (3.31) provides an oracle inequality with respect to the risk $\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|$; oracle inequalities of form (3.29) with respect to the squared risk are studied in [22, 36] using exponential weights procedures, and in [20, 21], [5], Theorem 2.1, using a convex relaxation of $\widehat{\mathsf{SURE}}$ named Q-aggregation and introduced in [43]. For linear estimators, these works exhibit an error term involving $\log m$ thanks to the availability of the Hanson–Wright inequality, which provides exponential concentration bounds for random variables of the form $\boldsymbol{\varepsilon}^\top(\boldsymbol{A}_j-\boldsymbol{A}_k)\boldsymbol{\varepsilon}$.

As the optimal remainder term for an oracle inequality of the form (3.29) satisfied by any estimator of the form $\widehat{\boldsymbol{\mu}}^{(\hat{j})}$, $\hat{j}\in[m]$ is of order $\sigma^2(n\log m)^{1/2}$ for deterministic vectors $\widehat{\boldsymbol{\mu}}^{(j)}=\boldsymbol{b}_j$ [44], Theorem 2.1, the dependence in n of (3.29) is optimal when m is smaller than an absolute constant.

Compared to these existing results, the novelty of Theorem 3.7 lies in its scope: Theorem 3.7 and (3.30) are applicable to to any collection of L-Lipschitz nonlinear estimators $\{\widehat{\boldsymbol{\mu}}^{(j)},j\in[m]\}$. The L-Lipschitz assumption is mild (cf. Remark 3.3) and Theorem 3.7 is thus applicable far beyond the case of linear estimators studied in the aforementioned literature.

A drawback of Theorem 3.7 compared to (3.31) is the suboptimality of the dependence in m. The difficulty to obtain a rate logarithmic in m is due to the unavailability of exponential concentration equalities for random variables of the form $\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}}^{(j)}-\widehat{\boldsymbol{\mu}}^{(k)})$ for nonlinear $\widehat{\boldsymbol{\mu}}^{(j)}$, $\widehat{\boldsymbol{\mu}}^{(k)}$.

Finally, we note that an oracle inequality for the squared risk can be obtained using the Q-aggregation procedure from [20, 21, 43] and its analysis in [5]. Indeed, let $\tilde{\boldsymbol{\mu}}_Q = \sum_{j=1}^m \hat{\theta}_j \hat{\boldsymbol{\mu}}^{(j)}$ where

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{\theta} = 1}{\arg\min} \left[\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{y}\|^2 + 2\sigma^2 \sum_{j=1}^m \theta_j \operatorname{div} \hat{\boldsymbol{\mu}}^{(j)} + \frac{1}{2} \sum_{j=1}^m \theta_j \|\hat{\boldsymbol{\mu}}^{(j)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\|^2 \right]$$

and $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \sum_{j=1}^m \theta_j \hat{\boldsymbol{\mu}}^{(j)}$ for every $\boldsymbol{\theta}$ in the simplex $\{\boldsymbol{\theta} \in \mathbb{R}^m : \boldsymbol{\theta} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{\theta} = 1\}$ in $\mathbb{R}^m$. Then $\tilde{\boldsymbol{\mu}}_Q$ satisfies under the assumptions of Theorem 3.7

$$\mathbb{E}[\|\tilde{\boldsymbol{\mu}}_Q - \boldsymbol{\mu}\|^2] - \min_{k \in [m]} \mathbb{E}[\|\hat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2] \leq 2\sigma^2 \big(\sqrt{s^* m} + (1+L)^2 m\big) + \sigma^2 L^2. \tag{3.32}$$

We refer to [5, 20, 21] for details on the construction of $\tilde{\boldsymbol{\mu}}_Q$. The proof of (3.32) is given at the end of Appendix D.

3.4. *The variance of the model size of the Lasso.* Consider a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.33}$$

where $\boldsymbol{\beta}$ is the true coefficient vector, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ is the noise and $\boldsymbol{X}$ is a deterministic design matrix. Consider the Lasso which solves the optimization problem

$$\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\max} \{ \|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}\|^2/(2n) + \lambda \|\boldsymbol{b}\|_1 \}. \tag{3.34}$$

Let $\hat{S} = \{j \in [p] : (\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_j \neq 0\}$ be the support of the Lasso. We are interested in the size of $\hat{S}$ denoted by $|\hat{S}|$. Even though the Lasso and sparse linear regression have been studied extensively in the last two decades, little is known about the stochastic behavior of the discrete random variable $|\hat{S}|$. Under the sparse Riesz or similar conditions, $|\hat{S}| \lesssim \|\boldsymbol{\beta}\|_0$ with high probability [59, 61, 63] but such results only imply a bound of the form $\operatorname{Var}[|\hat{S}|] \lesssim \|\boldsymbol{\beta}\|_0^2$ on the variance; we will see below that the variance of $|\hat{S}|$ is typically much smaller. There are trivial situations where the behavior of $|\hat{S}|$ is well understood: if λ is very large for instance, then $|\hat{S}| = 0$ with high probability. Or, under strong conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ that grants support recovery (cf. for instance, the conditions given in [40, 53, 56, 64]), $\hat{S} = \operatorname{supp}(\boldsymbol{\beta})$ holds with probability at least $1 - 1/p^2$ and in this case $\operatorname{Var}[|\hat{S}|] \leq \mathbb{E}[(|\hat{S}| - s_0)^2] \leq 1$.

Outside of these situations, studying $|\hat{S}|$ appears delicate; for instance, our previous attempts at studying the variance of $|\hat{S}|$ went as follows. Let $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$ be the canonical basis in $\mathbb{R}^p$ and let $\boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{e}_j$ for all $j = 1, \ldots, p$. The KKT conditions of the Lasso are given by

$$\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})/(n\lambda) \begin{cases} = \operatorname{sgn}((\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_j) & \text{if } (\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_j \neq 0, \\ \in [-1, 1] & \text{if } (\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_j = 0. \end{cases}$$

At a given point $\boldsymbol{y}$, to understand the stability of $\hat{S}$, a natural avenue is to identity how close the quantities $\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})/(n\lambda)$ are from ± 1 for the indices $j \notin \hat{S}$. If many indices $j \notin \hat{S}$ are such that $\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})/(n\lambda)$ is extremely close to ± 1, then a tiny variation in $\boldsymbol{y}$ may push some of the quantities $\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})/(n\lambda)$ toward ± 1 resulting in many new variables entering the support for this tiny variation in $\boldsymbol{y}$. The current model size $|\hat{S}|$ is noninformative about how many indices $j \notin \hat{S}$ are such that $\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})/(n\lambda)$ is extremely close to ± 1 and the random variable $|\hat{S}|$ appears prone to instability.

With the second-order Stein formula (2.3) and the tools developed in the previous section, the variance of $|\hat{S}|$ can be bounded as follows. First, we need to describe a condition on the deterministic matrix $\boldsymbol{X}$ which ensures that the KKT conditions of the Lasso hold strictly with probability 1. We say that the KKT conditions hold strictly if

$$\forall j \notin \widehat{S}, \quad -1 < \frac{1}{\lambda n} \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}) < 1. \tag{3.35}$$

ASSUMPTION 3.1. For all $\delta_1, \dots, \delta_p \in \{-1, 1\}$ and $1 \le j_0 < j_1 < \cdots < j_n \le p$,

$$\text{rank} \begin{pmatrix} \boldsymbol{x}_{j_0} & \boldsymbol{x}_{j_1} & \cdots & \boldsymbol{x}_{j_n} \\ \delta_{j_0} & \delta_{j_1} & \cdots & \delta_{j_n} \end{pmatrix}_{(n+1)\times(n+1)} = n + 1.$$

PROPOSITION 3.9. *If $\boldsymbol{X}$ satisfies the above assumption, then the set $B = \{j \in [p] : |\boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})| = \lambda n\}$ is such that $\boldsymbol{X}_B$ has rank $|B|$ and the solution $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ to the optimization problem* (3.34) *is unique. Furthermore, if $\mathbb{P}[\boldsymbol{v}^\top \boldsymbol{\varepsilon} = c] = 0$ for all vectors $\boldsymbol{v} \neq \boldsymbol{0}$ and real c, then the KKT conditions of the Lasso $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ hold strictly with probability* 1, *that is*, (3.35) *holds with probability* 1.

Expositions of the results in the first part of the above proposition exist in the literature; see, for instance, [59], Section 3, or [50, 52]. Compared with previous versions of the condition on the design, Assumption 3.1, which clearly holds with probability 1 when $\boldsymbol{X}$ is the realization of a continuous distribution over $\mathbb{R}^{n\times p}$, gives a natural interpretation in terms of the rank of specific matrices. The fact that the KKT conditions of the Lasso hold strictly with probability one is known although it is difficult to pinpoint an existing result in the form of Proposition 3.9 in the literature. We provide a short proof in Appendix E for completeness.

Next, define the function $f : \mathbb{R}^n \to \mathbb{R}^n$ by

$$f : \boldsymbol{\varepsilon} \to \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta}). \tag{3.36}$$

Then the function f is 1-Lipschitz and this property holds true for all convex penalized least-squares estimators [3], Proposition 3. Consequently, almost everywhere the partial derivatives of f exist and $\nabla f(\boldsymbol{\varepsilon})$ has operator norm at most one. It is enough to compute the gradient of f Lebesgue almost everywhere and by the above proposition, the KKT conditions holds strictly for almost every point $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^n$.

If the KKT conditions of the Lasso hold strictly for $\boldsymbol{\varepsilon}_0$, then by Lipschitz continuity of $\boldsymbol{\varepsilon} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ the KKT conditions also hold strictly in small enough nontrivial neighbourhood of $\boldsymbol{\varepsilon}_0$. In this small neighborhood, the sign and support of $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ are unchanged and we have for $\|\boldsymbol{h}\|$ small enough

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}(\boldsymbol{\varepsilon}_0 + \boldsymbol{h}) = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}(\boldsymbol{X}_{\widehat{S}}^\top(\boldsymbol{\varepsilon}_0 + \boldsymbol{h} + \boldsymbol{X}\boldsymbol{\beta}) - \lambda n \,\text{sign}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}(\boldsymbol{\varepsilon}_0))),$$

where $\widehat{S}$ denotes the locally constant support equal to the support of $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}(\boldsymbol{\varepsilon}_0)$. In this neighborhood, the map $\boldsymbol{h} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}(\boldsymbol{\varepsilon}_0 + \boldsymbol{h})$ as well as the map $\boldsymbol{h} \to \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}(\boldsymbol{\varepsilon}_0 + \boldsymbol{h}) - \boldsymbol{\beta})$ are locally affine with linear part equal to the orthogonal projection

$$\boldsymbol{P}_{\widehat{S}} = \boldsymbol{X}_{\widehat{S}}(\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top. \tag{3.37}$$

We conclude this calculation with the following lemma.

PROPOSITION 3.10. *Let $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ be the Lasso estimator* (3.34) *with data $(\boldsymbol{X}, \boldsymbol{y})$ satisfying $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Define $f(\boldsymbol{\varepsilon}) = \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})$ as in* (3.36). *Suppose Assumption* 3.1 *holds and $\mathbb{P}\{\boldsymbol{v}^\top \boldsymbol{\varepsilon} = c\} = 0$ for all deterministic $\boldsymbol{v} \in \mathbb{R}^n$ and real c. Then almost surely*

$$\nabla \hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} = \begin{pmatrix} (\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}_{\widehat{S}}^\top \\ \boldsymbol{0}_{\widehat{S}^c \times n} \end{pmatrix}_{p\times n}$$

as well as

$$\nabla f(\boldsymbol{\varepsilon}) = \boldsymbol{P}_{\widehat{S}} \quad \textit{and} \quad \operatorname{div} f(\boldsymbol{\varepsilon}) = \|\boldsymbol{P}_{\widehat{S}}\|_F^2 = \operatorname{trace} \boldsymbol{P}_{\widehat{S}} = |\widehat{S}|,$$

where $\widehat{S} = \operatorname{supp}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}})$ *and* $\boldsymbol{P}_{\widehat{S}}$ *is as in* (3.37).

3.4.1. *Variance formula and relative stability.* Using Theorem 2.2, we obtain the following result whose proof is given in Appendix F.

THEOREM 3.11. *Consider the linear model* (3.33) *and* $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}$ *in* (3.34), *with deterministic design* $\boldsymbol{X}$ *satisfying Assumption* 3.1, *true target vector* $\boldsymbol{\beta}$ *and noise* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. *Then the variance of the size of the selected support satisfies*

$$\begin{aligned} \operatorname{Var}[|\widehat{S}|] &= \mathbb{E}[|\widehat{S}|] + \mathbb{E}[\|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\|^2/\sigma^2] + \operatorname{Var}(g(\boldsymbol{\varepsilon})) - \sigma^2 \mathbb{E}[\|\nabla g(\boldsymbol{\varepsilon})\|^2] \\ &\le \mathbb{E}[|\widehat{S}|] + \mathbb{E}[\|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\|^2/\sigma^2], \end{aligned} \tag{3.38}$$

where $g(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}} - \boldsymbol{\beta})/\sigma^2$. *Consequently,* $\operatorname{Var}[|\widehat{S}|] \le 2n$ *as well as*

$$\begin{aligned} \operatorname{Var}[|\widehat{S}|] &\le 3\mathbb{E}[|\widehat{S}|] + 4\mathbb{E}\left[|\widehat{S}| \log\left(\frac{ep}{1 \vee |\widehat{S}|}\right)\right] \\ &\le 3\mathbb{E}[|\widehat{S}|] + 4\mathbb{E}[|\widehat{S}|] \log\left(\frac{ep}{1 \vee \mathbb{E}[|\widehat{S}|]}\right). \end{aligned} \tag{3.39}$$

A significant feature of the above theorem, and Theorem 2.2 as well, is the requirement of no condition on the true $\boldsymbol{\beta}$, the penalty level λ or the design matrix $\boldsymbol{X}$ beyond Assumption 3.1. In particular, the restricted eigenvalue condition is not required.

Using $a/b + b/(a \vee 1) - 2 \le (a-b)^2/(b(a \vee 1))$, an implication of Theorem 3.11 is the confidence interval

$$\mathbb{P}\left(\frac{|\widehat{S}|}{\mathbb{E}[|\widehat{S}|]} + \frac{\mathbb{E}[|\widehat{S}|]}{|\widehat{S}| \vee 1} - 2 \le C_\alpha\left(\frac{3}{|\widehat{S}| \vee 1} + \frac{4\log(ep)}{|\widehat{S}| \vee 1}\right)\right) \ge 1 - \alpha \tag{3.40}$$

for $\mathbb{E}[|\widehat{S}|]$ with conservative $C_\alpha = 1/\alpha$, although $\mathbb{E}[|\widehat{S}|]$ is not a conventional parameter due to its dependence on the specific choice of $\hat{\boldsymbol{\beta}}$.

A sequence of nonnegative random variables $(Z_q)_{q \ge 1}$ is said to be relatively stable if $Z_q/\mathbb{E}[Z_q]$ converges to 1 in probability. A direct consequence of Theorem 3.11 is that the model size $|\widehat{S}|$ is relatively stable provided that $\mathbb{E}[|\widehat{S}|]$ is not pathologically small. If the setting and assumptions of Theorem 3.11 are fulfilled, then

$$\mathbb{E}\left[\left(\frac{|\widehat{S}|}{\mathbb{E}|\widehat{S}|} - 1\right)^2\right] \le \frac{3}{\mathbb{E}|\widehat{S}|} + \frac{4\log(ep/\mathbb{E}|\widehat{S}|)}{\mathbb{E}|\widehat{S}|} \le \frac{3 + 4\log(ep)}{\mathbb{E}|\widehat{S}|}.$$

Consequently, if one considers a sequence of regression problems such that $\mathbb{E}[|\widehat{S}|]/\log(ep) \to +\infty$, then $|\widehat{S}|/\mathbb{E}|\widehat{S}|$ converges to 1 in L_2 and in probability.

While the stability of $|\widehat{S}|$ supports the use of the Lasso, it is guaranteed by Theorem 3.11 only when $\mathbb{E}[|\widehat{S}|]$ is of greater order than $\log p$. In practice, this means observing a sufficiently large $|\widehat{S}|$ in view of (3.40). In general large $\mathbb{E}[|\widehat{S}|]$ happens when the penalty level is low or the number of significant coefficients is large. It is known that $\mathbb{P}\{|\widehat{S}| \ge k\} \approx 1$ when $\boldsymbol{X}$ has i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$ rows with $(\boldsymbol{\Sigma})_{jj} = 1$, $\boldsymbol{\beta} = \mathbf{0}$ and $\lambda = \lambda(k) = (\sigma/n^{1/2})\{L(k/p) - 1\}$ under a mild side condition on $\boldsymbol{\Sigma}$, where $L(t)$ is defined by $\mathbb{P}\{N(0,1) > L(t)\} = t$ [48], Proposition 14(ii). As $|\widehat{S}|$ is expected to be larger when $\boldsymbol{\beta} \ne \mathbf{0}$, it would be reasonable to expect $\mathbb{E}[|\widehat{S}|] \gg \log p$ when $\lambda = \lambda(k_n)$ with $k_n \gg \log p$. Additionally, since the Lasso is known to satisfy estimation error bounds of the form $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \le C_0^2 \|\boldsymbol{\beta}\|_0 \lambda^2$ with high probability for some constant $C_0 > 0$, on this event it must be that $|\hat{S}| \ge (1 - b^{-2} - \pi_0)\|\boldsymbol{\beta}\|_0$ under the signal strength condition $\#\{j : |\beta_j| > bC_0\lambda\} \ge (1 - \pi_0)\|\boldsymbol{\beta}\|_0$. This implies that $|\hat{S}| \ggg \log p$ holds with high probability when $\|\boldsymbol{\beta}\|_0 \ggg \log p$ and $b^{-2} + \pi_0 < 1$.

An informative benchmark to study the tightness of inequality (3.39) for the support of the Lasso is the case $\boldsymbol{X} = \sqrt{n}\boldsymbol{I}_p$ which reduces to the Gaussian sequence model. Then $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ is the soft-thresholding operator and $|\widehat{S}|$ is the sum of p i.i.d. Bernoulli random variables with parameters $q_1, \dots, q_p \in (0, 1)$ and $\operatorname{Var}[|\widehat{S}|] = \sum_{j=1}^p q_j(1 - q_j)$. Under mild assumption on the probabilities q_j (e.g., $q_j \le 1/2$ for all j), the variance $\operatorname{Var}[|\widehat{S}|]$ is of the same order as $\mathbb{E}|\widehat{S}|$. Hence the bound (3.39) is sharp up to a logarithmic factor.

3.4.2. *Linearity of the variance in the true model size.* From (3.39), we can also obtain more explicit bounds on the variance of $|\widehat{S}|$ by bounding from above $\mathbb{E}[|\widehat{S}|]$. We provide below upper bounds on $\mathbb{E}[|\widehat{S}|]$ under two assumptions on $\boldsymbol{X}$: the Sparse Riesz Condition (SRC) [59, 61] and the Restricted Eigenvalue (RE) condition [12]. Under both conditions, if the tuning parameter of the Lasso is large enough then the squared risk $\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2/n$ is bounded from above by $C(\boldsymbol{X})s_0\lambda^2$ with high probability [12, 59, 61] and in expectation [3, 6, 58], where $C(\boldsymbol{X})$ is a multiplicative constant that depends on $\boldsymbol{X}$. We refer the reader to the books [16, 27, 28] and the references therein for surveys of existing results. Throughout the rest of this section, denote by $s_0 = \|\boldsymbol{\beta}\|_0$ the number of nonzero coefficients, or sparsity, of the unknown coefficient vector $\boldsymbol{\beta}$.

The Sparse Riesz Condition (SRC) [59, 61] on the design $\boldsymbol{X}$ holds if for certain $\eta \in (0, 1)$ and nonnegative reals $\{\epsilon_1, \epsilon_2\}$ and integers $\{m, k\}$,

$$|S| + \epsilon_2 k \le \min_{B \subset [p]: |B \setminus S| \le m} \frac{2(1-\eta)^2 m - \epsilon_1 k}{\phi_{\text{cond}}(\boldsymbol{X}_B^\top \boldsymbol{X}_B) - 1}, \tag{3.41}$$

where S denotes the support of $\boldsymbol{\beta}$ and $\phi_{\text{cond}}(\cdot)$ denotes the condition number. Let $\phi(d) \ge \max\{\phi_{\text{cond}}(\boldsymbol{X}_B^\top \boldsymbol{X}_B) : |B| = d\}$ be an upper bound for the d-sparse condition number of the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$. Given d and $\{\eta, \epsilon_1, \epsilon_2\}$, the SRC can be viewed as a sparsity condition on $\boldsymbol{\beta}$ as it holds with $k = \|\boldsymbol{\beta}\|_0$ when $\|\boldsymbol{\beta}\|_0 \le 2d(1-\eta)^2/\{(1+\epsilon_2)(\phi(d) - 1) + 2(1 - \eta)^2 + \epsilon_1\}$. In particular, the SRC holds under the RIP condition $\delta_{2s_0} < 1/2$ or $\delta_{3s_0} < 2/3$ for sufficiently small $\{\eta, \epsilon_1, \epsilon_2\}$ where δ_k is the maximum spectrum norm of $\boldsymbol{X}_B^\top \boldsymbol{X}_B/n - \boldsymbol{I}_B$ over $|B| \le k$. When $\boldsymbol{X}$ has i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows, we may take $d = a_1 n/\log p$ such that $\phi(d) = (1 + a_2)\phi_{\text{cond}}(\boldsymbol{\Sigma})$ is a valid upper bound for the d-sparse condition number of $\boldsymbol{X}^\top \boldsymbol{X}/n$ with probability $1 - e^{-a_2 n}$ for some small positive constants a_1 and a_2 [61].

Note that the original SRC in [61] is stated in terms of ratio of maximal and minimal sparse eigenvalues instead of sparse condition number as in (3.41). A common feature on the works on the SRC [59, 61] is that $|\widehat{S}| \lesssim s_0$ with large probability (up to constants depending on the SRC constants and the tuning parameter). We obtain $\mathbb{E}|\widehat{S}| \lesssim s_0$ as a consequence, provided that $\mathbb{P}(|\widehat{S}| \lesssim s_0)$ is large enough.

PROPOSITION 3.12. *Let $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and $s_0 = \|\boldsymbol{\beta}\|_0$ in the linear model* (3.33). *Let the tuning parameter of the Lasso* (3.34) *satisfy*

$$\lambda \ge (\sigma/\eta)\sqrt{(1+\tau)(2/n)\log(p/k)\}} \tag{3.42}$$

with $\eta \in (0, 1)$. Assume that $\boldsymbol{X}$ is deterministic with $\max_{j \in [p]} \|\boldsymbol{X}\boldsymbol{e}_j\|^2/n \le 1$ and that (3.41) *holds with $\epsilon_1 = 8(1-\eta)/\{\sqrt{2\pi}t_0\}$, $\epsilon_2 = 4/\{t_0^2(2\pi t_0^2 + 4)^{1/2}\}$ and some nonnegative integers $\{m, k\}$, where $t_0 = \eta^{-1}\sqrt{2\log(p/k)}$. Then, $\mathbb{P}(|\widehat{S} \setminus S| < m) \ge 1 - (k/p)^\tau/(1 + (\eta t_0)^2)$ and $\mathbb{E}[|\widehat{S}|] \le s_0 + m + p(k/p)^\tau/(1 + (\eta t_0)^2)$. Consequently, if $\tau \ge 1$ then $\mathbb{E}|\widehat{S}| \le s_0 + m + k/(1 + (\eta t_0)^2)$.*

The proof is given in Appendix H. In the above result, $\tau > 0$, $\eta < 1$ are thought of as absolute constants while k, m as sparsity levels typically proportional to the true sparsity s_0 of $\boldsymbol{\beta}$. For instance, the required lower bound on the tuning parameter in (3.42) reads $\lambda \geq 1.01\sigma\sqrt{4\log(p/s_0)/n}$ for $\eta = 1/1.01$, $\tau = 1$ and $k = s_0$. In particular if $\tau \geq 1$, $s_0 \approx m$ and $k \lesssim s_0$ then the SRC grants $\mathbb{E}|\widehat{S}| \lesssim s_0$. We refer to [8, 59, 61] for more detailed discussions on the SRC.

The expectation $\mathbb{E}[|\widehat{S}|]$ can also be bounded from above if both Assumption 3.1 and the Restricted Eigenvalue (RE) condition [12] hold.

THEOREM 3.13. *Consider the linear model* (3.33) *with* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ *and* $s_0 = \|\boldsymbol{\beta}\|_0$. *Let* $\tau, \gamma > 0$, $\omega = \sigma(1+\tau)/\sqrt{n}$ *and* $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ *be as in* (3.34) *with*

$$\lambda = \sigma(1+\tau)(1+\gamma)\sqrt{(2/n)\log\big(ep/(s_0 \vee 1)\big)}. \tag{3.43}$$

Assume that the columns of $\boldsymbol{X}$ *are normalized such that* $\max_{j\in[p]} \|\boldsymbol{X}\boldsymbol{e}_j\|^2 \leq n$. *Let* $\widehat{S}$ *be the support of* $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$. *Then the Lasso satisfies*

$$\mathbb{E}\big[2\tau\boldsymbol{\varepsilon}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta}) + \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2\big]/n \leq \frac{s_0(\lambda^2+\omega^2) + (s_0 \vee 1)\omega^2}{\text{RE}^2(S, c_0)} + \frac{\omega^2}{2},$$

where $\text{RE}(S, c_0) = \inf_{\boldsymbol{u}\in\mathbb{R}^p: \|\boldsymbol{u}_{S^c}\|_1 \leq c_0\sqrt{s_0 \vee 1}\|\boldsymbol{u}\|}(n^{-1/2}\|\boldsymbol{X}\boldsymbol{u}\|/\|\boldsymbol{u}\|)$ *and* S *is the support of* $\boldsymbol{\beta}$, *provided that* $c_0 \geq \gamma^{-1}\sqrt{2(1+2\omega^2/\lambda^2)}$, *for example*, $c_0 = 2/\gamma$. *If in addition Assumption* 3.1 *holds*, *then*

$$\begin{aligned}
&\mathbb{E}\big[|\widehat{S}| + \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2/(2\sigma^2\tau)\big] \\
&\quad \leq (\sqrt{\tau} + 1/\sqrt{\tau})^2 \bigg[\frac{(1+\gamma)^2\{s_0\log(ep/(s_0 \vee 1)) + (s_0 \vee 1)\}}{\text{RE}^2(S, 2/\gamma)} + \frac{1}{4}\bigg].
\end{aligned} \tag{3.44}$$

The proof of Theorem 3.13 is given in Appendix I. Here, $\tau, \gamma, c_0 > 0 > 0$ are thought of as absolute constants, and $\text{RE}(S, c_0) > 0$ as of constant order. For instance, the tuning parameter (3.43) reads $\lambda = 1.01\sigma\sqrt{2\log(ep/s_0)/n}$ for $\tau = \gamma = \sqrt{1.01} - 1$. The scaling $\sigma\sqrt{\log(p/s_0)}$ for the tuning parameter visible in (3.42) and (3.43) allows for smaller tuning parameters than the universal parameter $\sigma\sqrt{2\log p}$ studied in earlier works on the Lasso (e.g., [12]). Tuning parameters of order $\sigma\sqrt{\log(p/s_0)}$ have been previously studied in [4, 6, 8, 35, 47].

The Gaussian concentration theorem is used in [3, 6] to obtain bounds on $\mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2]$ as well as higher order moments of the squared risk; similar arguments are used to derive Theorem 3.13. If $\boldsymbol{X}$ satisfies Assumption 3.1 then $\mathbb{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})] = \mathbb{E}[|\widehat{S}|]$, so that the argument leads to (3.44). Informally, this implies $\mathbb{E}[|\widehat{S}|] \lesssim 1 + s_0\log(ep/(s_0 \vee 1))$ up to a multiplicative constant that depends only on γ, τ and the restricted eigenvalue. To our knowledge, this bound on the size of the model selected by the Lasso under the RE condition is new. Previous upper bounds of the form $|\widehat{S}| \lesssim s_0$ require that both maximal and minimal sparse eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}/n$ are bounded away from 0 and $+\infty$; cf. Proposition 3.12 above or [61], [59], Lemma 1, [12], (7.9), [11], Theorem 3, among others. The major difference between such conditions and the RE condition is that the RE condition does not require any bounds on the maximal sparse eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}/n$. Inequality (3.44) reveals that the RE condition is sufficient to control $\mathbb{E}[|\widehat{S}|]$ by s_0 times a logarithmic factor. Under the RE condition, assumptions on the maximal sparse eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}/n$ are unnecessary to control $\mathbb{E}[|\widehat{S}|]$.

The above bounds on $\mathbb{E}|\widehat{S}|$ under the SRC or the RE condition yield the following on the variance of $|\widehat{S}|$ in virtue of (3.39). If Assumption 3.1 holds:

(i) If λ is as in (3.43) for some $\gamma, \tau > 0$ and $\max_{j=1,\dots,p} \|\boldsymbol{X}\boldsymbol{e}_j\|^2 \le n$ then

$$\mathrm{Var}[|\widehat{S}|] \le \frac{(s_0 \vee 1)C}{\mathrm{RE}^2(S, 2/\gamma)} \log\left(\frac{ep}{1 \vee s_0}\right)^2,$$

where $C = C(\gamma, \tau) > 0$ only depends on γ, τ.

(ii) If $\boldsymbol{X}$ satisfies the SRC (3.41) for some $\eta > 0$ and $0 \le m, k \le p$, with λ is as in (3.42) with $\tau \ge 1$ then for some absolute constant $C' > 0$,

$$\mathrm{Var}[|\widehat{S}|] \le C'\left((s_0 + m)\log\left(\frac{ep}{s_0 + m}\right)\right).$$

In other words, under the RE condition or the SRC with $m \approx s_0$ the standard deviation of the size of the model $|\widehat{S}|$ is smaller than $\sqrt{s_0}$ up to logarithmic factors. The bound is sharper under the SRC by a logarithmic factor.

3.4.3. *Variance of degrees-of-freedom of penalized estimators.* Some techniques above are not specific to the Lasso. For instance, for any estimator defined as the solution of a convex optimization problem of the form given in Remark 3.3, the map $f : \boldsymbol{\varepsilon} \to \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is 1-Lipschitz and satisfies

$$\begin{aligned}\mathbb{E}\big[(\sigma^2\hat{\mathsf{df}} - \boldsymbol{\varepsilon}^\top f(\boldsymbol{\varepsilon}))^2\big] &\le \sigma^2\mathbb{E}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + \sigma^4\mathbb{E}\|\nabla f(\boldsymbol{\varepsilon})\|_F^2 \\ &\le \sigma^2\mathbb{E}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + \sigma^4 n\end{aligned} \tag{3.45}$$

by Theorem 2.1 where $\hat{\mathsf{df}} = \operatorname{div} f(\boldsymbol{\varepsilon})$. Similarly, by Theorem 2.2 we have

$$\begin{aligned}\mathrm{Var}[\hat{\mathsf{df}}] &\le \mathbb{E}\|\nabla f(\boldsymbol{\varepsilon})\|_F^2 + \mathbb{E}\|\nabla f(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}\|^2/\sigma^2 \\ &\le 2n.\end{aligned} \tag{3.46}$$

3.5. *SURE for SURE in high-dimensional linear regression.* Again we consider linear regression with deterministic design $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. With the notation of Section 3.1, consider the sequence model $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ and the unknown mean is $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, as in the linear model (3.33).

3.5.1. *Lasso.* Set $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}$ with the Lasso estimator (3.34). We have derived in the previous section the gradient of $\boldsymbol{\varepsilon} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}$ almost everywhere under Assumption 3.1. It is instructive to use these calculations to make explicit SURE for SURE from Section 3.1 in the Lasso case. Under Assumption 3.1, $|\widehat{S}| = \operatorname{div}\widehat{\boldsymbol{\mu}} = \operatorname{trace}((\nabla\widehat{\boldsymbol{\mu}})^2)$ by Proposition 3.10, so that Stein's unbiased risk estimate is

$$\widehat{\mathsf{SURE}} = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}\|^2 + \sigma^2(2|\widehat{S}| - n) \tag{3.47}$$

as in (3.7). Moreover, by Theorem 3.1, SURE for SURE in the Lasso case is

$$\widehat{R}_{\mathrm{SURE}} = 4\sigma^2\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}\|^2 + \sigma^4(4|\widehat{S}| - 2n) \tag{3.48}$$

which is an unbiased estimator of $R_{\mathrm{SURE}} = \mathbb{E}[(\widehat{\mathsf{SURE}} - \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}} - \boldsymbol{\beta})\|^2)^2]$. The identity $\mathbb{E}[\widehat{R}_{\mathrm{SURE}}] = R_{\mathrm{SURE}}$ for the Lasso appeared previously in [24].

As $\nabla(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\mathrm{LASSO}}) = \boldsymbol{P}_{\widehat{S}}$ is a random projection, Theorem 3.2 applies with $\widehat{R}'_{\mathrm{SURE}} = \widehat{R}_{\mathrm{SURE}}$, so that SURE for SURE is consistent. We explicitly state the consequences of this result in the following proposition.

PROPOSITION 3.14. *Consider the sequence model* $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ *and let* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *satisfy Assumption* 3.1. *Let* $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ *be the Lasso* (3.34). *Consider SURE in* (3.47) *and SURE for SURE in* (3.48). *Then for any tuning parameter* $\lambda \ge 0$, *the following holds*:

(i) $\mathrm{Var}(\widehat{\mathsf{SURE}}) = \mathbb{E}[(\widehat{\mathsf{SURE}} - \mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}\|^2])^2] \le \mathbb{E}[\hat{R}_{\text{SURE}}] + \sigma^4 n$.

(ii) $\mathbb{E}[\hat{R}_{\text{SURE}}] = 2\sigma^2 \mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}\|^2 + \widehat{\mathsf{SURE}}] \ge n\sigma^4$.

(iii) *The self-bounding property* $\mathrm{Var}[\hat{R}_{\text{SURE}}] \le 16\sigma^4 \mathbb{E}[\hat{R}_{\text{SURE}}]$ *holds*.

(iv) *Inequality* $\mathbb{E}[|\hat{R}_{\text{SURE}}/\mathbb{E}[\hat{R}_{\text{SURE}}] - 1|^2] \le 16/n$ *holds so that the ratio* $\hat{R}_{\text{SURE}}/\mathbb{E}[\hat{R}_{\text{SURE}}]$ *converges to* 1 *in* L_2 *and in probability as* $n, p \to \infty$.

(v) *For any* $\alpha, \gamma \in (0, 1)$, *with probability at least* $1 - \alpha - \gamma$,

$$\bigl|\widehat{\mathsf{SURE}} - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\mu}\|^2\bigr|^2 \le \gamma^{-1}\hat{R}_{\text{SURE}}\bigl(1 - 4(n\alpha)^{-1/2}\bigr)^{-1}.$$

The proof is given in Appendix J. In the above, (i) provides in terms of SURE for SURE an upper bound for the mean squared error of SURE when the prediction risk is viewed as the estimation target of SURE, and (iv) is a consistency result for SURE for SURE in the Lasso case. The nonasymptotic Proposition 3.14 holds with no restriction on (n, p) and the tuning parameter λ. The assumption-free nature of Proposition 3.14 is striking: SURE for SURE is consistent even if the Lasso itself is not consistent for prediction in the sense that $\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\mu}\|^2/(n\sigma^2)]$ is bounded away from 0.

To bound the variance of $\hat{R}_{\text{SURE}}$, Proposition 3.14(iv) leverages the fact that for $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$, $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ is a random orthogonal projection and $\hat{R}_{\text{SURE}} = \hat{R}'_{\text{SURE}}$, cf. the discussion following (3.9). While $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ for $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ is not a projection for other estimators $\hat{\boldsymbol{\beta}}$ such as the Elastic-Net studied in Section 3.5.3, the upper bounds in Theorem 3.2 still apply to $\hat{R}'_{\text{SURE}}$ for any convex regularized least-squares as explained in Remark 3.3.

A drawback of the confidence region in (v) is the conservative constant factor γ^{-1}. This can be fixed under common regularity assumptions made in sparse linear regression as follows with the approach of Section 3.2. Let $\{\tau, \gamma, \lambda\}$ be as in Theorem 3.13 and define $C_{\tau,\gamma} = \max(1, 2\tau)(\sqrt{\tau} + 1/\sqrt{\tau})^2\{4(1+\gamma)^2 + 5\}$. Under the conditions of Theorem 3.13 including Assumption 3.1, (3.44) implies that the right-hand side of (3.18) is bounded by $\sigma^4 2n\epsilon_n^*$ with $\epsilon_n^* = C_{\tau,\gamma}(s_0 \vee 1)\{\log(p/(s_0 \vee 1))\}/\{n\,\mathrm{RE}^2(S, 2/\gamma)\}$, so that

$$\mathbb{P}\bigl\{\bigl|\widehat{\mathsf{SURE}} - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2\bigr| \le 1.96\sigma^2\sqrt{2n}\bigr\} \approx 95\%$$

by Theorem 3.4 when $\epsilon_n^* = o(1)$, with $v_0^2 = \epsilon_n = \sqrt{\epsilon_n^*}$, and similarly

$$\mathbb{P}\bigl\{\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 \le \widehat{\mathsf{SURE}} + 1.645\sigma^2\sqrt{2n}\bigr\} \approx 95\%.$$

3.5.2. *Two or more Lasso estimators*. For the comparison of two Lasso estimators $\hat{\boldsymbol{\beta}}^{(\lambda_1)}_{\text{LASSO}}$ and $\hat{\boldsymbol{\beta}}^{(\lambda_2)}_{\text{LASSO}}$ with $\lambda_1 \ne \lambda_2$,

$$\widehat{\mathsf{SURE}}^{(\text{diff})} = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_1)}_{\text{LASSO}} - \boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_2)}_{\text{LASSO}} - \boldsymbol{y}\|^2 + 2\sigma^2\bigl(|\widehat{S}^{(\lambda_1)}| - |\widehat{S}^{(\lambda_2)}|\bigr) \tag{3.49}$$

provides $\mathbb{E}[\widehat{\mathsf{SURE}}^{(\text{diff})}] = \mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda_1)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2 - \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda_2)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2]$, where $\widehat{S}^{(\lambda_j)} = \mathrm{supp}(\hat{\boldsymbol{\beta}}^{(\lambda_j)}_{\text{LASSO}})$. If $\boldsymbol{P}_A$ is the projection onto the column space of $\boldsymbol{X}_A$,

$$\widehat{R}^{(\text{diff})}_{\text{SURE}} = 4\sigma^2\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda_1)}_{\text{LASSO}} - \hat{\boldsymbol{\beta}}^{(\lambda_2)}_{\text{LASSO}})\|^2 + 4\sigma^4\,\mathrm{trace}\bigl((\boldsymbol{P}_{\widehat{S}^{(\lambda_1)}} - \boldsymbol{P}_{\widehat{S}^{(\lambda_2)}})^2\bigr) \tag{3.50}$$

provides

$$\mathbb{E}\widehat{R}^{(\text{diff})}_{\text{SURE}} = R^{(\text{diff})}_{\text{SURE}} = \mathbb{E}\bigl[\bigl(\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda_1)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2 - \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda_2)}_{\text{LASSO}} - \boldsymbol{\beta})\|^2 - \widehat{\mathsf{SURE}}^{(\text{diff})}\bigr)^2\bigr].$$

The results of Section 3.3 are also directly applicable to SURE-tuned Lasso estimators: If $\hat{\lambda}$ is the tuning parameter among $\{\lambda_1, \dots, \lambda_m\}$ with the smallest SURE, then for some absolute constant $C > 0$,

$$\mathbb{E}\Big[n^{-1/2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\hat{\lambda})}_{\text{LASSO}} - \boldsymbol{\beta})\| - \min_{\lambda\in\{\lambda_1,\dots,\lambda_m\}} n^{-1/2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta})\|\Big] \le C(m/n)^{1/4}.$$

3.5.3. *Elastic net*. Similar computations can be carried out for other estimators such as the Group Lasso or the Elastic Net. For instance, consider the Elastic Net estimator $\hat{\boldsymbol{\beta}}_{\text{EN}}$ defined as the solution of the optimization problem

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \arg\max_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}\|^2/2n + \lambda\|\boldsymbol{b}\|_1 + \gamma\|\boldsymbol{b}\|^2/2, \tag{3.51}$$

where $\lambda, \gamma > 0$. Set $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{EN}}$. Then by similar arguments as in the Lasso case, the KKT conditions of the optimization problem (3.51) hold strictly almost everywhere in $\boldsymbol{y}$. By differentiating the KKT conditions on a neighborhood where the KKT conditions hold strictly (the details are omitted), the gradient of $\boldsymbol{y} \to \hat{\boldsymbol{\beta}}_{\text{EN}}$ is given by

$$\nabla\hat{\boldsymbol{\beta}}_{\text{EN}} = \begin{pmatrix} (\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}} \\ \boldsymbol{0}_{\widehat{S}^c\times n} \end{pmatrix}_{p\times n}, \tag{3.52}$$

and the gradient of $\boldsymbol{y} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{EN}}$ is given by

$$\nabla(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{EN}}) = \boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}, \tag{3.53}$$

where $\widehat{S} \subset [p]$ is the set of nonzero coefficients of $\hat{\boldsymbol{\beta}}_{\text{EN}}$. Stein's unbiased risk estimate is given by

$$\widehat{\mathsf{SURE}} = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{EN}}\|^2 + 2\sigma^2\operatorname{trace}[\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}] - \sigma^2 n, \tag{3.54}$$

and SURE for SURE in the Elastic-Net case is

$$\widehat{R}_{\text{SURE}} = 4\sigma^2\|\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{EN}} - \boldsymbol{y}\|^2 + 4\sigma^4\|\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}\|_F^2 - 2\sigma^4 n. \tag{3.55}$$

By Theorem 3.1, this is an unbiased estimate of $\mathbb{E}[(\widehat{\mathsf{SURE}} - \|\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{EN}})\|^2)^2]$. SURE for SURE $\widehat{R}^{(\text{diff})}_{\text{SURE}}$ for the difference between two E-nets or between the Lasso and E-net can be derived similarly as in (3.50). We omit the details.

REMARK 3.4. Let $\hat{\mathsf{df}} = \operatorname{trace}[\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}]$. Since $\hat{\mathsf{df}}$ is the divergence of the function $\boldsymbol{\varepsilon} \to \boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\text{EN}} - \boldsymbol{\beta})$, Theorem 2.2 implies that

$$\begin{aligned}\operatorname{Var}[\hat{\mathsf{df}}] \le{}& \mathbb{E}[\|\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}\|_F^2] \\ &+ \mathbb{E}[\|\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{\varepsilon}\|^2]/\sigma^2.\end{aligned}$$

If $\boldsymbol{P}_{\widehat{S}}$ is the orthogonal projection onto the span of the columns of $\boldsymbol{X}_{\widehat{S}}$ then the second term satisfies $\mathbb{E}[\|\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{\varepsilon}\|^2]/\sigma^2 \le \mathbb{E}[\|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\|^2]/\sigma^2$. Since the right-hand side of (3.38) is no greater than that of (3.39) for any random $\widehat{S}$ by the proof of Theorem 3.11, we obtain

$$\begin{aligned}\operatorname{Var}[\hat{\mathsf{df}}] \le{}& \mathbb{E}[\|\boldsymbol{X}_{\widehat{S}}(\gamma\boldsymbol{I}_{\widehat{S}} + \boldsymbol{X}^\top_{\widehat{S}}\boldsymbol{X}_{\widehat{S}})^{-1}\boldsymbol{X}^\top_{\widehat{S}}\|_F^2] \\ &+ \mathbb{E}[2|\widehat{S}| + 4|\widehat{S}|\log(ep/\{1\vee|\widehat{S}|\})] \\ \le{}& 3\mathbb{E}[|\widehat{S}|] + 4\mathbb{E}[|\widehat{S}|\log(ep/\{1\vee|\widehat{S}|\})].\end{aligned}$$

3.6. *Debiasing nonlinear estimators in linear regression.* Consider a linear regression model

$$y = X\beta + \varepsilon \tag{3.56}$$

with an unknown target vector $\beta \in \mathbb{R}^p$, a Gaussian noise vector $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, and a Gaussian design matrix $X \in \mathbb{R}^{n\times p}$ with i.i.d. $N(\mathbf{0}, \Sigma)$ rows. We assume that the covariance matrix Σ is known and invertible.

This section explains how to construct an estimate of a linear contrast

$$\theta = \langle a_0, \beta\rangle \tag{3.57}$$

from an initial estimator $\hat{\beta}$. Here and in the sequel, $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbb{R}^n$. Define

$$u_0 = \Sigma^{-1}a_0/\langle a_0, \Sigma^{-1}a_0\rangle, \qquad z_0 = X u_0, \qquad Q_0 = I_{p\times p} - u_0 a_0^\top \tag{3.58}$$

and assume for simplicity that a_0 is normalized such that

$$\langle a_0, \Sigma^{-1}a_0\rangle = 1.$$

By definition of u_0, $z_0 \sim N(\mathbf{0}, I_n)$ and z_0 is independent of XQ_0.

We assume throughout this section that we are given an initial estimator $\hat{\beta}$. Since $X = z_0 a_0^\top + XQ_0$ and the two random vectors z_0, XQ_0 are independent, we view $\hat{\beta}$ as a function with three arguments $\hat{\beta} = \hat{\beta}(y, z_0, XQ_0)$ and we assume that the partial derivatives $(\partial/\partial y)\hat{\beta}$ and $(\partial/\partial z_0)\hat{\beta}$ exist almost everywhere.

The estimator $\hat{\beta}$ provides an initial estimate of the unknown parameter θ (3.57) by the plug-in $\langle a_0, \hat{\beta}\rangle$. However, this estimator may be biased, and a first attempt to fix the bias is the following one-step MLE correction in the direction given by the one-dimensional model $\{\hat{\beta} + t u_0, t \in \mathbb{R}\}$,

$$\langle a_0, \hat{\beta}\rangle + \frac{\langle z_0, y - X\hat{\beta}\rangle}{\|z_0\|^2}. \tag{3.59}$$

Variants of the above debiasing scheme have been considered in [10, 15, 30, 31, 54, 60, 62], among others. We multiply by $\|z_0\|^2$ to avoid random denominators; the random variables $\|z_0\|^2$ is chi-square with n degrees of freedom, equal to $n + O(\sqrt{n})$ with overwhelming probability so that $\|z_0\|^2 \approx n$ describes the number of observations.

When constructing the estimator (3.59) above by the one-step MLE correction, the statistician hopes that the quantity

$$\|z_0\|^2\langle a_0, \hat{\beta} - \beta\rangle + \langle z_0, y - X\hat{\beta}\rangle \tag{3.60}$$

is asymptotically standard normal; this is the ideal result to construct confidence intervals for the unknown parameter (3.57) at the $\sqrt{n}$-adjusted rate.

By simple algebra, we have

$$\|z_0\|^2\langle a_0, \hat{\beta} - \beta\rangle + \langle z_0, y - X\hat{\beta}\rangle = z_0^\top \varepsilon - z_0^\top XQ_0(\hat{\beta} - \beta). \tag{3.61}$$

The random variable $z_0^\top\varepsilon$ in the right-hand side is mean-zero and $z_0^\top\varepsilon/(\sigma/\sqrt{n})$ is asymptotically standard normal. It remains to understand the bias term $z_0^\top XQ_0(\hat{\beta} - \beta)$. For the derivation below, we will argue conditionally on (ε, XQ_0) and define $f : \mathbb{R}^n \to \mathbb{R}^n$ by

$$f(z_0) = XQ_0(\hat{\beta} - \beta).$$

The quantity $f(z_0)$ is still biased and Stein's formula lets us quantify the remaining bias in (3.61) exactly as follows:

$$\mathbb{E}[z_0^\top \boldsymbol{X}\boldsymbol{Q}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon}] = \mathbb{E}[z_0^\top f(z_0)|\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon}] = \mathbb{E}[\operatorname{div} f(z_0)|\boldsymbol{X}\boldsymbol{Q}_0, \boldsymbol{\varepsilon}].$$

The partial derivatives $(\partial/\partial z_{0i}) f_i$ where f_i is the ith coordinate of f can be computed by the chain rule

$$\frac{\partial f_i}{\partial z_{0i}} = \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{Q}_0\left[\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle \frac{\partial \hat{\boldsymbol{\beta}}}{\partial y_i} + \frac{\partial \hat{\boldsymbol{\beta}}}{\partial z_{0i}}\right].$$

Hence, the divergence of f, which quantifies the remaining bias in (3.61) is $\operatorname{div} f = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle \hat{\nu} + \hat{B}$, where

$$\hat{\nu} = \operatorname{trace}\left[\boldsymbol{X}\boldsymbol{Q}_0 \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}}\right], \qquad \hat{B} = \operatorname{trace}\left[\boldsymbol{X}\boldsymbol{Q}_0 \frac{\partial \hat{\boldsymbol{\beta}}}{\partial z_0}\right]. \tag{3.62}$$

It will be convenient to write $\operatorname{div} f$ instead as

$$\operatorname{div} f = \langle \boldsymbol{a}_0, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle \hat{\nu} + \hat{A} \quad \text{where } \hat{A} = \hat{B} + \langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}\rangle \hat{\nu}. \tag{3.63}$$

The quantities $\hat{\nu}$, $\hat{A}$ and $\hat{B}$ above can be constructed from the observed data since they only depend on $\boldsymbol{X}$, $\boldsymbol{Q}_0$, $\boldsymbol{y}$ and the derivatives of $\hat{\boldsymbol{\beta}}$. However, the quantity $\langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ is unknown; it is the parameter of interest that we wish to estimate. This motivates the estimator of $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ defined by

$$\hat{\theta} = \langle \boldsymbol{a}_0, \hat{\boldsymbol{\beta}}\rangle + \frac{z_0^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \hat{A}}{\|z_0\|^2 - \hat{\nu}} \tag{3.64}$$

with $\hat{A}$ and $\hat{\nu}$ as in (3.62) and (3.63). This estimator $\hat{\theta}$ is constructed so that the random variable

$$\begin{aligned}(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta) - z_0^\top \boldsymbol{\varepsilon} &= \hat{A} + \hat{\nu}\langle \boldsymbol{a}_0, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle - z_0^\top f(z_0) \\ &= \operatorname{div} f(z_0) - z_0^\top f(z_0)\end{aligned} \tag{3.65}$$

is exactly mean-zero by the first-order Stein's formula (1.1). Furthermore, the variance of this random variable can be expressed exactly in terms of the derivatives of f thanks to the second-order Stein formula (2.3). Similarly, the above equality can be rewritten as

$$(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta) = \operatorname{div} f(z_0) - z_0^\top (f(z_0) - \boldsymbol{\varepsilon}), \tag{3.66}$$

which is equal to $\operatorname{div} g(z_0) - z_0^\top g(z_0)$ for $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{\varepsilon}$ since f and g have the same divergence. Hence the random variable (3.66) is exactly mean zero by the first-order Stein's formula, and the second-order Stein formula (2.3) provides an exact identity for its variance. We gather the above derivation in the following theorem.

THEOREM 3.15. *Let $\hat{\boldsymbol{\beta}}$ be an estimator such that, if we write it as a function $\hat{\boldsymbol{\beta}}(\boldsymbol{y}, z_0, \boldsymbol{X}\boldsymbol{Q}_0)$, all partial derivatives of the function $\boldsymbol{X}\boldsymbol{Q}_0\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{y}$ and z_0 exist and are in L_2. Define the estimator $\hat{\theta}$ of $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ by* (3.64), *with $\hat{\nu}$ and $\hat{A}$ as in* (3.62) *and* (3.63). *Then the random variable*

$$\frac{(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta)}{\sigma\sqrt{n}} - \frac{z_0^\top \boldsymbol{\varepsilon}}{\sigma\sqrt{n}} \tag{3.67}$$

is exactly mean-zero and its variance is exactly equal to

$$\frac{1}{n\sigma^2}\big(\mathbb{E}\big[\|\boldsymbol{X}\boldsymbol{Q}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2\big] + \mathbb{E}\big[\operatorname{trace}\big((\nabla f(z_0))^2\big)\big]\big), \tag{3.68}$$

where $f(z_0) = \boldsymbol{X}\boldsymbol{Q}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. *Furthermore, the random variable* $(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta)$ *is also mean-zero with variance equal to*

$$\mathbb{E}\big[\big\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y} - z_0\boldsymbol{a}_0^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big\|^2 + \operatorname{trace}\big((\nabla f(z_0))^2\big)\big]. \tag{3.69}$$

Theorem 3.15 is a direct consequence of the second-order Stein formula (2.3) and the analysis summarized in (3.65).

The random variable $z_0^\top\boldsymbol{\varepsilon}/(\sigma\sqrt{n})$ is asymptotically standard normal. The first claim of the above result implies that $((\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta))/\sigma\sqrt{n}$ is also asymptotically normal if (3.68) converges to 0 as $n \to +\infty$. This provides a general strategy to derive asymptotic normality results; however the calculation of the gradients of $\hat{\boldsymbol{\beta}}$ and f has to be carried out on case-by-case basis which is outside of the scope of the present paper.

The above construction provides a general scheme to de-bias an initial estimator $\hat{\boldsymbol{\beta}}$ for the estimation of a linear contrast $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$ when the covariance matrix $\boldsymbol{\Sigma}$ is known because $(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta)$ is mean-zero and its variance is exactly given by (3.69). Notably, both $(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta)$ and the quantity inside the expectation in (3.69) only depends on the unknown parameter of interest $\theta = \boldsymbol{a}_0^\top\boldsymbol{\beta}$ and known observable quantities $\boldsymbol{a}_0^\top\hat{\boldsymbol{\beta}}$, $\hat{A}$, $\hat{\nu}$, $\hat{\boldsymbol{\beta}}$ and its derivatives. One may consider the z-score $(\|z_0\|^2 - \hat{\nu})(\hat{\theta} - \theta)/V^*(\theta)^{1/2}$ where $V^*(\theta)$ is the quantity inside the expectation in (3.69). First- and second-order Stein formulae provide this z-score as the starting point to de-bias the general estimator $\hat{\boldsymbol{\beta}}$ and normalize its variance, as soon as the partial derivatives of $\hat{\boldsymbol{\beta}}$ exist. Studying the asymptotic distribution of this z-score requires tools beyond the scope of the present work and will be the subject of a forthcoming paper.

A notable feature of the above result is the random variable $\hat{\nu}$ whose role is to adjust multiplicatively the random variable $(\hat{\theta} - \theta)$ so that $(\|z_0\| - \hat{\nu})(\hat{\theta} - \theta)$ is exactly mean-zero. This adjustment accounts for the degrees-of-freedom of the initial estimator $\hat{\boldsymbol{\beta}}$. We refer to our concurrent paper [8] for theory of degrees-of-freedom adjustment in semiparametric inference about a preconceived one-dimensional parameter $\theta = \langle \boldsymbol{a}_0, \boldsymbol{\beta}\rangle$.

3.7. *Monte Carlo approximation of divergence.* The second-order Stein formula and the techniques presented in this paper also suggest a Monte Carlo method to approximate the divergence in the general case.

Suppose we are interested in the approximation of $\operatorname{div} f(\boldsymbol{y})$ at the currently observed vector $\boldsymbol{y}$. Assume that the function $f(\cdot)$ is 1-Lipschitz and its value can be quickly computed for small perturbations of $\boldsymbol{y}$, say, $f(\boldsymbol{y} + az)$ for small az. For example, when $f(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ in the linear model with a convex regularized least-squares estimator $\hat{\boldsymbol{\beta}}$, the 1-Lipschitz condition holds automatically [3] as discussed in Remark 3.3, and if $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ has already been computed by an iterative algorithm, the computation of $\hat{\boldsymbol{\beta}}(\boldsymbol{y} + az)$ would typically be fast as one can use $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ as a starting point ("warm start") to compute $\hat{\boldsymbol{\beta}}(\boldsymbol{y} + az)$. Next, with the 1-Lipschitz function $h(z) = a^{-1}(f(\boldsymbol{y} + az) - f(\boldsymbol{y}))$ and $z \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$ independent of $\boldsymbol{y}$, if $\mathbb{E}_z$ denotes the expectation with respect to z conditionally on $\boldsymbol{y}$, we have by the Gaussian Poincaré inequality that

$$\mathbb{E}_z\big[\big(z^\top h(z) - D_0\big)^2\big] \le \mathbb{E}_z\big[\big\|h(z) + \nabla h(z)z\big\|^2\big] \le 4n$$

with $D_0 = \mathbb{E}_z \operatorname{div} h(z) = \int_{\mathbb{R}^n} (2\pi)^{-n/2} e^{-\|\boldsymbol{x}\|^2/2} \operatorname{div} f(\boldsymbol{y} + a\boldsymbol{x})\, d\boldsymbol{x}$. Hence If we compute $f(\boldsymbol{y} + az_j)$ at m independent Gaussian perturbations $z_1, \dots, z_m \sim N(0, \boldsymbol{I}_n)$, inequality

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m z_j^\top h(z_j) - D_0\right)^2 \,\Big|\, \boldsymbol{y}\right] \le \frac{4n}{m}$$

holds. Here, the function $\boldsymbol{y} \to \operatorname{div}(\boldsymbol{y})$ is locally integrable and almost surely bounded by n thanks to the Lipschitzness of f. For almost every $\boldsymbol{y}$, $\boldsymbol{y}$ is a Lebesgues point of $\operatorname{div} f$ so that $D_0 \to \operatorname{div}(\boldsymbol{y})$ as $a \to 0$ by the Lebesgues differentiation theorem. Hence $D_0 \approx \operatorname{div} f(\boldsymbol{y})$ for small enough $a > 0$ and $\frac{1}{m}\sum_{j=1}^m \boldsymbol{z}_j^\top h(\boldsymbol{z}_j)$ provides a useful approximation of the divergence thanks to the above conditional variance bound. For large m, more precise results can be obtained by the central limit theorem. Note that by the second-order Stein formula, $\frac{1}{m}\sum_{j=1}^m \boldsymbol{z}_j^\top h(\boldsymbol{z}_j)$ is also close to the empirical average $\bar{D}^{(m)} = \frac{1}{m}\sum_{j=1}^m \operatorname{div} f(\boldsymbol{y}+a\boldsymbol{z}_j)$ thanks to

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m \boldsymbol{z}_j^\top h(\boldsymbol{z}_j) - \bar{D}^{(m)}\right)^2\right] = \frac{1}{m}\mathbb{E}\left[\|h(\boldsymbol{z}_1)\|^2 + \operatorname{trace}\{[\nabla h(\boldsymbol{z}_1)]^2\}\right] \le \frac{2n}{m}.$$

We apply this probabilistic procedure to the Elastic-Net and Singular Value Thresholding (SVT) [17] for which explicit formulae for $\hat{\mathsf{df}} = \operatorname{trace}[\nabla f(\boldsymbol{y})]$ are available.

Indeed, [17], equation (1.8), provides an explicit formula for the degrees-of-freedom of SVT with tuning parameter λ: If $\hat{\boldsymbol{B}}$ soft-thresholds the singular values of an observed matrix $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ with tuning parameter λ, the divergence of $\hat{\boldsymbol{B}}$ with respect to $\boldsymbol{Y}$ is given by

$$\hat{\mathsf{df}} = \sum_{i=1}^{q\wedge n}\{I_{\{\sigma_i>\lambda\}} + |q-n|(1-\lambda/\sigma_i)_+\} + 2\sum_{i=1}^{q\wedge n}\sum_{j=1, j\neq i}^{q\wedge n}\frac{\sigma_i(\sigma_i-\lambda)_+}{\sigma_i^2-\sigma_j^2},$$

where $\sigma_1, \dots, \sigma_{n\wedge q}$ are the singular values of $\boldsymbol{Y}$. We then compare on simulated data this exact formula to the above random approximation scheme. For $n = 100$, $q = 101$, $\lambda = 10.0$, with $\boldsymbol{Y}$ being the sum of standard normal noise plus a ground-truth rank-10 matrix, we apply the above algorithm with m perturbations $(\boldsymbol{Y} + a\boldsymbol{Z}_j)_{j=1,\dots,m}$ for various values of m and compute $\hat{\mathsf{df}}_{\text{approx}} = \frac{1}{m}\sum_{j=1}^m \operatorname{trace}\{\boldsymbol{Z}_j^\top h(\boldsymbol{Z}_j)\}$ where $h(\boldsymbol{Z}_j) = a^{-1}(\hat{\boldsymbol{B}}(\boldsymbol{Y}+a\boldsymbol{Z}_j) - \hat{\boldsymbol{B}}(\boldsymbol{Y}))$ with $a = 0.0001$ as explained above. The results are in Figure 1.

In the case of the Elastic-Net with ℓ_1 parameter $\lambda > 0$ and ℓ_2 parameter $\gamma > 0$, we draw a similar experiment with the exact formula for degrees-of-freedom being given by $\hat{\mathsf{df}} = \operatorname{trace}[\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top\boldsymbol{X}_{\hat{S}} + n\gamma)^{-1}\boldsymbol{X}_{\hat{S}}^\top]$ as in Section 3.5.3. With $n = 500$, $p = 400$, and again $a = 0.001$, $\boldsymbol{X}$ having independent symmetric ± 1, $\lambda = 0.8\sqrt{4\log(p)/n}$, $\gamma = 0.2\sqrt{4\log(p)/n}$, we obtain the standard errors and boxplots in Figure 2.

The experiments show that the above approximation scheme provides good approximations in these special cases where exact formula are available. Hence it could also be useful for estimators where no exact formula is available for the divergence.

APPENDIX A: NONSMOOTH FUNCTIONS

PROOF OF THEOREM 2.1(ii) FOR LIPSCHITZ FUNCTIONS. If f is Lipschitz, then each component f_i of f is also Lipschitz. Hence f_i belongs to the space $W^{1,2}(\gamma_n)$ defined above (2.2) and the weak gradient of f_i is equal almost everywhere to its gradient in the sense of Frechet differentiability (cf., e.g., [26], Theorems 4–6, pp. 279–281). Thus (ii) is a consequence of (iii). □

PROOF OF THEOREM 2.1(iii) FOR $f_i \in W^{1,2}(\gamma_n)$. Since $W^{1,2}(\gamma_n)$ is the completion with respect to the norm (2.2) of the space $C_0^\infty(\mathbb{R}^n)$ of smooth functions with compact support, for each coordinate $i = 1, \dots, n$ there exists a sequence $(g_{i,q})_{q\ge 1}$ of $C_0^\infty(\mathbb{R}^n)$ functions with $\max_{i=1,\dots,n}\mathbb{E}[(f_i - g_{i,q})^2 + \|\nabla f_i - \nabla g_{i,q}\|^2] \to 0$ as $q \to +\infty$. Define $g_q : \mathbb{R}^n \to \mathbb{R}^n$ as the function with components $g_{1,q}, \dots, g_{n,q}$. By considering a subsequence, we may assume that for all $q \ge 1$,

$$\mathbb{E}[\|g_q(\boldsymbol{z}) - f(\boldsymbol{z})\|^2] + \mathbb{E}[\|\nabla g_q(\boldsymbol{z}) - \nabla f(\boldsymbol{z})\|_F^2] \le 2^{-q-2}$$

m	10	25	50	75	100	125	150	175	200	225	df_exact
mean	2600.98	2600.17	2598.41	2597.98	2597.39	2597.74	2597.91	2597.78	2597.90	2597.73	2598.75
std	22.06	11.96	9.52	8.21	6.91	5.98	5.41	4.72	5.02	5.56	NA

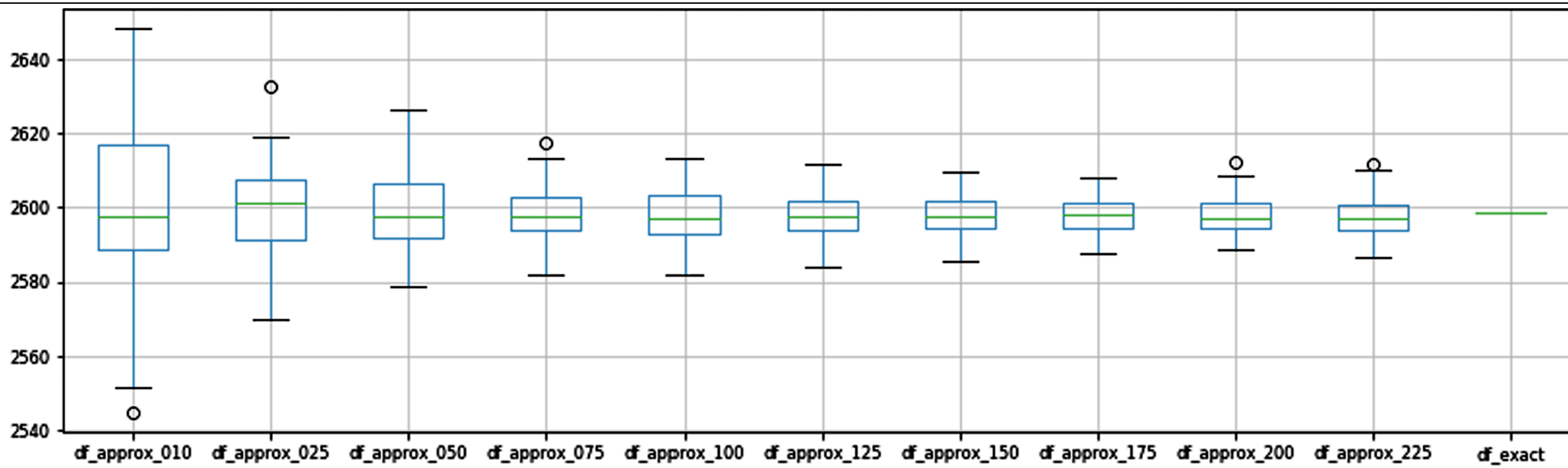

FIG. 1. *Approximate* $\hat{\mathsf{df}}$ *for SVT, computed over* 50 *realisations of* $(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_m)$ *for various values of* m. *The rightmost column is the excact formula from* [17]. *The value of* $\boldsymbol{Y}$ *is the same over all* 50 *realisations.*

m	10	25	50	75	100	125	150	175	200	225	250	df_exact
mean	52.16	52.06	52.23	52.39	52.34	52.33	52.40	52.35	52.41	52.40	52.43	52.45
std	2.91	1.77	1.36	1.18	1.02	0.88	0.89	0.88	0.77	0.72	0.68	NA

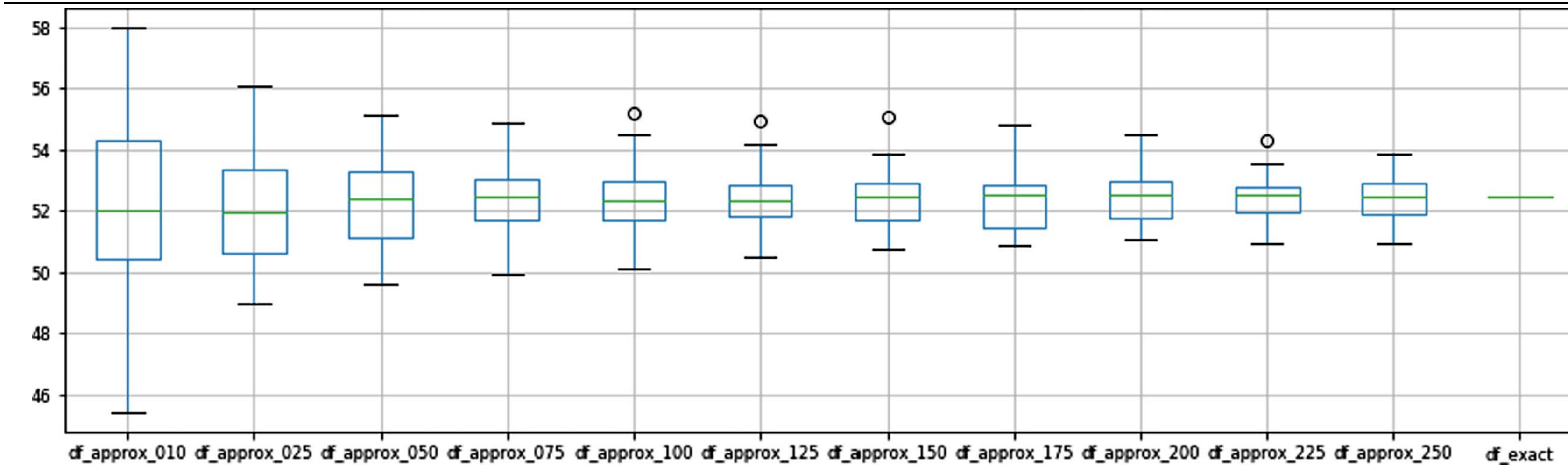

FIG. 2. *Approximate* div$(X\hat{\beta})$ *for the Elastic-Net, computed over* 50 *realisations of* $(z_1, \dots, z_m)$ *for various values of m The rightmost column is the excact formula from* [17]. *Corresponding boxplots are visible in Figure* 1. *The value of* (X, y) *is the same over all* 50 *realisations*.

which implies that $g_q \to f$ and $\nabla g_q \to \nabla f$ pointwise almost surely by the Borel–Cantelli lemma. Let $X_q = z^\top g_q(z) - \operatorname{div} g_q(z)$ and $X = z^\top f(z) - \operatorname{div} f(z)$. Then $X_q \to X$ almost surely. The triangle inequality and Theorem 2.1(i) applied to $g_k - g_{k+1}$ yields

$$\begin{aligned}
&\{\mathbb{E}[(X_q - X)^2]\}^{1/2} \\
&\quad \le \sum_{k=q}^{\infty} \{\mathbb{E}[(X_k - X_{k+1})^2]\}^{1/2} \\
&\quad \le \sum_{k=q}^{\infty} \{\mathbb{E}[\|g_k(z) - g_{k+1}(z)\|^2] + \mathbb{E}[\|\nabla g_k(z) - \nabla g_{k+1}(z)\|_F^2]\}^{1/2} \\
&\quad \le \sum_{k=q}^{\infty} 2^{-k/2} \to 0.
\end{aligned}$$

Hence, with another application of Theorem 2.1(i),

$$\begin{aligned}
&\mathbb{E}[(z^\top f(z) - \operatorname{div} f(z))^2] \\
&\quad = \lim_{q\to\infty} \mathbb{E}[(z^\top g_q(z) - \operatorname{div} g_q(z))^2] \\
&\quad = \lim_{q\to\infty} \mathbb{E}[\|g_q(z)\|_2^2 + \operatorname{trace}((\nabla g_q)^2(z))] \\
&\quad = \mathbb{E}[\|f(z)\|_2^2 + \operatorname{trace}((\nabla f)^2(z))].
\end{aligned}$$

For the first and last equality, we use the fact that if two sequences $(Z_q)_{q\ge1}$ and $(Y_q)_{q\ge1}$ and two random variables Y_∞, Z_∞ are such that $\mathbb{E}[(Y_q - Y_\infty)^2] \to 0$ and $\mathbb{E}[(Z_q - Z_\infty)^2] \to 0$ as $q \to +\infty$ then $\mathbb{E}[Y_q^2] \to \mathbb{E}[Y_\infty^2]$ and $\mathbb{E}[Y_q Z_q] \to \mathbb{E}[Y_\infty Z_\infty]$. This completes the proof. □

APPENDIX B: PROOF OF CONSISTENCY

PROOF OF THEOREM 3.2. (i) Since $\mathbb{E}[\widehat{\mathsf{SURE}}] = \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$,

$$\mathbb{E}[\widehat{R}'_{\mathrm{SURE}}] = \mathbb{E}[2\sigma^2\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + 2\sigma^2\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] \ge \sigma^2\mathbb{E}[\|\boldsymbol{\varepsilon}\|^2] = \sigma^4 n.$$

(ii) As $((\widehat{\mathsf{SURE}})_+^{1/2} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|)^2 \le |\widehat{\mathsf{SURE}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2|$, by the triangle inequality

$$\begin{aligned}
&\mathbb{E}[((\widehat{\mathsf{SURE}})_+{}^{1/2} - \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]^{1/2})^4]^{1/4} \\
&\quad \le R_{\mathrm{SURE}}^{1/4} + \mathbb{E}[(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]^{1/2})^4]^{1/4}.
\end{aligned}$$

If $\boldsymbol{y} \to \widehat{\boldsymbol{\mu}}$ is a 1-Lipschitz function, then $\boldsymbol{\varepsilon} \to \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$ is also 1-Lipschitz and the second term above is bounded from above as follows:

$$\begin{aligned}
&\mathbb{E}[(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]^{1/2})^4]^{1/4} \\
&\quad \le \mathbb{E}[(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|)^4]^{1/4} + \mathbb{E}[(\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]^{1/2})^4]^{1/4} \\
&\quad \le (2+1)\sigma = 3\sigma
\end{aligned}$$

by $\int_0^\infty \mathbb{P}(|(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|| > \sigma x) 4x^3\,dx \le \int_0^\infty 8e^{-x^2/2}x^3\,dx = 16$ for the first term, and by $0 \le \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] - (\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|)^2 \le \sigma^2$ thanks to the Gaussian Poincaré inequality for the second term. This yields (3.10).

(iii) Let $\boldsymbol{h} = \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ and $\hat{\mathsf{df}} = \operatorname{div} \widehat{\boldsymbol{\mu}}$. By (3.7), $\widehat{\mathsf{SURE}}$ is a variable of form (2.10) with $f(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon} - 2\boldsymbol{h}$ and $g(\boldsymbol{\varepsilon}) = -\|\boldsymbol{h}\|^2$. As $\nabla f(\boldsymbol{\varepsilon}) = \boldsymbol{I}_n - 2\nabla\widehat{\boldsymbol{\mu}}$ and $\nabla g(\boldsymbol{\varepsilon}) = -2(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{h}$, (2.12) implies

$$
\begin{aligned}
&\operatorname{Var}(\widehat{\mathsf{SURE}})\\
&\quad\le \mathbb{E}\big[\sigma^2\|\boldsymbol{\varepsilon} - 2\boldsymbol{h} + 2(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{h}\|^2 + \sigma^4 \operatorname{trace}\big((\boldsymbol{I}_n - 2\nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\
&\quad= R_{\mathrm{SURE}} + 4\sigma^2\mathbb{E}\big[(\boldsymbol{\varepsilon} - 2\boldsymbol{h})^\top(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{h} + \|(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{h}\|^2\big]\\
&\quad= R_{\mathrm{SURE}} + \sigma^2\mathbb{E}\big[\boldsymbol{\varepsilon}^\top(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{\varepsilon} - (\boldsymbol{\varepsilon} - 2\boldsymbol{h})^\top(\nabla\widehat{\boldsymbol{\mu}})(\boldsymbol{\varepsilon} - 2\boldsymbol{h}) + 4\boldsymbol{h}^\top\{(\nabla\widehat{\boldsymbol{\mu}})^2 - \nabla\widehat{\boldsymbol{\mu}}\}\boldsymbol{h}\big]\\
&\quad\le R_{\mathrm{SURE}} + \sigma^4 n,
\end{aligned}
$$

where the first equality follows from (3.6), the second is simple algebra and the last inequality follows from $\mathbf{0}_{n\times n} \preceq (\nabla\widehat{\boldsymbol{\mu}})^2 \preceq \nabla\widehat{\boldsymbol{\mu}} \preceq \boldsymbol{I}_n$.

(iv) Similar to (iii), $\widehat{R}'_{\mathrm{SURE}}/(4\sigma^2) = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + \sigma^2\hat{\mathsf{df}} - \sigma^2 n/2$ is a variable of form (2.10) with $f(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon} - \boldsymbol{h}$ and $g(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^\top\boldsymbol{h} - \|\boldsymbol{h}\|^2 - \sigma^2 n/2$. Thus, (2.12) implies

$$
\begin{aligned}
&\operatorname{Var}\big(\widehat{R}'_{\mathrm{SURE}}/(4\sigma^2)\big)\\
&\quad\le \mathbb{E}\big[\sigma^2\|(\nabla\widehat{\boldsymbol{\mu}} - \boldsymbol{I}_n)(2\boldsymbol{h} - \boldsymbol{\varepsilon})\|^2 + \sigma^4 \operatorname{trace}\big((\boldsymbol{I}_n - \nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\
&\quad\le \mathbb{E}\big[\sigma^2\|2\boldsymbol{h} - \boldsymbol{\varepsilon}\|^2 + \sigma^4 \operatorname{trace}\big((\boldsymbol{I}_n - \nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\
&\quad= \sigma^2\mathbb{E}\big[4\|\boldsymbol{\varepsilon} - \boldsymbol{h}\|^2 + 2\sigma^2(\hat{\mathsf{df}} - n) + \sigma^2 \operatorname{trace}\big((\nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\
&\quad= (3/4)R'_{\mathrm{SURE}} + R_{\mathrm{SURE}}/4 - \sigma^4\mathbb{E}[\hat{\mathsf{df}}]\\
&\quad= \mathbb{E}\big[\widehat{R}''_{\mathrm{SURE}}\big]
\end{aligned}
$$

with the $\widehat{R}''_{\mathrm{SURE}}$ in (3.13). As $\widehat{R}_{\mathrm{SURE}} \le \widehat{R}'_{\mathrm{SURE}}$, $\operatorname{Var}(\widehat{R}'_{\mathrm{SURE}}/(4\sigma^2)) \le \mathbb{E}[\widehat{R}'_{\mathrm{SURE}}]$. □

PROOF OF REMARK 3.3. The claim that $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ is 1-Lipschitz is proved in [3], Proposition 3. The symmetry and positivity of $\nabla\widehat{\boldsymbol{\mu}}$ is proved in [9], Proposition J.1, with the argument outlined here: $\widehat{\boldsymbol{\mu}}(\boldsymbol{y}) \in \partial u(\boldsymbol{y})$ where the function $u(\boldsymbol{y}) = \|\boldsymbol{y}\|^2/2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{y}) - \boldsymbol{y}\|^2/2 - g(\hat{\boldsymbol{\beta}}(\boldsymbol{y}))$ is convex, Alexandrov's theorem on the almost sure second-order differentiability of convex functions given, for instance, in [42], Theorem D.2.1, grants that the Hessian of u is almost surely symmetric positive semidefinite, and the Hessian of u equals $\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y})$ almost surely. □

APPENDIX C: PROOFS: CONFIDENCE REGIONS WITH SURE

PROOF OF THEOREM 3.4 AND COROLLARY 3.5. Assume $\sigma = 1$ without loss of generality. As $f(\boldsymbol{y}) = \widehat{\boldsymbol{\mu}} - \boldsymbol{y} = (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \boldsymbol{\varepsilon}$,

$$
\begin{aligned}
\widehat{\mathsf{SURE}} - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 &= \|\widehat{\boldsymbol{\mu}} - \boldsymbol{y}\|^2 - \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 + 2\operatorname{div}(\widehat{\boldsymbol{\mu}} - \boldsymbol{y}) + n\\
&= \|\boldsymbol{\varepsilon}\|^2 - 2\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) + 2\operatorname{div}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - n,
\end{aligned}
$$

so that (3.18) is a direct consequence of Theorem 2.1. By Markov' inequality,

$$
\begin{aligned}
&\mathbb{P}\big\{\big|\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \operatorname{div}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big| \ge v_0\sqrt{n/2}\big\}\\
&\quad\le \mathbb{E}\big[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + \operatorname{trace}\big((\nabla\widehat{\boldsymbol{\mu}}(\boldsymbol{y}))^2\big)\big]/(v_0^2 n/2) \le \epsilon_n.
\end{aligned}
$$

The conclusion follows from the definition of v_α and the union bound.

Corollary 3.5 follows from Theorem 3.4 with $v_0^2 = 2\gamma_n^{1/2}$, $\epsilon_n = \gamma_n^{1/2}$ and the continuous mapping theorem. □

PROOF OF THEOREM 3.6. Assume $\sigma = 1$ without loss of generality. Set $\hat{\mathsf{df}} = \operatorname{trace}[\nabla\widehat{\boldsymbol{\mu}}]$, $\boldsymbol{h} = \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ and $W = \widehat{\mathsf{SURE}} - \mathbb{E}[\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] - \|\boldsymbol{\varepsilon}\|^2 + n$. As W is of the form (2.10) with $f(\boldsymbol{\varepsilon}) = -2\boldsymbol{h}$ and $g(\boldsymbol{\varepsilon}) = \mathbb{E}[\|\boldsymbol{h}\|^2] - \|\boldsymbol{h}\|^2$,

$$\begin{aligned}\operatorname{Var}(W) &\le \mathbb{E}\big[\|-2\boldsymbol{h} + (\nabla\widehat{\boldsymbol{\mu}})(2\boldsymbol{h})\|^2 + \operatorname{trace}\big((2\nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\ &\le \mathbb{E}\big[4\|\boldsymbol{h}\|^2 + 4\operatorname{trace}\big((\nabla\widehat{\boldsymbol{\mu}})^2\big)\big]\\ &\le 4\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]\end{aligned}$$

by (2.12). This gives (3.22). Let $X_n = |W|/\mathbb{E}[W^2]^{1/2}$. We have

$$|W| \le 2X_n\big(\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]\big)^{1/2}. \tag{C.1}$$

As $\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}$ is of form (2.10) with $f(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon} - 3\boldsymbol{h}$ and $g(\boldsymbol{\varepsilon}) = -\boldsymbol{\varepsilon}^\top\boldsymbol{h} - \|\boldsymbol{h}\|^2$, (2.12) yields

$$\operatorname{Var}(\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}) \le \mathbb{E}\big[\|\boldsymbol{\varepsilon} - 2\boldsymbol{h} + (\nabla\widehat{\boldsymbol{\mu}})(2\boldsymbol{h} + \boldsymbol{\varepsilon})\|^2 + \operatorname{trace}\big((\boldsymbol{I}_n - 3\nabla\widehat{\boldsymbol{\mu}})^2\big)\big].$$

Using $\operatorname{trace}[(\nabla\widehat{\boldsymbol{\mu}})^2] \le \hat{\mathsf{df}}$, the second term is bounded by $n - 6\hat{\mathsf{df}} + 9\hat{\mathsf{df}} = n + 3\hat{\mathsf{df}}$. Since $(2\boldsymbol{h} + \boldsymbol{\varepsilon})(\nabla\widehat{\boldsymbol{\mu}})^2(2\boldsymbol{h} + \boldsymbol{\varepsilon}) \le (2\boldsymbol{h} + \boldsymbol{\varepsilon})(\nabla\widehat{\boldsymbol{\mu}})(2\boldsymbol{h} + \boldsymbol{\varepsilon})$, expanding the square yields

$$\begin{aligned}&\operatorname{Var}(\widehat{\mathsf{SURE}} + \hat{\mathsf{df}})\\ &\quad\le \mathbb{E}\big[\|\boldsymbol{\varepsilon} - 2\boldsymbol{h}\|^2 + (2\boldsymbol{h} + \boldsymbol{\varepsilon})(\nabla\widehat{\boldsymbol{\mu}})(2\boldsymbol{h} + \boldsymbol{\varepsilon}) + 2(\boldsymbol{\varepsilon} - 2\boldsymbol{h})^\top(\nabla\widehat{\boldsymbol{\mu}})(2\boldsymbol{h} + \boldsymbol{\varepsilon})\big]\\ &\qquad + \mathbb{E}[n + 3\hat{\mathsf{df}}]\\ &\quad= \mathbb{E}\big[n + 3\hat{\mathsf{df}} + \|\boldsymbol{\varepsilon} - 2\boldsymbol{h}\|^2 + 4\boldsymbol{\varepsilon}^\top(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{\varepsilon}\big] - \mathbb{E}\big[(\boldsymbol{\varepsilon} - 2\boldsymbol{h})^\top(\nabla\widehat{\boldsymbol{\mu}})(\boldsymbol{\varepsilon} - 2\boldsymbol{h})\big]\\ &\quad\le \mathbb{E}\big[4\widehat{\mathsf{SURE}} + (6n - \hat{\mathsf{df}})\big],\end{aligned}$$

where the last inequality follows from $\boldsymbol{\varepsilon}^\top(\nabla\widehat{\boldsymbol{\mu}})\boldsymbol{\varepsilon} \le \|\boldsymbol{\varepsilon}\|^2$ and Stein's formulae $\mathbb{E}[-4\boldsymbol{\varepsilon}^\top\boldsymbol{h}] = \mathbb{E}[-4\hat{\mathsf{df}}]$ and $\mathbb{E}[\|\boldsymbol{h}\|^2] = \mathbb{E}[\widehat{\mathsf{SURE}}]$. Hence there exists a random variable $Y_n \ge 0$ with $\mathbb{E}[Y_n^2] \le 1$ such that almost surely $\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}] \le \widehat{\mathsf{SURE}} + \hat{\mathsf{df}} + 2Y_n\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]^{1/2} + Y_n\sqrt{6n}$. By completing the square,

$$\big(\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]^{1/2} - Y_n\big)^2 \le \widehat{\mathsf{SURE}} + \hat{\mathsf{df}} + Y_n^2 + Y_n\sqrt{6n}. \tag{C.2}$$

Combining (C.1) and (C.2) above, we get almost surely

$$|W| \le 2X_n\big[(\widehat{\mathsf{SURE}} + \hat{\mathsf{df}})_+^{1/2} + 2Y_n + \sqrt{Y_n}(6n)^{1/4}\big]$$

which is equivalent to (3.23). As $X_n \le 1/\sqrt{\beta_1}$ and $Y_n \le 1/\sqrt{\beta_2}$ with probability at least probability $1 - \beta_1 - \beta_2$, the conclusions follow.

For the estimation of $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$, we set $W = \widehat{\mathsf{SURE}} - \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\varepsilon}\|^2 + n$ with $\mathbb{E}[W^2] \le 4\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]$ in virtue of $\operatorname{trace}(\{\nabla\widehat{\boldsymbol{\mu}}\}^2) \le \hat{\mathsf{df}}$ by 1-Lipschitzness of $\widehat{\boldsymbol{\mu}}$ so that the upper bounds for $\mathbb{E}[\widehat{\mathsf{SURE}} + \hat{\mathsf{df}}]$ still apply. □

APPENDIX D: PROOFS OF THEOREM 3.7, PROPOSITION 3.8 AND (3.32)

The proof of Theorem 3.7 requires the following lemma.

LEMMA D.1. *Consider the sequence model* $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. *Let* $\widehat{\boldsymbol{\mu}}^{(1)}$, $\widehat{\boldsymbol{\mu}}^{(2)}$ *be two estimators that are L-Lipschitz functions of* $\boldsymbol{y}$ *and let* $f(\boldsymbol{y}) = \widehat{\boldsymbol{\mu}}^{(2)} - \widehat{\boldsymbol{\mu}}^{(1)}$. *Let* $\widetilde{\boldsymbol{\mu}}$ *be the estimator among* $\{\widehat{\boldsymbol{\mu}}^{(1)}, \widehat{\boldsymbol{\mu}}^{(2)}\}$ *with the smallest* $\widehat{\mathsf{SURE}}$. *Then either*

$$\mathbb{E}[\Delta^4] \le 8\sigma^4 \mathbb{E}\,\mathrm{trace}[\{\nabla f(\boldsymbol{y})\}^2] \quad or \quad \mathbb{E}[\Delta^2] \le 8(\sqrt{2}L + 1)\sigma^2$$

holds, where $\Delta = \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{j=1,2} \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|$.

PROOF OF LEMMA D.1. Let $\xi = \|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{\mu}\|^2 - \widehat{\mathsf{SURE}}^{(1)} + \widehat{\mathsf{SURE}}^{(2)} = 2\,\mathrm{div}\, f(\boldsymbol{y}) - 2\boldsymbol{\varepsilon}^\top f(\boldsymbol{y})$. By definitions of $\widetilde{\boldsymbol{\mu}}$ and Δ,

$$\begin{aligned} |\xi| &\ge \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{j=1,2} \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2 \\ &= \{\|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}\| + \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{\mu}\|\}\Delta \\ &\ge (\|f(\boldsymbol{y})\|\Delta) \vee \Delta^2. \end{aligned}$$

If $\sigma^2 \mathbb{E}\,\mathrm{trace}[\{\nabla f(\boldsymbol{y})\}^2] \ge \mathbb{E}[\|f(\boldsymbol{y})\|^2]$, then $\mathbb{E}[\Delta^4] \le \mathbb{E}[\xi^2]$ and the identity (2.7) for $\mathbb{E}[\xi^2]/4$ yields the bound on $\mathbb{E}[\Delta^4]$.

If now $\sigma^2 \mathbb{E}\,\mathrm{trace}[\{\nabla f(\boldsymbol{y})\}^2] \le \mathbb{E}[\|f(\boldsymbol{y})\|^2]$, as $\boldsymbol{y} \to \|f(\boldsymbol{y})\|$ is $2L$-Lipschitz

$$\begin{aligned} \mathbb{E}[\|f(\boldsymbol{y})\|^2]\mathbb{E}[\Delta^2] &\le \mathbb{E}[\|f(\boldsymbol{y})\|^2\Delta^2] + \{\mathrm{Var}(\|f(\boldsymbol{y})\|^2)\mathbb{E}[\Delta^4]\}^{1/2} \\ &\le \mathbb{E}[\xi^2] + \{\mathrm{Var}(\|f(\boldsymbol{y})\|^2)\mathbb{E}[\xi^2]\}^{1/2} \\ &\le 8\sigma^2\mathbb{E}[\|f(\boldsymbol{y})\|^2] + \sqrt{(4L)^2 8}\sigma^2\mathbb{E}[\|f(\boldsymbol{y})\|^2], \end{aligned}$$

by the Cauchy–Schwarz inequality for the first inequality and the Gaussian Poincaré inequality $\mathrm{Var}[\|f(\boldsymbol{y})\|^2] \le \mathbb{E}[\|2\{\nabla f(\boldsymbol{y})\} f(\boldsymbol{y})\|^2] \le (4L)^2\mathbb{E}[\|f(\boldsymbol{y})\|^2]$ for the second. Hence $\mathbb{E}[\Delta^2] \le 8(\sqrt{2}L + 1)\sigma^2$. $\square$

PROOF OF THEOREM 3.7. Let $j_0 = \arg\min_{j=1,\dots,m} \mathbb{E}\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|$. For $k \in [m]$, let

$$\Delta_k = I_k(\|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\|)_+,$$

where I_k is the indicator of the event that the $\widehat{\mathsf{SURE}}$ of $\widehat{\boldsymbol{\mu}}^{(k)}$ is smaller than the $\widehat{\mathsf{SURE}}$ of $\widehat{\boldsymbol{\mu}}^{(j_0)}$. Then by Lemma D.1 we have $\mathbb{E}[A_k] \le 1$ with

$$A_k = \min\{\Delta_k^4/(8\sigma^4 s^*), \Delta_k^2/(8\sigma^2(\sqrt{2}L + 1))\}$$

and $X = m^{-1}\max_{k\in[m]} A_k$ has also $\mathbb{E}[X] \le 1$. Then almost surely,

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\| \le \Delta_{\hat{k}} \le \sigma \max\{(8s^*mX)^{1/4}, (8(\sqrt{2}L + 1)mX)^{1/2}\} \tag{D.1}$$

so that $\mathbb{P}(X > 1/\alpha) \le \alpha$ yields (i).

(ii) Set $W_k = (\|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|)_+$. For each $k \in [m]$, the function $\boldsymbol{y} \to \|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|$ is $2L$-Lipschitz with negative expectation so that $\mathbb{P}(W_k > 2L\sigma\sqrt{2x}) \le e^{-x}$ for all $x > 0$ by Gaussian concentration and $\mathbb{P}(\max_{k\in[m]} W_k > 2L\sigma\sqrt{2\log(m/\delta)}) \le \delta$ by the union bound.

(iii) is obtained using (D.1), $\mathbb{E}[X] \le 1$ and $\mathbb{E}[\max_{k\in[m]} W_k^2] \le 8L^2\sigma^2\log(em)$ by integration.

(iv) For (3.29), let $k_0 = \arg\min_{j\in[m]} \mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2]$. For all $j, k \in [m]$, we have for $\xi_{j,k} = \|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2 - \widehat{\mathsf{SURE}}^{(k)} - \|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2 + \widehat{\mathsf{SURE}}^{(j)}$ that

$$\begin{aligned} \mathbb{E}[\xi_{j,k}^2] &= 4\mathbb{E}[\sigma^2\|\widehat{\boldsymbol{\mu}}^{(j)} - \widehat{\boldsymbol{\mu}}^{(k)}\|^2 + 4\sigma^4\,\mathrm{trace}(\{\nabla\widehat{\boldsymbol{\mu}}^{(k)} - \nabla\widehat{\boldsymbol{\mu}}^{(j)}\}^2)] \\ &\le 16L^2\sigma^4 n + 16L^2\sigma^4 n = 32L^2\sigma^4 n \end{aligned}$$

by assumption. Since $G = \max_{k\in[m]} \xi_{k_0,k}^2/(32L^2\sigma^4 mn)$ has $\mathbb{E}[G] \le 1$ and $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(k_0)} - \boldsymbol{\mu}\|^2 \le (32Gmn)^{1/2} L\sigma^2$ holds a.s., we get (3.29). □

PROOF OF PROPOSITION 3.8. Let $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\|^2 = \sigma^2\sqrt{n}$. Choose

$$\boldsymbol{\mu} = \mathbf{0}, \qquad \widehat{\boldsymbol{\mu}}^{(1)}(\boldsymbol{y}) = \mathbf{0}, \qquad \widehat{\boldsymbol{\mu}}^{(2)}(\boldsymbol{y}) = \boldsymbol{v} + G(\boldsymbol{y}),$$

where $G : \mathbb{R}^n \to \mathbb{R}^n$ with $G(\boldsymbol{y})_i = g(y_i)$ and g is defined as the only function that is $2^{n+1}\sigma$ periodic, symmetric ($g(-u) = g(u)$) and with $g(u) = \sigma(\frac{u}{\sigma} \wedge (2^n - \frac{u}{\sigma}))$ on $[0, 2^n\sigma]$. Observe that g and G are both 1-Lipschitz. Furthermore, $\mathbb{P}(g'(y_i) = \pm 1) = 1/2$ by symmetry of g and $\operatorname{div} G(\boldsymbol{y}) = \sum_{i=1}^n r_i$ where $(r_1, \dots, r_n)$ are i.i.d. with $\mathbb{P}(r_i = \pm 1) = 1/2$. Since $\widehat{\boldsymbol{\mu}}^{(1)}$ is the oracle, that is, it has smaller risk than $\widehat{\boldsymbol{\mu}}^{(2)}$, we now study the event $\Omega = \{\widehat{\mathsf{SURE}}^{(2)} < \widehat{\mathsf{SURE}}^{(1)}\}$ in which $\widetilde{\boldsymbol{\mu}}$ selects the worse estimator. Event Ω can be rewritten $\{\|\boldsymbol{v} + G(\boldsymbol{y})\|^2 - 2\boldsymbol{\varepsilon}^\top(\boldsymbol{v} + G(\boldsymbol{y})) + 2\sigma^2 \operatorname{div} G(\boldsymbol{y}) < 0\}$. We have $\mathbb{P}\{\operatorname{div} G(\boldsymbol{y}) \le -\sqrt{n}\} \ge C_0 > 0$ by the reverse of Hoeffding inequality given in [39], Theorem 7.3.2, while $\|\boldsymbol{v} + G(\boldsymbol{y})\|^2 - 2\boldsymbol{\varepsilon}^\top(\boldsymbol{v} + G(\boldsymbol{y})) = \sigma^2\sqrt{n}(1 + o_{\mathbb{P}}(1))$ since $\|\boldsymbol{v}\|^2 = \sigma^2\sqrt{n}$, $|G| \le \sigma 2^{-n}$ and $\boldsymbol{\varepsilon}^\top\boldsymbol{v} = O_{\mathbb{P}}(\sigma\|\boldsymbol{v}\|)$. This proves that $\mathbb{P}(\Omega) \ge C_3 > 0$ for instance with $C_3 = C_0/2$ and any $n \ge C_1$ for large enough C_1. On Ω, we have $\widetilde{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}^{(2)}$ as well as

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \|\widehat{\boldsymbol{\mu}}^{(1)} - \boldsymbol{\mu}\| = \|\widehat{\boldsymbol{\mu}}^{(2)} - \boldsymbol{\mu}\| = \|\boldsymbol{v} + G(\boldsymbol{y})\| \ge C_2\sigma n^{1/4}. \qquad \square$$

PROOF OF (3.32) ASSUMING $\boldsymbol{y} \to \widehat{\boldsymbol{\mu}}^{(j)}(\boldsymbol{y})$ IS L-LIPSCHITZ FOR ALL j. Let $k \in [m]$ be fixed. Proposition 3.2 in [5] states that $\widetilde{\boldsymbol{\mu}}_Q$ satisfies

$$\begin{aligned}
&\|\widetilde{\boldsymbol{\mu}}_Q - \boldsymbol{\mu}\|^2 - \|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\|^2 \\
&\quad \le \max_{k\in[m]} 2\{\boldsymbol{\varepsilon}^\top(\widehat{\boldsymbol{\mu}}^{(k)} - \widehat{\boldsymbol{\mu}}^{(j_0)}) - \sigma^2 \operatorname{div}(\widehat{\boldsymbol{\mu}}^{(k)} - \widehat{\boldsymbol{\mu}}^{(j_0)}) - \|\widehat{\boldsymbol{\mu}}^{(j_0} - \widehat{\boldsymbol{\mu}}^{(k)}\|^2/4\}.
\end{aligned}$$

Let $W_{j_0,k}$ be the random variable inside the maximum, which is of the form (2.10) with $f(\boldsymbol{\varepsilon}) = (\widehat{\boldsymbol{\mu}}^{(k)} - \widehat{\boldsymbol{\mu}}^{(j_0)})$ and $g(\boldsymbol{\varepsilon}) = \|f(\boldsymbol{\varepsilon})\|^2/4$. Then $G = \max_{j\in[m]}(W_{j_0,k} - \mathbb{E}[W_{j_0,k}])_+^2/(m \operatorname{Var}[W_{j_0,k}])$ has $\mathbb{E}[G] \le 1$ and $\operatorname{Var}[W_{j_0,k}] \le \sigma^2\|(\boldsymbol{I}_n - \frac{1}{2}\nabla f)f(\boldsymbol{\varepsilon})\|^2 + \sigma^4 s^* \le (1 + L)^2\sigma^2\mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{j_0} - \widehat{\boldsymbol{\mu}}^{(k)}\|^2] + \sigma^4 s^*$ by (2.12). Then almost surely

$$\begin{aligned}
W_{j_0,k} &\le \sqrt{Gm}\operatorname{Var}[W_{j,k}]^{1/2} - \mathbb{E}[\|\widehat{\boldsymbol{\mu}}^j - \widehat{\boldsymbol{\mu}}^{(k)}\|^2/4] \\
&\le \sigma^2\sqrt{Gms^*} + (1 + L)\sigma\sqrt{G}\mathbb{E}[\|\widehat{\boldsymbol{\mu}}^j - \widehat{\boldsymbol{\mu}}^{(k)}\|^2]^{1/2} - \mathbb{E}[\|\widehat{\boldsymbol{\mu}}^j - \widehat{\boldsymbol{\mu}}^{(k)}\|^2/4] \\
&\le \sigma^2\sqrt{Gms*} + (1 + L)^2\sigma^2 G.
\end{aligned}$$

The proof is complete using $\mathbb{E}[G] \le 1$ and $\mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{(j_0)} - \boldsymbol{\mu}\|^2] - \min_{j\in[m]}\mathbb{E}[\|\widehat{\boldsymbol{\mu}}^{(j)} - \boldsymbol{\mu}\|^2] \le \sigma^2 L^2$ by definition of j_0 and L-Lipschitzness of $\widehat{\boldsymbol{\mu}}^{(j_0)}$. □

APPENDIX E: STRICTNESS OF THE KKT CONDITIONS

PROOF OF PROPOSITION 3.9. Assume that $\boldsymbol{X}_B$ has rank strictly less than $|B|$. Then there must exist some $j \in B$ and $A \subseteq B \setminus \{j\}$ with $\boldsymbol{x}_j = \sum_{k\in A}\gamma_k\boldsymbol{x}_k$ and $\operatorname{rank}(\boldsymbol{X}_A) = \min(|A|, n)$. By the definition of B,

$$\lambda n\delta_j = \boldsymbol{x}_j^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\lambda)}) = \lambda n\sum_{k\in A}\gamma_k\delta_k,$$

where $\delta_k = \boldsymbol{x}_k^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\lambda)})/(\lambda n) \in \{-1, 1\}$. This is impossible by Assumption 3.1 on $\boldsymbol{X}$. Hence $\boldsymbol{X}_B$ has rank $|B|$. For the uniqueness, consider two Lasso solutions $\hat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\lambda)}$ and $\hat{\boldsymbol{b}}$ of

(3.34). It is easily seen that $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} = \boldsymbol{X}\hat{\boldsymbol{b}}$ by the strict convexity of the squared loss in $\boldsymbol{X}\boldsymbol{b}$ in (3.34); actually the function $\boldsymbol{y} \to \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ is 1-Lipschitz (cf., for instance, [6]). Furthermore, both $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$, $\hat{\boldsymbol{b}}$ must be supported on B. Hence $\boldsymbol{X}_B(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_B = \boldsymbol{X}_B\hat{\boldsymbol{b}}_B$ which implies that $\hat{\boldsymbol{b}}_B = (\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_B$ because $\boldsymbol{X}_B$ has rank $|B|$.

It remains to show that for any $j \notin \widehat{S}$, the KKT conditions on coordinate j holds strictly with probability one. As $\boldsymbol{X}_B$ has rank $|B|$, it suffices to consider the case of $|\widehat{S}| < n$. By the KKT conditions, $(\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_{\widehat{S}} = (\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{X}_{\widehat{S}})^{-1}\{\boldsymbol{X}_{\widehat{S}}^\top \boldsymbol{y} - n\lambda\,\text{sgn}((\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}})_{\widehat{S}})\}$. As $\mathbb{P}[\boldsymbol{v}^\top \boldsymbol{y} = c] = 0$ for all deterministic $\boldsymbol{v} \neq \boldsymbol{0}$ and real c,

$$\mathbb{E}\big(\mathbb{P}\big[\boldsymbol{x}_j^\top\{\boldsymbol{y} - \boldsymbol{X}_S(\boldsymbol{X}_S^\top\boldsymbol{X}_S)^{-1}(\boldsymbol{X}_S^\top\boldsymbol{y} - n\lambda\boldsymbol{u}_S)/n\} = \pm\lambda|\boldsymbol{X}\big]\big) = 0$$

for all deterministic $\{S, j, \boldsymbol{u}\}$ satisfying $\text{rank}(\boldsymbol{X}_S) = |S| < n$, $\text{rank}(\boldsymbol{X}_{S\cup\{j\}}) = |S| + 1$ and $\boldsymbol{u}_S \in \{\pm 1\}^S$. Hence, $\mathbb{P}[|B| > |\widehat{S}|] = 0$, which means that the KKT conditions of $\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}}$ must hold strictly with probability one. □

APPENDIX F: PROOF: BOUND ON THE VARIANCE OF $|\hat{S}|$

PROOF OF THEOREM 3.11. Assume $\sigma = 1$ without loss of generality due to scale invariance. The first claim follows from (3.37) and the discussion leading to it, combined with Theorem 2.2. Next, we first use the rough bounds $|\widehat{S}| \le n$ and $\|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\| \le \|\boldsymbol{\varepsilon}\|$ to obtain $\text{Var}[|\widehat{S}|] \le 2n$. For the right term of the minimum, for a fixed $A \subset [p]$, the random variable $\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2$ has chi-squared distribution with at most $|A|$ degrees of freedom and a classical tail bound (cf., for instance, [34], Lemma 1) states that

$$\mathbb{P}\big(\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2 > 2|A| + 3x\big) \le \mathbb{P}\big(\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2 > |A| + 2\sqrt{x|A|} + 2x\big) \le e^{-x}.$$

Consequently, by the union bound over all $\binom{p}{m} \le (\frac{ep}{m})^m$ supports A of size m,

$$\mathbb{P}\left(\max_{A\subset[p]:|A|=m} \|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2 > 2m + 3\left(m\log\left(\frac{ep}{m}\right) + x\right)\right) \le e^{-x}.$$

By a second union bound over all possible support sizes $m = 1, \dots, p$,

$$\mathbb{P}\left(\max_{A\subset[p]}\left\{\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2 - 2|A| - 3\left(|A|\log\left(\frac{ep}{|A|\vee 1}\right)\right)\right\} > 3(\log p + x)\right) \le e^{-x}.$$

Finally, let $X = (1/3)\max_{A\subset[p]}\{\|\boldsymbol{P}_A\boldsymbol{\varepsilon}\|^2 - 2|A| - 3(|A|\log(\frac{ep}{|A|}) + \log p)\}$ so that $\mathbb{P}(X > x) \le e^{-x}$ holds. The identity $\mathbb{E}[\max(X, 0)] = \int_0^\infty \mathbb{P}(X > x)\,dx \le 1$ yields

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{P}_{\widehat{S}}\boldsymbol{\varepsilon}\|^2\big] &\le 2\mathbb{E}|\widehat{S}| + 3\mathbb{E}\big[|\widehat{S}|\log\big(ep/\{|\widehat{S}|\vee 1\}\big) + \log(ep)\big] \\ &\le 2\mathbb{E}|\widehat{S}| + 4\mathbb{E}\big[|\widehat{S}|\log\big(ep/\{|\widehat{S}|\vee 1\}\big)\big].\end{aligned} \tag{F.1}$$

The proof is complete as the second inequality in (3.39) follows from the concavity of the function $x \to x\log(ep/(x\vee 1))$. □

APPENDIX G: PRELIMINARIES FOR BOUNDS ON $\mathbb{E}|\widehat{S}|$

LEMMA G.1. *Let Z be a standard normal random variable. Then*

$$\mathbb{P}[Z > t] \le \frac{e^{-t^2/2}}{(2\pi t^2 + 4)^{1/2}}, \quad \forall t \ge 0,$$

$$\mathbb{E}\big[(|Z| - t)_+\big] \le \frac{2e^{-t^2/2}}{(2\pi)^{1/2}(t^2 + 1)}, \quad \forall t \ge 0,$$

and

$$\mathbb{E}\big[(|Z|-t)_+^2\big] \le \frac{4e^{-t^2/2}}{(t^2+2)(2\pi t^2+4)^{1/2}}, \quad \forall t \ge 0. \tag{G.1}$$

REMARK. Compared with the usual tail probability bounds for standard Gaussian, the upper bounds in Lemma G.1 is sharp at both $t=0$ and $t\to\infty$.

PROOF. Let $t>0$. Let $\varphi(t)$ and $\Phi(t)$ respectively be the density and cumulative distribution function of Z. With $u = tx + x^2/2$ and $du = (t+x)\,dx$,

$$\frac{\Phi(-t)}{\varphi(t)} = \int_0^\infty e^{-tx-x^2/2}\,dx = \int_0^\infty \frac{e^{-u}\,du}{(t^2+2u)^{1/2}} = \int_0^\infty f_t(u^{-1/2})e^{-u}\,du,$$

where $f_t(x) = (t^2+2/x^2)^{-1/2} = x(t^2x^2+2)^{-1/2}$ is a concave function of x. Thus,

$$\Phi(-t)/\varphi(t) \le f_t\big(\Gamma(1/2)\big) = f_t(\sqrt{\pi}) = \big(t^2+2/\pi\big)^{-1/2}.$$

This gives the tail probability bound. From the well known $t(1+t^2)^{-1}\varphi(t) \le \Phi(-t)$, we also have

$$\mathbb{E}\big[(|Z|-t)_+\big] = 2\big\{\varphi(t) - t\Phi(-t)\big\} \le 2\varphi(t)/\big(1+t^2\big).$$

Define $J_k(t) = \int_0^\infty x^k e^{-x-x^2/(2t^2)}\,dx$. For the second tail moment, we have $\mathbb{E}[(|Z|-t)_+^2] = 2\varphi(t)J_2(t)/t^3 = 2\Phi(-t)J_2(t)/\{t^2J_0(t)\}$. As in Proposition 10(i) in [48] and its proof, (G.1) follows from $J_2(t)/J_0(t) \le 1/(1/2+1/t^2)$ due to the recursion $J_{k+1}(t) + t^{-2}J_{k+2}(t) = (k+1)J_k(t)$ for $k\ge 0$. $\square$

APPENDIX H: UPPER BOUND ON THE SPARSITY OF THE LASSO UNDER THE SRC

PROOF OF PROPOSITION 3.12. The SRC (3.41) can be written as

$$\Big(\max_{B\subset[p]:|B\setminus S|\le m} \phi_{\text{cond}}(\boldsymbol{X}_B^\top \boldsymbol{X}_B) - 1\Big)(|S| + \epsilon_2 k) + \epsilon_1 k \le 2(1-\eta)^2 m. \tag{H.1}$$

Let $\boldsymbol{g} = (n\lambda)^{-1}\boldsymbol{X}^\top \boldsymbol{P}_S^\perp \boldsymbol{\varepsilon}$ with elements $g_j \sim N(0,\sigma_j^2)$ satisfying $1/\sigma_j \ge t_\tau = \eta^{-1} \times \sqrt{(1+\tau)2\log(p/k)}$. Let $C_0 = \max_{B:|B\setminus S|\le m}\phi_{\text{cond}}(\boldsymbol{X}_B^\top\boldsymbol{X}_B) - 1$ and

$$\Omega = \big\{4(1-\eta)\big\|(|\boldsymbol{g}|-\eta)_+\big\|_1 + C_0\big\|(|\boldsymbol{g}|-1)_+\big\|_2^2 \;<\; \epsilon_1 k + C_0\epsilon_2 k\big\}.$$

It follows from Lemma H.1 below that $|\widehat{S}\setminus S| < m$ in this event Ω. Applying Lemma G.1, we find that

$$4(1-\eta)\mathbb{E}\big[\big\|(|\boldsymbol{g}|-\eta)_+\big\|_1\big] \le \frac{8(1-\eta)pe^{-(\eta t_\tau)^2/2}}{\sqrt{2\pi}t_\tau(1+(\eta t_\tau)^2)} \le \frac{\epsilon_1 k(k/p)^\tau}{1+(\eta t_0)^2}$$

with $\epsilon_1 = 8(1-\eta)/\{\sqrt{2\pi}t_0\}$ and

$$\mathbb{E}\big[\big\|(|\boldsymbol{g}|-1)_+\big\|_2^2\big] \le \frac{4e^{-t_\tau^2/2}}{t_\tau^2(t_\tau^2+2)(2\pi t_\tau^2+4)^{1/2}} \le \frac{\epsilon_2 k(k/p)^\tau}{1+(\eta t_0)^2}$$

with $\epsilon_2 = 4/\{t_0^2(2\pi t_0^2+4)^{1/2}\}$. Thus, $\mathbb{P}\{\Omega\}$ is no smaller than $1-(k/p)^\tau/(1+(\eta t_0)^2)$. Finally, $\mathbb{E}|\widehat{S}|$ is bounded from above by $|S| + m + \mathbb{E}[I_{\Omega^c}|\widehat{S}|]$ so that combining $\mathbb{E}[I_{\Omega^c}|\widehat{S}|] \le p\mathbb{P}(\Omega^c)$ with the previous bound on $\mathbb{P}(\Omega^c)$ gives the upper bound on $\mathbb{E}|\widehat{S}|$. $\square$

The following lemma is a slight modification of [8], Proposition 7.4.

LEMMA H.1 (Deterministic lemma). *Let* $\lambda, \epsilon_1, \epsilon_2 > 0$ *and* $\eta \in (0,1)$. *Let* $\hat{\boldsymbol{\beta}}$ *be the Lasso* (3.34) *with* $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ *and* $S = \mathrm{supp}(\boldsymbol{\beta})$. *If for* $\boldsymbol{g} = \boldsymbol{X}^\top \boldsymbol{P}_S^\perp \boldsymbol{\varepsilon}/(n\lambda)$ *and* $C_0 = \max_{B:|B\setminus S|\le m} \phi_{\mathrm{cond}}(\boldsymbol{X}_B^\top \boldsymbol{X}_B) - 1$, *we have*

$$4(1-\eta)\big\|(|\boldsymbol{g}|-\eta)_+\big\|_1 + C_0\big\|(|\boldsymbol{g}|-1)_+\big\|_2^2 \;<\; \epsilon_1 k + C_0\epsilon_2 k$$

and the SRC (3.41) *holds*, *then* $|\widehat{S}\setminus S| < m$.

PROOF. The SRC (3.41) can be written as (H.1). Let $\overline{\boldsymbol{\beta}}$ the oracle LSE satisfying $\mathrm{supp}(\overline{\boldsymbol{\beta}}) \subseteq S$ and $\boldsymbol{X}\overline{\boldsymbol{\beta}} = \boldsymbol{P}_S\boldsymbol{y}$. Let B satisfy

$$S \cup \widehat{S} \subseteq B \subseteq S \cup \big\{j : \big|\boldsymbol{x}_j^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\big| = \lambda\big\}.$$

Let $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$, $B_1 = B\setminus S$, $\boldsymbol{u} = (\hat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}})/\lambda$ and $\boldsymbol{v} = \overline{\boldsymbol{\Sigma}}\boldsymbol{u}$. By algebra,

$$\boldsymbol{v}_B^\top \overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\boldsymbol{v}_B + \boldsymbol{v}_{B_1}^\top\big(\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\big)_{B_1,B_1}\boldsymbol{v}_{B_1} = \boldsymbol{v}_S^\top\big(\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\big)_{S,S}\boldsymbol{v}_S + 2\boldsymbol{u}_{B_1}^\top\boldsymbol{v}_{B_1}.$$

Let $\boldsymbol{s} = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n\lambda) = \boldsymbol{g} - \boldsymbol{v}$. By the KKT conditions, $s_j = \mathrm{sgn}(u_j)$ for $j \in B_1$, so that $\boldsymbol{u}_{B_1}^\top\boldsymbol{v}_{B_1} = \sum_{j\in B\setminus S}|u_j|(s_j g_j - 1) \le \boldsymbol{v}_B^\top\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\boldsymbol{w}_B$, where $\boldsymbol{w}$ is the vector with elements $w_j = I\{j\in B_1\}s_j(s_j g_j - 1)_+$. By algebra,

$$\begin{aligned}(\boldsymbol{v}-\boldsymbol{w})_B^\top\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}(\boldsymbol{v}-\boldsymbol{w})_B &+ \boldsymbol{v}_{B_1}^\top\big(\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\big)_{B_1,B_1}\boldsymbol{v}_{B_1}\\ &\le \boldsymbol{v}_S^\top\big(\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\big)_{S,S}\boldsymbol{v}_S + \boldsymbol{w}_B\overline{\boldsymbol{\Sigma}}_{B,B}^{-1}\boldsymbol{w}_B.\end{aligned}$$

It follows that

$$\|\boldsymbol{v}_S\|_2^2 + \|\boldsymbol{v}_{B_1} - \boldsymbol{w}_{B_1}\|_2^2 + \|\boldsymbol{v}_{B_1}\|_2^2 \le \phi_{\mathrm{cond}}(\overline{\boldsymbol{\Sigma}}_{B,B})\big(\|\boldsymbol{v}_S\|_2^2 + \|\boldsymbol{w}_{B_1}\|_2^2\big).$$

Moreover, as $v_j - w_j = -s_j(s_j g_j - 1)_-$ for $j\in B_1$, $\|\boldsymbol{v}_{B_1}\|_2^2 = \|\boldsymbol{v}_{B_1} - \boldsymbol{w}_{B_1}\|_2^2 + \|\boldsymbol{w}_{B_1}\|_2^2$. This and the inequality above imply

$$2\|\boldsymbol{v}_{B_1} - \boldsymbol{w}_{B_1}\|_2^2 \le \phi_{\mathrm{cond}}(\overline{\boldsymbol{\Sigma}}_{B,B} - 1)\big(\|\boldsymbol{v}_S\|_2^2 + \|\boldsymbol{w}_{B_1}\|_2^2\big).$$

As $\boldsymbol{v} = \boldsymbol{g} - \boldsymbol{s}$ and $\boldsymbol{g}_S = 0$, $\|\boldsymbol{v}_S\|_2^2 = |S|$. We also have $\|\boldsymbol{w}_{B_1}\|_2^2 \le \sum_{j=1}^p (s_j g_j - 1)_+^2$ so that $2\|\boldsymbol{v}_{B_1} - \boldsymbol{w}_{B_1}\|_2^2 \le C_0(|S| + \|(|\boldsymbol{g}|-1)_+\|_2^2)$. As $s_j = \mathrm{sgn}(u_j)$ and $|v_j - w_j| = (1 - s_j g_j)_+$ in B_1,

$$\begin{aligned}2(1-\eta)^2|B_1| &\le 2\|\boldsymbol{v}_{B_1} - \boldsymbol{w}_{B_1}\|_2^2 + 4(1-\eta)\sum_{j=1}^p (s_j g_j - \eta)_+\\ &\le C_0\big(|S| + \big\|(|\boldsymbol{g}|-1)_+\big\|_2^2\big) + 4(1-\eta)\big\|(|\boldsymbol{g}|-\eta)_+\big\|_1\\ &< C_0\big(|S| + \epsilon_2 k\big) + \epsilon_1 k.\end{aligned}$$

Thus, by (H.1), $|B_1| \le m$ implies $|B_1| < m$. As $|B_1|$ is allowed to change one-at-a-time along the Lasso path from $\lambda = \infty$ and the condition on $\boldsymbol{g}$ is monotone in λ, $|B_1| < m$ holds for all penalty levels satisfying the condition on $\boldsymbol{g}$. For details of this argument, see [59], Proof of Lemma 1. □

APPENDIX I: UPPER BOUND ON THE SPARSITY OF THE LASSO UNDER THE RE CONDITION

PROOF OF THEOREM 3.13. Let $\mu = (1+\tau)\sigma\sqrt{2\log(ep/(s_0 \vee 1))/n}$ and $\lambda = (1+\gamma)\mu$. For each $j \in [p]$ set $g_j = (\tau+1)\boldsymbol{\varepsilon}^\top \boldsymbol{X}\boldsymbol{e}_j/n$. By the KKT conditions of the Lasso, we have

$$(\tau\boldsymbol{\varepsilon}^\top \boldsymbol{X}\boldsymbol{h} + \|\boldsymbol{X}\boldsymbol{h}\|^2)/n \le \boldsymbol{g}^\top \boldsymbol{h} - \lambda\|\boldsymbol{h}_{S^c}\|_1 - \lambda\operatorname{sgn}(\boldsymbol{\beta}_S)^\top \boldsymbol{h}_S,$$

where $\boldsymbol{h} = \hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\beta}$. Define $T = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}_{S^c}\|_1 < c_0\sqrt{s_0 \vee 1}\|\boldsymbol{u}\|\}$ as well as

$$f_{\lambda,\mu}(\boldsymbol{\varepsilon}) = \sup_{\boldsymbol{u}\in T} \frac{(\boldsymbol{g}^\top \boldsymbol{u} - \mu\|\boldsymbol{u}_{S^c}\|_1 - \lambda\operatorname{sgn}(\boldsymbol{\beta}_S)^\top \boldsymbol{u}_S)_+}{\|\boldsymbol{u}\|},$$

$$g_{\lambda,\mu}(\boldsymbol{\varepsilon}) = \sup_{\boldsymbol{u}\notin T, \|\boldsymbol{X}\boldsymbol{u}\|>0} \frac{(\boldsymbol{g}^\top \boldsymbol{u} - \mu\|\boldsymbol{u}_{S^c}\|_1 - \lambda\operatorname{sgn}(\boldsymbol{\beta}_S)^\top \boldsymbol{u}_S - \gamma\lambda c_0\sqrt{s_0\vee 1}\|\boldsymbol{u}\|)_+}{\|\boldsymbol{X}\boldsymbol{u}\|/\sqrt{n}},$$

where $a_+ = \max(a, 0)$ for any real a. Let Ω be the event $\Omega = \{\boldsymbol{h} \in T\}$ and I_Ω be its indicator function. Using the elementary inequality $2ab - b^2 \le a^2$, we get

$$(2\tau\boldsymbol{\varepsilon}^\top \boldsymbol{X}\boldsymbol{h} + \|\boldsymbol{X}\boldsymbol{h}\|^2)/n \le I_\Omega f_{\lambda,\lambda}(\boldsymbol{\varepsilon})^2/\operatorname{RE}(S, c_0)^2 + I_{\Omega^c} g_{\lambda,\lambda}(\boldsymbol{\varepsilon})^2.$$

We now bound the expectation $\mathbb{E}[f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2]$. By simple algebra on each coordinate and the Cauchy–Schwarz inequality,

$$f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2 = \sum_{j\in S}(g_j - \lambda\operatorname{sgn}(\beta_j))^2 + \sum_{j\notin S}(|g_j| - \mu)^2.$$

Each g_j is centered, normal with variance at most $\omega^2 = (1+\tau)^2\sigma^2/n$, hence Lemma G.1 implies that $\mathbb{E}[f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2] \le s_0(\lambda^2 + \omega^2) + (s_0 \vee 1)\omega^2$, which is then bounded by $2^{-1}(\gamma c_0)^2(s_0 \vee 1)\lambda^2$ by the condition on c_0.

Note that by construction, the function $\boldsymbol{\varepsilon} \to g_{\lambda,\mu}(\boldsymbol{\varepsilon})$ is $((1+\tau)/\sqrt{n})$-Lipschitz, so that by the Gaussian concentration inequality (see, e.g., [14], Theorem 10.17),

$$\mathbb{E}[g_{\lambda,\mu}(\boldsymbol{\varepsilon})^2] = \int_0^{+\infty} \mathbb{P}[g_{\lambda,\mu}(\boldsymbol{\varepsilon}) > \sqrt{t}]\,dt \le \int_0^{+\infty} \mathbb{P}[\omega N(0,1) > \sqrt{t}]\,dt = \omega^2/2,$$

provided that the median of $g_{\lambda,\mu}(\boldsymbol{\varepsilon})$ is zero. We now prove that the median is indeed zero. The event $\{f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2 \le 2\mathbb{E}[f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2]\}$ has probability at least $1/2$ thanks to Markov's inequality. Furthermore, we proved above that $2\mathbb{E}[f_{\lambda,\mu}(\boldsymbol{\varepsilon})^2] \le (c_0\gamma)^2(s_0 \vee 1)\lambda^2$. On this event of probability at least $1/2$, for any $\boldsymbol{u} \in T$ we have

$$(\boldsymbol{g}^\top \boldsymbol{u} - \mu\|\boldsymbol{u}_{S^c}\|_1 - \lambda\operatorname{sgn}(\boldsymbol{\beta}_S)^\top \boldsymbol{u}_S - (c_0\gamma)\lambda\sqrt{s_0 \vee 1}\|\boldsymbol{u}\|)_+ = 0.$$

We have established that the median of $g_{\lambda,\mu}$ is nonpositive and the proof is complete. □

APPENDIX J: CONSISTENCY OF SURE FOR SURE IN THE LASSO CASE

PROOF OF PROPOSITION 3.14. (i)–(iv) follow directly from Theorem 3.2. For (v), Markov's inequality with the bound in (iii) followed by the bound $\sigma^4 n \le \mathbb{E}[\hat{R}_{\text{SURE}}]$ from (ii) implies

$$|\hat{R}_{\text{SURE}} - \mathbb{E}[\hat{R}_{\text{SURE}}]|^2 \le 4(n\alpha)^{-1}\sigma^4 n\mathbb{E}[\hat{R}_{\text{SURE}}] \le 4(n\alpha)^{-1}\mathbb{E}[\hat{R}_{\text{SURE}}]^2$$

with probability at least $1-\alpha$, and $\mathbb{E}[\hat{R}_{\text{SURE}}] \le (1 - 4(n\alpha)^{-1/2})^{-1}\hat{R}_{\text{SURE}}$ on this event by the triangle inequality. The proof is completed using $\mathbb{P}(|\widehat{\mathsf{SURE}} - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}_{\text{LASSO}} - \boldsymbol{\mu}\|^2| \le \gamma^{-1}\mathbb{E}[\hat{R}_{\text{SURE}}]) \ge 1-\gamma$. which follows by another application of Markov's inequality. □

Acknowledgments. The first author thanks Julien Reygner and Yair Shenfeld for pointing out useful references regarding Section 2.4.

Funding. The first author was supported in part by NSF Grant DMS-1811976 and NSF CAREER award DMS-1945428. The second author was supported in part by NSF Grants DMS-1513378, IIS-1407939, DMS-1721495, IIS-1741390 and CCF-1934924.

REFERENCES

[1] ARLOT, S. and BACH, F. R. (2009). Data-driven calibration of linear estimators with minimal penalties. *Adv. Neural Inf. Process. Syst.* 46–54.

[2] BAKRY, D., GENTIL, I. and LEDOUX, M. (2014). *Analysis and Geometry of Markov Diffusion Operators. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **348**. Springer, Cham. MR3155209 https://doi.org/10.1007/978-3-319-00227-9

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

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

[5] BELLEC, P. C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46** 30–59. MR3766945 https://doi.org/10.1214/17-AOS1540

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

[7] BELLEC, P. C. and YANG, D. (2019). The cost-free nature of optimally tuning Tikhonov regularizers and other ordered smoothers. Preprint. Available at arXiv:1905.12517.

[8] BELLEC, P. C. and ZHANG, C.-H. (2018). De-biasing the lasso with degrees-of-freedom adjustment. Preprint.

[9] BELLEC, P. C. and ZHANG, C.-H. (2019). Second order Poincaré inequalities and de-biasing arbitrary convex regularizers when $p/n \to \gamma$. Preprint. Available at arXiv:1912.11943.

[10] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

[11] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.* **42** 757–788. MR3210986 https://doi.org/10.1214/14-AOS1204

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

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

[14] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[15] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

[16] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[17] CANDÈS, E. J., SING-LONG, C. A. and TRZASKO, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.* **61** 4643–4657. MR3105401 https://doi.org/10.1109/TSP.2013.2270464

[18] CHATTERJEE, S. (2014). A short survey of Stein's method. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. IV* 1–24. Kyung Moon Sa, Seoul. MR3727600

[19] CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's Method. Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2732624 https://doi.org/10.1007/978-3-642-15007-4

[20] DAI, D., RIGOLLET, P., XIA, L. and ZHANG, T. (2014). Aggregation of affine estimators. *Electron. J. Stat.* **8** 302–327. MR3192554 https://doi.org/10.1214/14-EJS886

[21] DAI, D., RIGOLLET, P. and ZHANG, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.* **40** 1878–1905. MR3015047 https://doi.org/10.1214/12-AOS1025

[22] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085 https://doi.org/10.1214/12-AOS1038

[23] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[24] DOSSAL, C., KACHOUR, M., FADILI, M. J., PEYRÉ, G. and CHESNEAU, C. (2013). The degrees of freedom of the Lasso for general design matrix. *Statist. Sinica* **23** 809–828. MR3086657

[25] ERDOGDU, M. A. (2016). Newton–Stein method: An optimization method for GLMs via Stein's lemma. *J. Mach. Learn. Res.* **17** Paper No. 216, 52. MR3595150

[26] EVANS, L. C. (1999). Partial differential equations and Monge–Kantorovich mass transfer. In *Current Developments in Mathematics*, 1997 (*Cambridge*, *MA*) 65–126. Int. Press, Boston, MA. MR1698853

[27] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.* **27** 500–518. MR3025131 https://doi.org/10.1214/12-STS398

[28] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

[29] JANZAMIN, M., SEDGHI, H. and ANANDKUMAR, A. (2014). Score function features for discriminative learning: Matrix and tensor framework. Preprint. Available at arXiv:1412.2863.

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

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

[32] JOHNSTONE, I. (1988). On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics*, *IV*, *Vol.* 1 (*West Lafayette*, *Ind.*, 1986) 361–379. Springer, New York. MR0927112

[33] KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866. MR1292543 https://doi.org/10.1214/aos/1176325498

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

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

[36] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172

[37] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135 https://doi.org/10.1214/aos/1176347253

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

[39] MATOUSEK, J. and VONDRAK, J. The Probabilistic Method. Lecture notes.

[40] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[41] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170

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

[43] RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665. MR2933661 https://doi.org/10.1214/11-AOS961

[44] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134 https://doi.org/10.1214/12-STS393

[45] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. MR0402873

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

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

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

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

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

[51] TIBSHIRANI, R. J. and ROSSET, S. (2019). Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *J. Amer. Statist. Assoc.* **114** 697–712. MR3963173 https://doi.org/10.1080/01621459.2018.1429276

[52] TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948 https://doi.org/10.1214/12-AOS1003

[53] TROPP, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory* **52** 1030–1051. MR2238069 https://doi.org/10.1109/TIT.2005.864420

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

[55] VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. https://doi.org/10.1007/978-1-4757-2545-2_3

[56] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

[57] YANG, Z., BALASUBRAMANIAN, K., WANG, Z. and LIU, H. Learning non-Gaussian multi-index model via second-order Stein's method.

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

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

[60] ZHANG, C.-H. (2011). Statistical inference for high-dimensional data. *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models* **48** 28–31. Report.

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

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

[63] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135 https://doi.org/10.1214/12-STS399

[64] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449