Error estimation and adaptive tuning for unregularized robust M-estimator

Pierre C. Bellec PIERRE.BELLEC@RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854, USA

Takuya Koriyama TKORIYAM@UCHICAGO.EDU
Booth School of Business
The University of Chicago
Chicago, IL 60637, USA

Editor:

Abstract

We consider unregularized robust M-estimators for linear models under Gaussian design and heavy-tailed noise, in the proportional asymptotics regime where the sample size n and the number of features p are both increasing such that $p/n \to \gamma \in (0, 1)$. An estimator of the out-of-sample error of a robust M-estimator is analysed and proved to be consistent for a large family of loss functions that includes the Huber loss. As an application of this result, we propose an adaptive tuning procedure of the scale parameter $\lambda > 0$ of a given loss function ρ: choosing $\hat{\lambda}$ in a given interval I that minimizes the out-of-sample error estimate of the M-estimator constructed with loss $\rho_\lambda(\cdot) = \lambda^2\rho(\cdot/\lambda)$ leads to the optimal out-of-sample error over I. The proof relies on a smoothing argument: the unregularized M-estimation objective function is perturbed, or smoothed, with a Ridge penalty that vanishes as $n \to +\infty$, and show that the unregularized M-estimator of interest inherits properties of its smoothed version.

Keywords: Robust regression, proportional regime, Huber loss, adaptive tuning, high-dimensional statistics.

1 Introduction

Robust statistics originated with the foundational work of Huber (1964), which introduced methods for estimating a location parameter under heavy-tailed noise assumptions. Over time, it has evolved into a crucial area of study, providing tools to address the challenges posed by outliers and non-standard error distributions. The practical applications of robust statistics span diverse fields, including financial modeling (Lambert-Lacroix and Zwald (2011)) and genomic data analysis (Sun et al. (2020)), underscoring its broad utility across scientific and applied domains. For a comprehensive overview of the theoretical foundations and their practical implementations in data analysis, see Loh (2024), Maronna et al. (2019), and references therein.

In this paper, we consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}$ where the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has i.i.d rows $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ and the noise $\boldsymbol{\epsilon} \in \mathbb{R}^n$ has an i.i.d. marginal F_ϵ. Our assumption (Assumption 2 below) allows F_ϵ to be heavy-tailed, including distributions with no finite moments. We focus on the high-dimensional regime where the sample size n and the dimension p are both increasing such that $p/n \to \gamma \in (0,1)$. In this setting, we consider the unregularized robust M-estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{1}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a convex and differentiable loss such that its derivative $\psi = \rho' : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz. A well-studied example is the Huber loss, which is defined by

$$\forall x \in \mathbb{R}, \quad \rho(x) = \int_0^{|x|} \min(1, u) \mathrm{d}u = \begin{cases} x^2/2 & |x| \le 1 \\ |x| - 1/2 & |x| \ge 1 \end{cases}. \tag{2}$$

Our goal is to select a robust loss achieving a small out-of-sample error in a data-driven manner, i.e., by only looking at the observation $(\boldsymbol{y}, \boldsymbol{X})$. Here, the out-of-sample error refers to the random quantity

$$R := \big\| \boldsymbol{\Sigma}^{1/2} (\hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}^\star) \big\|_2^2 = \mathbb{E}\Big[\big\{ \boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \big\}^2 | \boldsymbol{X}, \boldsymbol{y} \Big], \tag{3}$$

where $\boldsymbol{x}_0$ is independent of $(\boldsymbol{y}, \boldsymbol{X})$ and has the same law as any row $\boldsymbol{x}_i$ of $\boldsymbol{X}$. In this paper, we occasionally refer to the random quantity (3) simply as the *Risk*. Since the calculation of the out-of-sample error requires the unobservable quantities $(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$, we need an observable proxy to the out-of-sample error for such data-driven selection of loss.

Classical statistics for M-estimation, for instance confidence intervals for the coefficients of $\boldsymbol{\beta}^\star$, require that $p/n \to 0$, that is, the dimension is negligible compared to the sample size. However, the last decade has identified ample evidence that such classical asymptotic results fail in moderate dimension where dimension and sample size are of the same order (Donoho and Montanari, 2016; El Karoui et al., 2013). Beyond linear models, this phenomenon is also clearly exposed in logistic regression Sur and Candès (2019) with clear departure from classical results as soon as $n \ge 10p$ is violated. From the point of view of classical statistics, a peculiar property of the proportional asymptotic regime where $\lim p/n$ is a positive constant is that *consistency fails*, and that the error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$ does not converge to 0 in probability.

1.1 Results at a glance

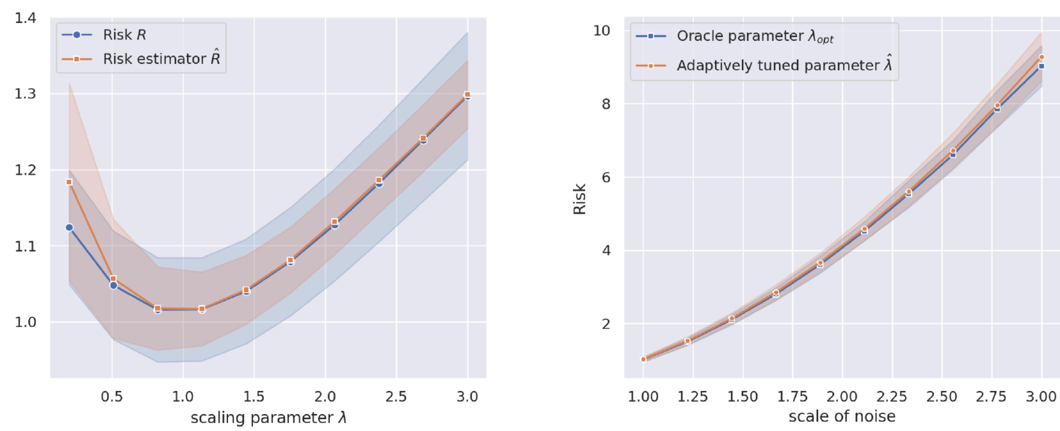
Our contribution is to propose a consistent estimator of the out-of-sample error for unregularized robust M-estimators. For $\psi = \rho' : \mathbb{R} \to \mathbb{R}$, let $\psi(\boldsymbol{r}) = (\psi(r_i))_{i=1}^n \in \mathbb{R}^n$ for any vector $\boldsymbol{r} \in \mathbb{R}^n$, in other words $\psi : \mathbb{R} \to \mathbb{R}$ acts componentwise on vectors. Then, under some regularity condition on the loss and noise distribution (see Assumption 1-2), we show that the random quantity

$$\hat{R} = \frac{p \|\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2^2}{\{\mathrm{tr}[\boldsymbol{V}]\}^2} \quad \text{with } \boldsymbol{V} := \frac{\partial \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{y}} = \Big(\frac{\partial \psi(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{\partial y_j} \Big)_{i,j} \in \mathbb{R}^{n \times n} \tag{4}$$

is consistent as an estimate of the out-of-sample error, in the sense that

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 = \hat{R} + o_P(1). \tag{5}$$

The matrix $\boldsymbol{V}$ in (4) is the Jacobian of the map $\boldsymbol{y} \mapsto \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, and its existence will be proved with probability one in Proposition 3. Note that the random quantity $\hat{R}$ in (4) is observable since both of $\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and $\operatorname{tr}[\boldsymbol{V}]$ can be computed by observation of $(\boldsymbol{y}, \boldsymbol{X})$ only. See Figure 1a for simulation. The formal statement of the consistency result (5) is in Theorem 1 below.



(a) Estimation of out-of-sample error

(b) Adaptive tuning

Figure 1: Figure 1a is the plot of the out-of-sample error R and its estimator $\hat{R}$ with the Huber loss for different scaling parameter $\lambda > 0$. Figure 1b is the plot of the oracle out-of-sample error $R(\lambda_{\text{opt}})$ and the out-of-sample error $R(\hat{\lambda})$ with $\hat{\lambda}$ being the minimizer of the estimator $\hat{R}$ among a finite grid I, as the scale of noise changes. See Section 4 for the details.

A first application of the above result in statistical inference is the ability to choose, between two competing loss functions ρ and $\tilde{\rho}$, the loss function yielding the M-estimator $\hat{\boldsymbol{\beta}}$ enjoying the narrowest confidence interval for a component $\beta_j^\star$ of $\boldsymbol{\beta}^\star$. Lemma 1 in El Karoui et al. (2013) establishes that $\sqrt{p}(\beta_j^\star - \hat{\beta}_j)/\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})\|$ is asymptotically normal with variance only depending on the j-th diagonal element of $\boldsymbol{\Sigma}^{-1}$. Thus, the size of the confidence interval is proportional $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})\|$, and the ability to choose a loss function among $\{\rho, \tilde{\rho}\}$ leading to the smallest error $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})\|$ through (5) will lead to the smallest confidence interval.

As a second application of the above result, we propose an adaptive tuning procedure of the scale parameter λ in a collection of loss functions $\{\rho_\lambda(\cdot) = \lambda^2\rho(\cdot/\lambda), \lambda \in I\}$, for some fixed interval $I \subset (0, \infty)$ and some robust loss ρ. Note that this hyperparameter λ controls the sensitivity of the loss ρ_λ to outliers. For each $\lambda > 0$, let $\hat{\boldsymbol{\beta}}(\lambda)$ be the M-estimator (1) with $\rho = \rho_\lambda$ and let $\hat{R}(\lambda)$ be the corresponding estimate of the out-of-sample error (4). Then, we show that there exists a finite subset $J \subset I$ (a finite grid made of regularly spaced parameters) such that

$$\hat{\lambda} \in \operatorname*{arg\,min}_{\lambda \in J} \hat{R}(\lambda) \quad \text{satisfies} \quad \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}(\hat{\lambda}) - \boldsymbol{\beta}^\star)\|^2 \le \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}(\lambda_{\text{opt}}) - \boldsymbol{\beta}^\star)\|^2 + o_P(1) \quad (6)$$

where λ_{opt} is the oracle parameter achieving the smallest out-of-sample error in probability in the interval I. See Figure 1b for simulation. The formal statement of the result (6) is in Theorem 3 below.

1.2 Related work

Under regularity conditions on the loss ρ and noise distribution F_ϵ, the out-of-sample error of the unregularized M-estimator (1) is known El Karoui et al. (2013); El Karoui (2018, 2013); Donoho and Montanari (2016); Thrampoulidis et al. (2018) to converge in probability to a deterministic value α^2, which is a solution to the following nonlinear system of equations with positive unknowns (α, κ):

$$\begin{aligned} \alpha^2\gamma &= \mathbb{E}[((\alpha Z + W) - \operatorname{prox}[\kappa\rho](\alpha Z + W))^2] \\ \alpha\gamma &= \mathbb{E}[((\alpha Z + W) - \operatorname{prox}[\kappa\rho](\alpha Z + W)) \cdot Z] \end{aligned} \tag{7}$$

where $Z \sim \mathcal{N}(0,1)$, $W \sim F_\epsilon$, and $\operatorname{prox}[f](x) := \arg\min_{u\in\mathbb{R}}(x-u)^2/2 + f(u)$. The convergence in probability of the out-of-sample error of $\hat{\boldsymbol{\beta}}$ to α^2 is granted by Donoho and Montanari (2016); Thrampoulidis et al. (2018) provided that (7) admits a unique solution. The existence of solutions to (7) was established by Donoho and Montanari (2016) for strongly convex losses, and in our recent paper Bellec and Koriyama (2023) for general Lipschitz losses on the side of the phase transition where exact recovery (i.e., $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$) is not possible. On the other hand, uniqueness is addressed in Thrampoulidis et al. (2018) by strict convexity arguments and in Bellec and Koriyama (2023). The working assumptions of the present paper stated in Section 2 ensure that a solution (α, κ) to (7) exists and is unique, so that the convergence in probability $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| \to^p \alpha$ as $n, p \to +\infty$ with $p/n \to \gamma < 1$ holds by the main result of Thrampoulidis et al. (2018).

Bean et al. (2013) studied the construction of the optimal loss that minimizes the solution α of (7) when F_ϵ is log-concave and known. However, constructing the optimal loss in this way requires the knowledge of F_ϵ to minimize the corresponding α in (7). Since F_ϵ is typically unknown, this construction cannot be implemented in practice.

In this paper, we focus on the underparametrized regime $p/n \to \gamma$ with $\gamma \in (0,1)$. In the overparametrized regime with $\gamma > 1$, the out-of-sample error of the unregularized M-estimator (1) explodes (cf. (Thrampoulidis et al., 2018, Remark 5.1.1)). This curse of dimensionality can be overcome by using a penalty function g and computing the corresponding penalized M-estimator $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} n^{-1}\sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}) + g(\boldsymbol{\beta})$, for which the out-of-sample error will be finite under suitable assumptions on the penalty g Bayati and Montanari (2011); Thrampoulidis et al. (2018); Loureiro et al. (2021). If $\gamma < 1$, however, some advantages of the unregularized M-estimator(1) include its invariance with respect to $(\boldsymbol{\Sigma}, \boldsymbol{\beta}^\star)$ (El Karoui et al., 2013, Lemma 1) and that confidence intervals for components $\beta_j^\star$ of $\boldsymbol{\beta}^\star$ using the asymptotic normality of $\hat{\beta}_j$ do not require knowledge of $\boldsymbol{\Sigma}$; on the other hand, the de-debiasing correction $\boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ necessary for asymptotic normality of a regularized M-estimator with penalty $g(\cdot)$ as above does require knowledge (or some estimate) of $\boldsymbol{\Sigma}$ (Bellec et al., 2022).

Most similar works to (5) are Bellec (2023); Bellec and Shen (2022); Bellec (2025); they show that for general pairs of loss ρ and penalty g, the out-of-sample error of the penalized

M-estimator $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} n^{-1}\sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}) + g(\boldsymbol{\beta})$ enjoys the approximation

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \approx \operatorname{tr}[\boldsymbol{V}]^{-2}(\|\hat{\boldsymbol{\psi}}\|_2^2(2\hat{\mathrm{df}} - p) + \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\hat{\boldsymbol{\psi}}\|_2^2), \tag{8}$$

where $\hat{\boldsymbol{\psi}} := \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^n$, $\hat{\mathrm{df}} := \operatorname{tr}[(\partial/\partial\boldsymbol{y})(\boldsymbol{X}\hat{\boldsymbol{\beta}})]$, and $\boldsymbol{V}$ as in (4). Furthermore, these derivatives have closed forms for a certain choice of (ρ, g); for instance when ρ is the Huber loss and g is the L_1 penalty, we have $\hat{\mathrm{df}} = |\hat{S}|$ and $\operatorname{tr}[\boldsymbol{V}] = |\hat{I}| - |\hat{S}|$, where $\hat{S} = \{j \in [p] : \hat{\beta}_j \neq 0\}$ is the active set and $\hat{I} = \{i \in [n] : |y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}| \le 1\}$ is the set of inliers (Bellec, 2023, Propositions 2.2 and 2.3).

With the KKT conditions of (1) giving $\boldsymbol{X}^\top\hat{\boldsymbol{\psi}} = \boldsymbol{0}_p$, we may regard (4) as a special case of (8) by setting $\hat{\mathrm{df}} = p$. However, the proof techniques in the aforementioned papers Bellec (2023); Bellec and Shen (2022); Bellec (2025) rely either on the strong convexity of the penalty g (excluding the $g = 0$ case we study here), or on a sufficiently sparse structure in $\boldsymbol{\beta}^\star$ in which case the L1 penalty also allows for approximations of the form (8) Celentano et al. (2023); Bellec (2023).

For the unregularized case (1) studied here, the results from the papers El Karoui et al. (2013); El Karoui (2018, 2013); Donoho and Montanari (2016); Thrampoulidis et al. (2018) are not sufficient to establish the approximation (5), say for the Huber loss, on the one hand because El Karoui et al. (2013); El Karoui (2018, 2013); Donoho and Montanari (2016) rely on strong convexity on either the loss or the penalty, but more importantly because these results do not study the trace of the Jacobian $\operatorname{tr}[\boldsymbol{V}]$ in (4) and how this quantity relates to the solution (α, κ) of the nonlinear system (7). For instance, the Convex Gaussian Min-max Theorem (CGMT) of Thrampoulidis et al. (2018) provides the limit in probability of $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|^2$, and reversing the argument (replacing the design matrix by its transpose) provides the limit in probability of $\|\boldsymbol{\psi}\|^2/n$ in (1), which equals $\alpha^2\gamma/\kappa^2$. However, characterizing a limit in probability for the quantity $\operatorname{tr}[\boldsymbol{V}]$ has so far remained out of reach of the CGMT. The trace of the Jacobian $\operatorname{tr}[\boldsymbol{V}]$ appears in the leave-one-out analysis of El Karoui et al. (2013); El Karoui (2013, 2018), however these works study the regularized estimate $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} n^{-1}\sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}) + \mu\|\boldsymbol{\beta}\|^2/2$ with an additive Ridge penalty $\mu\|\boldsymbol{\beta}\|^2/2$. Recently, Bellec (2025) proposed an alternative regularization technique to study unregularized estimates of the form (1), but this crucially requires the loss ρ to be twice continuously differentiable and $\rho''(x) > 0$ for all $x \in \mathbb{R}$, which rules out the Huber loss (2), one of the most common robust loss functions and a major application of the present paper.

1.3 Precise analysis of a perturbed M-estimator

To overcome these difficulties and prove (5) for the Huber loss and its variants, we use the "*Ridge-smoothing*" technique, which is used in previous works El Karoui (2013); Celentano and Montanari (2022); Loureiro et al. (2022). By rotational and translation invariance, assume without loss of generality that $\boldsymbol{\beta}^\star = \boldsymbol{0}_p$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then, (5) is reduced to

$$\|\hat{\boldsymbol{\beta}}\|_2^2 = p\|\boldsymbol{\psi}\|_2^2 \cdot (\operatorname{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}])^{-2} + o_P(1) \tag{9}$$

where $\boldsymbol{\psi} : \boldsymbol{\epsilon} \in \mathbb{R}^n \mapsto \psi(\boldsymbol{\epsilon} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^n$ is a vector field and $(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}$ is the Jacobian matrix. To show (9), we instead consider the ridge-regularized M-estimator $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ with a diminishing regularization parameter n^{-c} for some constant $c > 0$

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \frac{n^{-c}}{2} \|\boldsymbol{\beta}\|_2^2, \tag{10}$$

and define the vector field $\boldsymbol{\psi}_{\text{ridge}} : \boldsymbol{\epsilon} \mapsto \psi(\boldsymbol{\epsilon} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{ridge}})$. We prove the target (9) by the following two steps:

(I) Prove (9) for the regularized $\hat{\boldsymbol{\beta}}_{\text{ridge}}$, i.e., $\|\hat{\boldsymbol{\beta}}_{\text{ridge}}\|_2^2 \approx p\|\boldsymbol{\psi}_{\text{ridge}}\|_2^2 \cdot (\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_{\text{ridge}}])^{-2}$.

(II) Prove that each quantity in (9) is approximately the same for $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ and for $\hat{\boldsymbol{\beta}}$, as the Ridge penalty coefficient in (10) converges to 0 polynomially in n:

$$\|\hat{\boldsymbol{\beta}}_{\text{ridge}}\|_2^2 \approx \|\hat{\boldsymbol{\beta}}\|_2^2, \qquad \tfrac{1}{n}\|\boldsymbol{\psi}_{\text{ridge}}\|_2^2 \approx \tfrac{1}{n}\|\boldsymbol{\psi}\|_2^2, \qquad \tfrac{1}{n}\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_{\text{ridge}}] \approx \tfrac{1}{n}\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}].$$

We prove (I) by a chi-square type moment inequality given in (Bellec, 2023, Section 7). The more subtle and novel part of the proof is to derive the three approximations in (II), in particular the challenging approximation

$$\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_{\text{ridge}}] \approx \text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}]. \tag{11}$$

Indeed, the closeness of the two vector fields $(\boldsymbol{\psi}, \boldsymbol{\psi}_{\text{ridge}})$ in the Euclidean norm does not necessarily imply the closeness of their divergence. To show (11), we leverage the assumption that the noise distribution is sufficiently smooth: concretely, Assumption 2 below grants that the noise distribution is a convolution of two probability distributions $(F, \tilde{F})$, that is,

$$\boldsymbol{\epsilon} = \boldsymbol{z} + \boldsymbol{\delta}, \qquad \boldsymbol{z} \perp\!\!\!\perp \boldsymbol{\delta}, \qquad (z_i)_{i=1}^n \sim F, \qquad (\delta_i)_{i=1}^n \sim \tilde{F}, \tag{12}$$

where F has density $z \mapsto \exp(-\phi(z))$ such that $\phi : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable with bounded second derivative (on the other hand, $\tilde{F}$ is unrestricted and may have arbitrarily fat tails). Then, using a variant of the second order Stein's formula (Bellec and Zhang, 2021, Section 2.4) extended to the distribution with density $z \mapsto \exp(-\phi(z))$ (see Theorem 6 below), we will argue in Lemma 5 below that $\mathbb{E}[(\text{tr}[(\partial/\partial\boldsymbol{z})(\boldsymbol{\psi}_{\text{ridge}} - \boldsymbol{\psi})] - \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_{\text{ridge}} - \boldsymbol{\psi}))^2]$ is relatively small. As a consequence, using the chain rule for the Jacobian, we have

$$\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_{\text{ridge}}] - \text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}] = \text{tr}[(\partial/\partial\boldsymbol{z})(\boldsymbol{\psi}_{\text{ridge}} - \boldsymbol{\psi})] \approx \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_{\text{ridge}} - \boldsymbol{\psi}). \tag{13}$$

Roughly speaking, Assumption 2 below grants that the noise distribution is sufficiently smooth, and thanks to this assumption, closeness of the vector fields in the form $\|\boldsymbol{\psi} - \boldsymbol{\psi}_{\text{ridge}}\|^2/n = o_P(1)$ carries over to the two divergences, and implies

$$\text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}]/n = \text{tr}[(\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_{\text{ridge}}]/n + o_P(1). \tag{14}$$

It is not clear at this point if (14) holds without this smoothness assumption on the noise distribution.

Adaptive tuning of the scale parameter λ was investigated before in several papers, including Loh (2021); Wang et al. (2021) and references therein. These works do not assume a proportional asymptotics regime and do not aim for exact multiplicative constants: the resulting risk is only guaranteed to be less than $C \times$ (optimal risk) for a constant $C > 1$. On the other hand, our goal in the present paper is to achieve the optimal risk with multiplicative constant 1.

1.4 Organization

In Section 2, we derive several results on the consistency of an estimator of the out-of-sample error. In Section 3, we discuss the adaptive tuning of scale parameters. Section 4 is devoted to numerical simulations, and Section 5 gives an outline of the proof. The rigorous proofs are provided in appendix.

1.5 Notation

For any vector $\boldsymbol{u}$, we denote by $\|\boldsymbol{u}\|$ the Euclidean norm $\sqrt{\sum_i u_i^2}$. For any matrix $\boldsymbol{X}$, let $\|\boldsymbol{X}\|_{op}$ be the operator norm, i.e., the maximum singular value of $\boldsymbol{X}$. Given a vector field $\boldsymbol{f} : \boldsymbol{z} \in \mathbb{R}^n \mapsto \boldsymbol{f}(\boldsymbol{z}) \in \mathbb{R}^m$, let $(\partial/\partial \boldsymbol{z})\boldsymbol{f} \in \mathbb{R}^{m\times n}$ be the Jacobian matrix, and $\|\boldsymbol{f}\|_{\text{lip}}$ be the Lipschitz constant of $\boldsymbol{f}$ induced by the Euclidean norm. For any function $\psi : \mathbb{R} \to \mathbb{R}$, let $\psi(\boldsymbol{r}) = (\psi(r_i))_{i=1}^n \in \mathbb{R}^n$ for all $\boldsymbol{r} \in \mathbb{R}^n$, in other words $\psi : \mathbb{R} \to \mathbb{R}$ acts componentwise on vectors, and let $\|\psi\|_\infty = \sup_{x\in\mathbb{R}} |\psi(x)|$ be the sup norm. If we write $C = C_1(a, b, c)$, C is a constant depending on (a, b, c) only. If two random vectors $\boldsymbol{x}, \boldsymbol{y}$ are independent, we write $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. For a sequence of random variables $(U_n)_{n\ge1}$, we write $U_n \to^p U$ to denote convergence in probability to U and $U_n = o_P(1)$ if $U_n \to^p 0$. For a sequence of reals $r_n > 0$, we write $U_n = O_P(r_n)$ if for any $\epsilon > 0$ there exists K_ϵ such that $\sup_{n\ge1} \mathbb{P}(|U_n| > K_\epsilon r_n) < \epsilon$

2 Estimation of the out-of-sample error

Throughout, we assume that $(y_i, \boldsymbol{x}_i, \epsilon_i)_{i=1}^n$ are independently distributed according to

$$\forall i \in [n], \qquad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star + \epsilon_i, \qquad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma}), \qquad \epsilon_i \sim F_\epsilon, \quad \boldsymbol{x}_i \perp\!\!\!\perp \epsilon_i$$

where $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ is an unknown regression vector, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is some symmetric positive semi-definite matrix, and F_ϵ is a probability distribution. Since our interest is the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|^2$, and the out-of-sample error for the unregularized M-estimator $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ is invariant with respect to $(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$ (cf. El Karoui (2013) or Section 5), we do not require any assumptions on $(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})$.

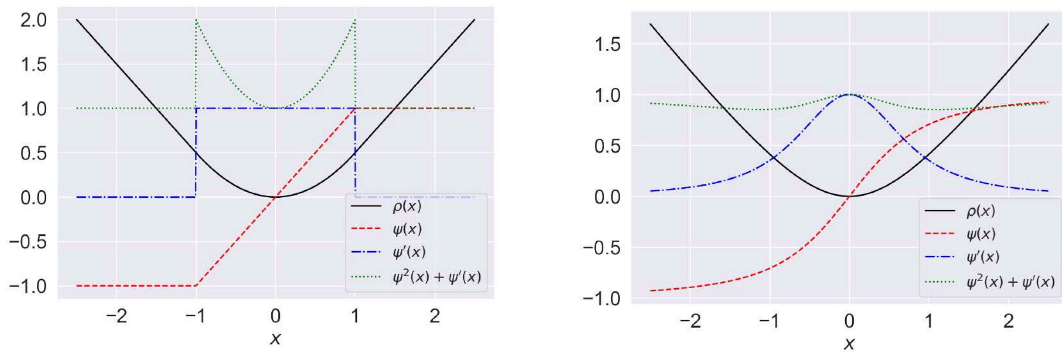
When taking a limit as $n \to \infty$, we implicitly assume that the number of features p and the sample size n are increasing such that $p/n \to \gamma \in (0, 1)$, while other quantities such ρ and F_ϵ are fixed. In this setting, it has been shown by El Karoui et al. (2013); El Karoui (2018, 2013); Donoho and Montanari (2016); Thrampoulidis et al. (2018) that if the nonlinear system of equations

$$\begin{aligned} \alpha^2\gamma &= \mathbb{E}[((\alpha Z + W) - \text{prox}[\kappa\rho](\alpha Z + W))^2] \\ \alpha\gamma &= \mathbb{E}[((\alpha Z + W) - \text{prox}[\kappa\rho](\alpha Z + W)) \cdot Z] \end{aligned} \quad \text{where } \begin{cases} Z \sim \mathcal{N}(0, 1), \\ W \sim F_\epsilon, Z \perp\!\!\!\perp W \end{cases} \tag{15}$$

admits a unique solution (α, κ), the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|^2$ converges to α^2 in probability. The existence of solutions to (15) was established in Donoho and Montanari (2016) for strongly convex ρ, and recently in our companion paper Bellec and Koriyama (2023) (see Appendix A for details).

In this section, we will argue that the random quantity $\hat{R}$ defined by

$$\hat{R} = \frac{p\|\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2^2}{\text{tr}[\boldsymbol{V}]^2}, \text{ where } \boldsymbol{V} := \frac{\partial\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{y}} \in \mathbb{R}^{n\times n}, \quad \psi = \rho' : \mathbb{R} \to \mathbb{R}, \tag{16}$$



(a) Huber $\rho(x) = \int_0^{|x|} \min(1, u)\mathrm{d}u$ (b) Pseudo Huber $\rho(x) = \sqrt{1 + x^2} - 1$

Figure 2: Example of loss satisfying Assumption 1

approximates well the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|^2$. Note that $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ is the Jacobian matrix of the map $\boldsymbol{y} \in \mathbb{R}^n \mapsto \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^n$. Before moving to the formal statement, we introduce our assumption on the loss ρ and noise distribution F_ϵ.

Assumption 1 *ρ is convex, differentiable, and $\{0\} = \arg\min_x \rho(x)$, as well as $\psi = \rho' : \mathbb{R} \to \mathbb{R}$ satisfies*

1. $\|\psi\|_\infty < +\infty$.

2. $\|\psi\|_{\mathrm{lip}} = 1$.

3. *$\exists \eta > 0$ such that $\psi(x)^2/\|\psi\|_\infty^2 + \psi'(x) \geq \eta$ for almost every $x \in \mathbb{R}$.*

Assumption 2 *F_ϵ is a convolution of $(F, \tilde{F})$, where $\tilde{F}$ is arbitrary while F has some density $z \mapsto \exp(-\phi(z))$ for some twice continuously differentiable function $\phi : \mathbb{R} \to \mathbb{R}$ with a bounded second derivative* $\sup_{x \in \mathbb{R}} |\phi''(x)| < +\infty$.

The Huber loss $\rho(x) = \int_0^{|x|} \min(1, u)\mathrm{d}u$ and its smooth approximations, such as the pseudo Huber loss $\rho(x) = \sqrt{1 + x^2} - 1$, satisfy Assumption 1 (See Figure 2). Assumption 2 requires that each component ϵ_i of the noise in the linear model is equal in distribution to $z_i + \delta_i$ for two independent random variables $z_i \sim F$ and $\delta_i \sim \tilde{F}$, where z_i has a smooth density $z \mapsto \exp(-\phi(z))$. Typical example of the distribution F in Assumption 2 is the normal distribution $\mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. We emphasize that there is no assumption on $\tilde{F}$, so that the noise distribution $F_\epsilon = F * \tilde{F}$ can be heavy-tailed. For instance, we allow $F_\epsilon = \mathcal{N}(0, 1) * \mathrm{Cauchy}(0, 1)$, in which F_ϵ has no finite moments.

Assumption 1(1) is the condition for Theorem 7 to be applicable; this is a sufficient condition to guarantee that the system (7) has a unique solution. Assumption 1(2) is mainly for the Jacobian $\boldsymbol{V}$ of the map $\boldsymbol{y} \in \mathbb{R}^n \mapsto \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^n$ in (16) to be well-defined. Indeed, under this assumption, $\boldsymbol{y} \mapsto \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is Lipschitz for almost every $\boldsymbol{X}$ (cf. Proposition 3 or Proposition 4.1 in Bellec (2023)), so that Rademacher's theorem guarantees the existence of the Jacobian $\boldsymbol{V}$ for almost every $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$. The Lipschitz assumption in

Assumption 1(2) can be relaxed to $\|\psi\|_{\text{lip}} < +\infty$ since the unregularized M-estimator (1) remains invariant under a rescaling of the loss function, i.e., $\rho \mapsto \rho/\|\psi\|_{\text{lip}}$. For simplicity and to streamline the proof arguments presented in Section 5 and the Appendix, we assume $\|\psi\|_{\text{lip}} = 1$ throughout the paper. Assumption 1(3) is used to show that $\text{tr}[\boldsymbol{V}]$ in the denominator of $\hat{R}$ in (16) is bounded from below by a positive constant times n with high probability.

Assumption 2 on F_ϵ is a technical condition to control $\text{tr}[\boldsymbol{V}]$; see the discussion surrounding (12). An equivalent formulation of Assumption 2 is that each entry ϵ_i of the noise is a sum $z_i + \delta_i$ of two independent random variables $z_i \sim F$ and $\delta_i \sim \tilde{F}$, where z_i has a smooth density $z \mapsto \exp(-\phi(z))$. Numerical simulations in Appendix D.1 suggest that our results still hold for some F_ϵ without the presence of the smooth noise component z_i, which suggests that it is an artifact of the proof. Some preliminary theoretical evidence that this assumption is not necesary is given in Lemma 6 where this assumption is replaced with $\epsilon_i = \sigma z_i + \delta_i$ with the same distributions as above but a vanishing amplitude σ for the smooth part z_i. However, our results in Section 3 require Assumption 2 (and cannot accomodate a vanishing σ) in Lemma 12.

The quantity $\text{tr}[\boldsymbol{V}]$ in (16) is observable and can be computed approximately by Monte Carlo schemes (see Section 2.11 in Bellec (2023) and the references therein) or off-the-shelf numerical methods to compute derivatives. For the Huber loss and the pseudo Huber loss, closed-form expressions for $\text{tr}[\boldsymbol{V}]$ are available: $\text{tr}[\boldsymbol{V}] = |\{i \in [n] : y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \in [-1, 1]\}| - p$ for the Huber loss and

$$\text{tr}[\boldsymbol{V}] = \sum_{i=1}^n \Big[\psi'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) - \psi'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})^2 \boldsymbol{x}_i^\top \Big(\sum_{l=1}^n \boldsymbol{x}_l \psi'(\boldsymbol{x}_l^\top \hat{\boldsymbol{\beta}}) \boldsymbol{x}_l^\top\Big)^{-1} \boldsymbol{x}_i\Big]$$

for the pseudo Huber loss $\rho(x) = \sqrt{1 + x^2} - 1$ or other twice-continuously differentiable loss functions with $\psi' = \rho''$ positive everywhere.

Remark 1 *Assumption 1-2 imply that the system (15) admits a unique solution. Indeed, according to Theorem 7, the sufficient conditions for the system to admit a unique solution are (1) ρ is convex, Lipschitz, and $\{0\} = \arg\min_{x\in\mathbb{R}} \rho(x)$, and (2) $\mathbb{P}(W \neq 0) > 0$ and $\inf_{\lambda>0} \mathbb{E}[\text{dist}(G, \lambda\partial\rho(W))^2] > 1 - \gamma$ for independent $W \sim F_\epsilon$ and $G \sim \mathcal{N}(0, 1)$. Here, condition (1) is readily satisfied by Assumption 1. For condition (2), thanks to Assumption 2, F_ϵ does not have any point mass, so in particular $\mathbb{P}_{W\sim F_\epsilon}(W \neq 0) = 1$. Since ρ is differentiable by Assumption 1, $\partial\rho(W)$ is always the singleton $\{\rho'(W)\}$, which gives $\inf_{\lambda>0} \mathbb{E}[\text{dist}(G, \lambda\partial\rho(W))^2] = \inf_{\lambda>0} \mathbb{E}[(G - \lambda\rho'(W))^2] = 1 > 1 - \gamma$.*

Now we claim that the random quantity $\hat{R}$ in (16) is a consistent estimate of the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|^2$.

Theorem 1 *Assume that (ρ, F_ϵ) satisfy Assumption 1 and 2. Let $\psi = \rho' : \mathbb{R} \to \mathbb{R}$ be the derivative of the loss, and $\boldsymbol{V}$ be the Jacobian matrix $(\partial/\partial \boldsymbol{y})\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^{n\times n}$. Then, as $n, p \to \infty$ with $p/n \to \gamma \in (0, 1)$, we have*

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|^2 = \hat{R} + o_P(1), \ \textit{where}\ \hat{R} := \frac{p\|\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2^2}{\text{tr}[\boldsymbol{V}]^2}. \tag{17}$$

An outline of the proof is given in Section 5 and the formal proof is given in Appendix B.4. We are able to relax Assumption 2 on the noise as the next proposition shows: the result (17) still holds if the smooth component of the noise has vanishing amplitude σ_n.

Proposition 1 *Let Assumption 1 be fulfilled, and assume that for each $i \in [n]$, the noise ϵ_i is equal in distribution to $\delta_i + \sigma_n z_i$ where $\delta_i \sim \tilde{F}$ and $z_i \sim F$ are independent and $(F, \tilde{F})$ are as in Assumption 2, and where $\sigma_n \to 0$ with $\sigma_n \geq n^{-1/8}$. Then (17) still holds as $n, p \to +\infty$ with $p/n \to \gamma \in (0, 1)$.*

The proof is given in Appendix B.5. Theorem 1 implies that the random quantity $\hat{R}$ is a consistent estimate of the out-of-sample error. Importantly, $\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and $\mathrm{tr}[\boldsymbol{V}]$ are both observable, i.e., they can be computed by observed data $(\boldsymbol{y}, \boldsymbol{X})$ only. Thus, $\hat{R}$ serves as a criterion to select different losses.

Corollary 1 *Assume either that F_ϵ satisfies Assumption 2, or that each iid component ϵ_i of the noise satisfies the relaxed condition in Proposition 1. For fixed integer K, consider K different loss functions $\rho_1, \ldots, \rho_K$ that satisfy Assumption 1. For each $k \in [K]$, let $\hat{\boldsymbol{\beta}}_k \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_k(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ be the M-estimator computed by the k-th loss and $\hat{R}_k$ be the corresponding criterion (16). Then, we have*

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_{\hat{k}} - \boldsymbol{\beta}^\star)\|^2 = \min_{k=1,\ldots,K} \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^\star)\|^2 + o_P(1), \quad \textit{where } \hat{k} \in \arg\min_{k=1,\ldots,K} \hat{R}_k.$$

See Appendix B.6 for the proof. Note that the left-hand side is the out-of-sample error of the M-estimator that minimizes the criterion $(\hat{R}_k)_{k=1}^K$ among K different losses, while $\min_{k=1,\ldots,K} \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^\star)\|^2$ on the right-hand side is the optimal out-of-sample error among the candidates. Thus, Corollary 1 implies that the M-estimator minimizing the criterion achieves the optimal out-of-sample error up to an error term that converges to 0 in probability.

3 Adaptive tuning of scale parameters

Let ρ be a fixed robust loss satisfying Assumption 1. For $\lambda > 0$, consider λ-*scaled* loss ρ_λ defined as

$$\forall x \in \mathbb{R}, \quad \rho_\lambda(x) := \lambda^2 \rho(x/\lambda).$$

This λ controls the sensitivity to outliers or heavy tails, and different λ lead to significanclty different performance as seen in Figures 3a and 5a. When the base loss ρ is the Huber loss or the pseudo Huber loss (see Figure 2), it holds that

$$\forall x \in \mathbb{R}, \quad \rho_\lambda(x) = \begin{cases} x^2 \cdot (x/\lambda)^{-2} \rho(x/\lambda) & \to x^2/2 \quad \text{as } \lambda \to +\infty \\ \lambda x \cdot (x/\lambda)^{-1} \rho(x/\lambda) & \sim \lambda|x| \quad \text{as } \lambda \to 0+ \end{cases},$$

thanks to $\lim_{u\to 0} \rho(u)/u^2 = 1/2$ and $\lim_{u \to \pm\infty} \rho(u)/u = \pm 1$. Informally speaking, the scaled loss ρ_λ behaves like the square loss for large λ and like the absolute loss $|x|$ for small λ.

Previous sections have so far discussed the consistency of the estimate $\hat{R}$ given by (16). In this section, we apply this consistency result to perform adaptive tuning of the parameter

λ. The goal of this section is to select some λ in a data-driven way so that the M-estimator $\hat{\boldsymbol{\beta}}_\lambda$ defined with the scaled loss ρ_λ,

$$\hat{\boldsymbol{\beta}}_\lambda \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho_\lambda(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \quad \text{with} \quad \rho_\lambda(\cdot) := \lambda^2 \rho(\cdot/\lambda) \tag{18}$$

achieves an asymptotically optimal risk. Let us introduce some useful notation. For all $\lambda > 0$, let $\psi_\lambda : \mathbb{R} \to \mathbb{R}$ be the derivative of the λ-scaled loss ρ_λ:

$$\psi_\lambda : x \mapsto \rho_\lambda'(x) = \lambda\psi(x/\lambda) \quad \text{with} \quad \psi = \rho'.$$

Note that by construction, $\|\psi_\lambda\|_{\text{lip}}$ is invariant with respect to λ. We define the random functions R, $\hat{R}$, and the deterministic function α as

$$R : (0,\infty) \to \mathbb{R}, \qquad \lambda \mapsto \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star)\|^2 \text{ with } \hat{\boldsymbol{\beta}}_\lambda \text{ being the M-estimator (18)} \tag{19}$$

$$\hat{R} : (0,\infty) \to \mathbb{R}, \qquad \lambda \mapsto p\frac{\|\psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2}{\operatorname{tr}[\boldsymbol{V}_\lambda]^2} \text{ with } \boldsymbol{V}_\lambda = \frac{\partial \psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)}{\partial \boldsymbol{y}} \in \mathbb{R}^{n\times n} \tag{20}$$

$$\alpha : (0,\infty) \to \mathbb{R}, \qquad \lambda \mapsto \alpha(\lambda) \text{ (the solution to (15) with } \rho(\cdot) = \rho_\lambda(\cdot)). \tag{21}$$

With the above notation, Donoho and Montanari (2016); Thrampoulidis et al. (2018) grants $R(\lambda) \to^p \alpha^2(\lambda)$ while Theorem 1 and Theorem 7 yield $\hat{R}(\lambda) \to^p \alpha^2$ and an explicit upper bound on $\alpha^2(\lambda)$ with respect to λ. We summarize these results in the following proposition.

Proposition 2 *Assume that (ρ, F_ϵ) satisfy Assumption 1 and 2. Then, the nonlinear system of equations (15) with $\rho = \rho_\lambda$ admits a unique solution for all $\lambda > 0$, so that the map $\lambda \mapsto \alpha^2(\lambda)$ in (21) is well-defined. Furthermore, as $n, p \to \infty$ with $p/n \to \gamma \in (0,1)$, we have*

$$\hat{R}(\lambda) \to^p \alpha^2(\lambda), \quad R(\lambda) \to^p \alpha^2(\lambda), \quad \alpha^2(\lambda) \le C_2(\gamma, F_\epsilon, \rho)(\lambda^2 + 1).$$

for all $\lambda > 0$.

See Appendix C.1 for the proof. Proposition 2 suggests that the M-estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ in (18) with $\hat{\lambda}$ minimizing the criterion $\hat{R}(\lambda)$ over a discrete grid achieves the optimal risk limit, i.e.,

$$\alpha^2(\hat{\lambda}) \approx \min_\lambda \alpha^2(\lambda) \text{ with } \hat{\lambda} \in \arg\min_\lambda \hat{R}(\lambda),$$

as long as the map $\lambda \mapsto \alpha^2(\lambda)$ is smooth enough and the grid fine enough, so that at least one element λ of the grid is sufficiently close to the optimal parameter (that may for instance fall between two consecutive elements of the grid). To ensure such smoothness condition on the function $\alpha^2(\cdot)$, we introduce an additional assumption on the loss ρ via $\psi = \rho'$.

Assumption 3 $\psi = \rho' : \mathbb{R} \to \mathbb{R}$ *satisfies the following:*

$$\sup_{\lambda,\tilde{\lambda}>0,\lambda\neq\tilde{\lambda}} \sup_{x\in\mathbb{R}} \frac{|\lambda\psi(x/\lambda) - \tilde{\lambda}\psi(x/\tilde{\lambda})|}{|\lambda - \tilde{\lambda}|} < +\infty.$$

Note that the Huber loss $\rho(x) = \int_0^{|x|} \min(1,u)du$ satisfies Assumption 3. If ρ is twice continuously differentiable, the sufficient condition for Assumption 3 is

$$\sup_{u\in\mathbb{R}} |\psi(u) - u\psi'(u)| < +\infty \tag{22}$$

Indeed, for all $\lambda, \lambda \in (0,\infty)$ with $\lambda \neq \tilde{\lambda}$ and for all $x \in \mathbb{R}$, there exists $\lambda_x > 0$ by the mean value theorem, such that

$$\frac{|\lambda\psi(x/\lambda) - \tilde{\lambda}\psi(x/\tilde{\lambda})|}{|\lambda - \tilde{\lambda}|} = \Big|\frac{d}{d\lambda}\Big(\lambda\psi\Big(\frac{x}{\lambda}\Big)\Big)\Big|_{\lambda=\lambda_x} = \Big|\psi\Big(\frac{x}{\lambda_x}\Big) - \frac{x}{\lambda_x}\psi'\Big(\frac{x}{\lambda_x}\Big)\Big| \leq \sup_{u\in\mathbb{R}}|\psi(u) - u\psi'(u)|,$$

which implies that (22) is a sufficient condition for Assumption 3. We can easily check that (22) is satisfied by the pseudo Huber loss $\rho(x) = \sqrt{1+x^2} - 1$.

Now we will claim that the map $\lambda \mapsto \alpha^2(\lambda)$ in (21) is localy 1/2-Hölder continuous. See Appendix C.2 for the proof.

Theorem 2 *Assume that (ρ, F_ϵ) satisfy Assumption 1, 2, and 3. For each $\lambda > 0$, let $\alpha^2(\lambda)$ be the solution to the nonlinear system of equations (15) with $\rho(\cdot) = \lambda^2\rho(\cdot/\lambda)$. Then, the map $\lambda \mapsto \alpha^2(\lambda)$ is locally 1/2-Hölder continuous in the sense that*

$$\sup_{\lambda_{\min}\leq\lambda\leq\lambda_{\max}} \frac{|\alpha^2(\lambda) - \alpha^2(\tilde{\lambda})|}{|\lambda - \tilde{\lambda}|^{1/2}} \leq C_3(\lambda_{\min}, \lambda_{\max}, \rho, F_\epsilon, \gamma)$$

for all $\lambda_{\min}, \lambda_{\max} \in (0,\infty)$ with $\lambda_{\min} < \lambda_{\max}$.

Remark 2 *Similar result to Theorem 2 is known for the regularization parameter λ of the Lasso $\lambda\|\boldsymbol{\beta}\|_1$ (see Celentano et al. (2023); Miolane and Montanari (2021)). This line of work shows a suitable smoothness of the associated nonlinear system with respect to λ, using the implicit function theorem. We proceed differently here for the proof of Theorem 2: we show the map*

$$\lambda \in (0,\infty) \mapsto \hat{R}(\lambda) = \frac{\|\psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2}{\mathrm{tr}[\boldsymbol{V}_\lambda]^2}$$

is Hölder continuous. This Hölder continuity of $\hat{R}(\cdot)$ implies that of $\alpha^2(\cdot)$ thanks to the approximation $\alpha^2(\lambda) = \hat{R}(\lambda) + o_P(1)$ by Proposition 2. Hölder continuity of $\hat{R}(\cdot)$ follows if $\lambda \mapsto \|\psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2$ and $\lambda \mapsto \mathrm{tr}[\boldsymbol{V}_\lambda]^2$ are both Hölder continuous. The proof in Appendix C.2 shows the Hölder continuity of the map $\lambda \mapsto \|\psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2$ using Assumption 3 and the KKT conditions. For the map $\lambda \mapsto \mathrm{tr}[\boldsymbol{V}_\lambda]^2$, we show that $\mathrm{tr}[\boldsymbol{V}_\lambda]$ inherits the Hölder continuity from $\psi_\lambda(\boldsymbol{y} - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_\lambda)$, by leveraging the assumption that the noise distribution is sufficiently smooth; see Theorem 6 and the discussion around (12) on how this smoothness assumption is also used for results in Section 2.

Finally, we apply Theorem 2 to the adaptive tuning of the scale parameter λ. Let us take $\lambda_{\min}$ and $\lambda_{\max}$ such that $0 < \lambda_{\min} < \lambda_{\max}$, and take the closed interval $I = [\lambda_{\min}, \lambda_{\max}]$. For this fixed interval I, we consider a finite grid I_N of $(N+1)$ points equispaced in log-scale, i.e.,

$$I_N := \Big\{\lambda_{\min}\Big(\frac{\lambda_{\max}}{\lambda_{\min}}\Big)^{i/N} : i = 0,\ldots,N\Big\} \subset I = [\lambda_{\min}, \lambda_{\max}]. \tag{23}$$

for all $N \in \mathbb{N}$. We define the oracle optimal parameter $\lambda_{\text{opt}} \in I$ and the data-driven selected parameter $\hat{\lambda}_N \in I_N$ as

$$\lambda_{\text{opt}} \in \arg\min_{\lambda \in I} \alpha^2(\lambda), \quad \hat{\lambda}_N \in \arg\min_{\lambda \in I_N} \hat{R}(\lambda).$$

Note that λ_{opt} exists thanks to the continuity of $\lambda \mapsto \alpha^2(\lambda)$ from Theorem 2 and the compactness of I. Here, λ_{opt} is the optimal scale parameter in the closed interval I, while $\hat{\lambda}_N$ is the minimizer of the criterion $\hat{R}(\lambda)$ among the finite grid I_N of $(N+1)$ points. With the above notation, we claim that $\hat{\lambda}_N$ achieves the theoretically optimal risk limit $\alpha^2(\lambda_{\text{opt}}) = \min_{\lambda \in I} \alpha^2(\lambda)$.

Theorem 3 *Assume that F_ϵ satisfies Assumption 2, and ρ satisfies Assumption 1 and 3. Let R, $\hat{R}$, and α^2 be the maps defined by (19), (20), and (21), respectively. Then, for any closed interval $I = [\lambda_{\min}, \lambda_{\max}] \subset (0, \infty)$, there exists a sequence of integers $(N_n)_{n=1}^\infty = (N_n(\gamma, \rho, F_\epsilon, \lambda_{\min}, \lambda_{\max}))_{n=1}^\infty$ such that as $n, p \to +\infty$ with $p/n \to \gamma \in (0,1)$, we have*

$$R(\hat{\lambda}_{N_n}) = \alpha^2(\hat{\lambda}_{N_n}) + o_P(1) = \alpha^2(\lambda_{\text{opt}}) + o_P(1),$$

where $\lambda_{\text{opt}} \in \arg\min_{\lambda \in I} \alpha^2(\lambda)$ and $\hat{\lambda}_N \in \arg\min_{\lambda \in I_N} \hat{R}(\lambda)$ with I_N given by (23).

See Appendix C for the proof. Theorem 3 implies that if the grid I_N is fine enough, the M-estimator $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ with $\hat{\lambda}$ minimizing the criteria $\hat{R}(\lambda)$ over the grid achieves the theoretically optimal risk $\alpha^2(\lambda_{\text{opt}}) = \min_{\lambda \in I} \alpha^2(\lambda)$ in the fixed interval I of λ. We will verify Theorem 3 by numerical simulations in Section 4.

4 Numerical simulations

We focus on the Huber loss $\rho(x) = \int_0^{|x|} \min(1, t) dt$ and consider the adaptive tuning of the scale parameter λ for the scaled loss $\rho_\lambda(\cdot) = \lambda^2 \rho(\cdot/\lambda)$. Note that the scaled loss ρ_λ and its derivative $\psi_\lambda := \rho_\lambda'$ can be written explicitly as

$$\forall x \in \mathbb{R}, \quad \rho_\lambda(x) := \begin{cases} x^2/2 & |x| \le \lambda \\ \lambda|x| - \lambda^2/2 & |x| \ge \lambda \end{cases}, \quad \psi_\lambda(x) = \begin{cases} x & |x| \le \lambda \\ \lambda \operatorname{sign}(x) & |x| \ge \lambda \end{cases}. \tag{24}$$

Below, we use the notation in (19), (20), and (21); let $\hat{\boldsymbol{\beta}}_\lambda \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_\lambda(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ be the unregularized M-estimator computed by the scaled loss ρ_λ, and let $R(\lambda)$ and $\hat{R}(\lambda)$ be the out-of-sample error and our proposed estimate

$$R(\lambda) := \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_*)\|, \qquad \hat{R}(\lambda) := \frac{p \sum_{i=1}^n \psi_\lambda(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda)^2}{|\{i \in [n] : |y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda| \le \lambda\}|^2},$$

where we have used the simplification $\operatorname{tr}[\boldsymbol{V}_\lambda] = \sum_{i=1}^n \mathbf{1}\{|y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_\lambda| \le \lambda\}$ in (20). Let $\alpha^2(\lambda)$ be the solution to the nonlinear system (15) with $\rho = \rho_\lambda$.

With the above notation, we first verify

$$\hat{R}(\lambda) \approx R(\lambda) \approx \alpha^2(\lambda). \tag{25}$$

We set $(n, p) = (4000, 1200)$, $F_\epsilon = \text{t-dist}(\text{df} = 2)$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and $\boldsymbol{\beta}^\star = \boldsymbol{0}_p$. Once we generate $(\boldsymbol{y}, \boldsymbol{X})$, we compute $(R(\lambda), \hat{R}(\lambda))$ for each λ in a finite grid. We repeat the above procedure 100 times and plot $(R(\lambda), \hat{R}(\lambda))$ in Figure 3a, and the relative error $|\hat{R}(\lambda)/R(\lambda) - 1|$ in Figure 3b. We also plot $\alpha^2(\lambda)$ in Figure 3a by solving the nonlinear system of equations (15) (see Remark 3 for details). Figure 1a in the introduction features the same experiment as Figure 3a, without the theoretical curve $\lambda \to \alpha^2(\lambda)$. These figures are consistent with (25). Next, we conduct the adaptive tuning of the scale parameter $\lambda > 0$. Let us take $I = [1, 10]$ and the finite grid I_N as $\{\lambda_i = 10^{i/100} : i \in 0, 1, \dots, 100\} \subset I$. Then, Corollary 1 and Theorem 3 implies

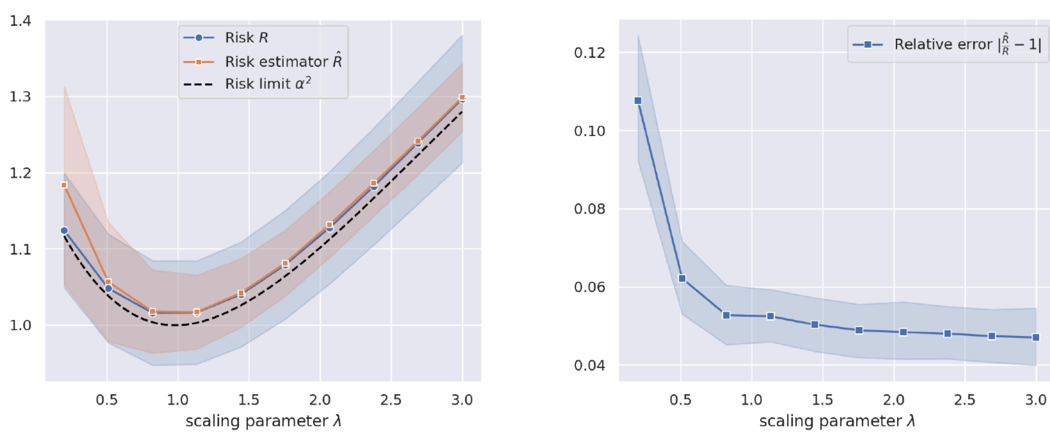
$$\min_{\lambda \in I_N} R(\lambda) \approx R(\hat{\lambda}) \approx \min_{\lambda \in I} \alpha^2(\lambda) \quad \text{where} \quad \hat{\lambda} \in \arg\min_{\lambda \in I_N} \hat{R}(\lambda). \tag{26}$$

Below, we verify (26) as we change the scale of noise distribution F_ϵ in the following way:

$$F_\epsilon := \sigma \cdot \text{t-dist}(\text{df} = 2) \text{ where } \sigma \in [1, 3].$$

For each σ in a finite grid over $[1, 3]$, we generate dataset $(\boldsymbol{X}, \boldsymbol{y})$ and calculate $R(\hat{\lambda}_N)$ and $\min_{\lambda \in I_N} R(\lambda)$. We repeat the above procedure 100 times and plot $R(\hat{\lambda})$ and $\min_{\lambda \in I} R(\lambda)$ in Figure 4a. We also plot $\min_{\lambda \in I} \alpha(\lambda)^2$ in the same figure. Figure 1b in the introduction has the same experiment as Figure 4a, without the theoretical limit.

In Figure 4b, we plot the two relative errors: $|\frac{R(\hat{\lambda}_N)}{\min_{\lambda \in I_N} R(\lambda)} - 1|$ and $|\frac{R(\hat{\lambda}_N)}{\min_{\lambda \in I} \alpha^2(\lambda)} - 1|$. These Figures are consistent with (26).



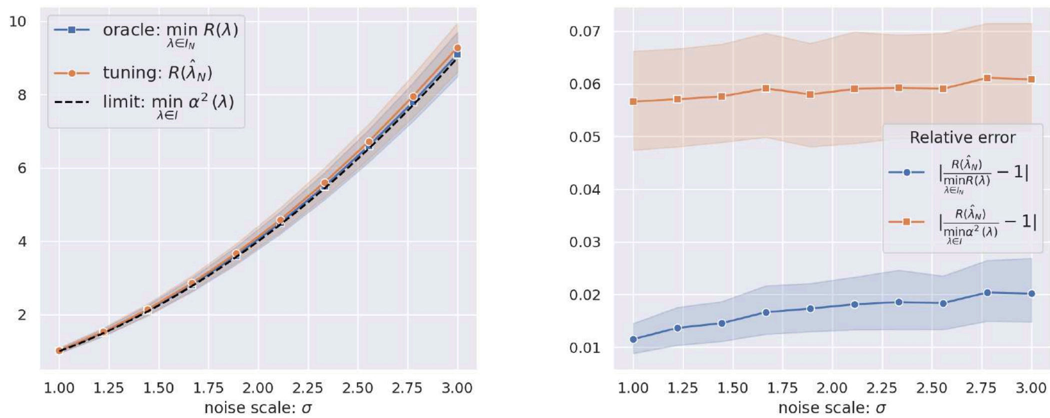
(a) $R(\lambda)$, $\hat{R}(\lambda)$ and $\alpha^2(\lambda)$.

(b) Relative error $|\hat{R}(\lambda)/R(\lambda) - 1|$

Figure 3: Plot of the out-of-sample error $R(\lambda)$ and estimator $\hat{R}(\lambda)$ over 100 repetitions, with $n = 4000$, $p = 1200$, for the Huber loss for different values of the scale parameters λ. The noise distribution is $\text{t-dist}(\text{df} = 2)$. $\alpha^2(\lambda)$ is the solution to the nonlinear system (15).

5 Outline of the proof

In this section, we give a sketch of proof of Theorem 1. See Appendix B for rigorous proofs. Let $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n \times p}$ so that $\boldsymbol{G}$ has i.i.d $\mathcal{N}(0, 1)$ entries, and we denote by $\boldsymbol{g}_i$ the i-th



(a) Adaptive tuning vs oracles (b) Relative error

Figure 4: Adaptive tuning with the scale of noise σ changing as $F_\epsilon = \sigma \cdot \text{t-dist}(\text{df} = 2)$. Here, $I = [1, 10]$, $n = 4000$, $p = 1200$, and I_N is the uniform grid in log-scale of length 101. $R(\hat{\lambda}_N)$ is the out-of-sample error with λ selected by $\hat{R}$, $\min_{\lambda \in I_N} R(\lambda)$ is the optimal out-of-sample error among I_N, and $\min_{\lambda \in I} \alpha^2(\lambda)$ is the theoretically optimal risk limit. We repeat 100 times.

row of $\boldsymbol{G}$. By the change of variable $\boldsymbol{\beta} \mapsto \boldsymbol{h} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)$, we write the M-estimator $\hat{\boldsymbol{\beta}}$ in (1) by $\hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{h}} + \boldsymbol{\beta}^\star$, where $\hat{\boldsymbol{h}}$ is defined as

$$\hat{\boldsymbol{h}}(\boldsymbol{\epsilon}, \boldsymbol{G}) \in \operatorname*{arg\,min}_{\boldsymbol{h} \in \mathbb{R}^p} \frac{1}{n} \sum_i \rho(y_i - \boldsymbol{x}_i^\top (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h} + \boldsymbol{\beta}^\star)) = \operatorname*{arg\,min}_{\boldsymbol{h} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - \boldsymbol{g}_i^\top \boldsymbol{h}). \tag{27}$$

Throughout, we regard $\hat{\boldsymbol{h}}$ as a function of $(\boldsymbol{\epsilon}, \boldsymbol{G})$. Then, quantities of interest such as $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star)\|$, $\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, and $\boldsymbol{V} = (\partial/\partial \boldsymbol{y})\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ can be expressed as functions of $(\boldsymbol{\epsilon}, \boldsymbol{G})$:

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\| = \|\hat{\boldsymbol{h}}\|, \quad \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}) \in \mathbb{R}^n, \quad \boldsymbol{V} = (\partial/\partial \boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}) \in \mathbb{R}^{n \times n}.$$

With the above notation, the statement of Theorem 1 is simplified to

$$\|\hat{\boldsymbol{h}}\|^2 = \frac{p}{\operatorname{tr}[\boldsymbol{V}]^2} \|\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}})\|^2 + o_P(1). \tag{28}$$

To prove (28), we consider the ridge-regularized M-estimator $\hat{\boldsymbol{h}}_\mu$

$$\hat{\boldsymbol{h}}_\mu(\boldsymbol{\epsilon}, \boldsymbol{G}) := \operatorname*{arg\,min}_{\boldsymbol{h} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \frac{\mu}{2}\|\boldsymbol{h}\|^2. \tag{29}$$

for a diminishing regularization parameter $\mu = \mu_n \to 0$. Note that $\hat{\boldsymbol{h}}_\mu$ coincides with $\hat{\boldsymbol{h}}$ defined by (27) when $\mu = 0$. The benefit of considering $\hat{\boldsymbol{h}}_\mu$ is its rich differentiability structures that will be explained in Section 5.1 ahead. Such derivative structure is helpful when we use probabilistic tools that involve derivatives, e.g., Stein's formula (Stein, 1981; Bellec and Zhang, 2021), the Gaussian Poincaré inequality (Boucheron et al., 2013, Theorem 3.20), and normal approximations (Chatterjee, 2009; Bellec and Zhang, 2023), etc.

5.1 Differentiability of M-estimators

In this section, we introduce a useful differentiable structure of the regularized M-estimator $\hat{\boldsymbol{h}}_\mu$ defined in (29). To begin with, we discuss the differentiability of $\psi(\boldsymbol{\epsilon}-\boldsymbol{G}\hat{\boldsymbol{h}}_\mu)$ with respect to $\boldsymbol{\epsilon}$.

Proposition 3 (Proposition 4.1 in Bellec (2023)) *Suppose that ρ is convex and differentiable and that $\psi = \rho'$ is 1-Lipschitz. Then, for any M-estimator of the form $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho(\boldsymbol{y}_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + g(\boldsymbol{\beta})$ where $g : \mathbb{R}^p \to \mathbb{R}$ is a convex penalty, the map $\boldsymbol{y} \in \mathbb{R}^n \mapsto \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \mathbb{R}^n$ is 1-Lipschitz for almost every $\boldsymbol{X} \in \mathbb{R}^{n\times p}$.*

Since $\psi = \rho'$ is 1-Lipschitz, Proposition 3 implies that the map $\boldsymbol{\epsilon} \mapsto \psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu(\boldsymbol{\epsilon}, \boldsymbol{G}))$ is 1-Lipschitz for all $\mu \geq 0$. As a consequence, Rademacher's theorem gives the following:

$$\forall \mu \geq 0, \quad \boldsymbol{V}_\mu := (\partial/\partial\boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu) \in \mathbb{R}^{n\times n} \text{ exists almost everywhere and } \|\boldsymbol{V}_\mu\|_{op} \leq 1. \tag{30}$$

We emphasize that (30) also holds for $\mu = 0$, so the original Jacobian matrix $\boldsymbol{V} = (\partial/\partial\boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}})$ is also well-defined and $\|\boldsymbol{V}\|_{op} \leq 1$.

The next theorem introduces a detailed differentiable structure of $\hat{\boldsymbol{h}}_\mu$ with respect to $\boldsymbol{G}$ when μ is strictly positive.

Theorem 4 (Theorem 1 in Bellec and Shen (2022)) *Suppose $\mu > 0$. Then, for almost every $(\boldsymbol{\epsilon}, \boldsymbol{G})$, the map $(\boldsymbol{\epsilon}, \boldsymbol{G}) \mapsto \hat{\boldsymbol{h}}_\mu(\boldsymbol{\epsilon}, \boldsymbol{G})$ defined by (29) is differentiable at almost every $(\boldsymbol{\epsilon}, \boldsymbol{G})$, and the derivatives are given by*

$$\forall i \in [n], \ \frac{\partial \hat{\boldsymbol{h}}_\mu}{\partial \epsilon_i}(\boldsymbol{\epsilon}, \boldsymbol{G}) = \hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{e}_i \psi'(r_i),$$
$$\forall i \in [n], \forall j \in [p], \ \frac{\partial \hat{\boldsymbol{h}}_\mu}{\partial g_{ij}}(\boldsymbol{\epsilon}, \boldsymbol{G}) = \hat{\boldsymbol{A}}_\mu \boldsymbol{e}_j \psi(r_i) - \hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{e}_i \psi'(r_i) \boldsymbol{e}_j^\top \hat{\boldsymbol{h}}_\mu,$$

where $\hat{\boldsymbol{A}}_\mu = (\boldsymbol{G}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G} + n\mu I_p)^{-1} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{r} = \boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu \in \mathbb{R}^n$.

See Bellec and Shen (2022) for the differentiability of regularized M-estimators for general strongly convex penalties. Thanks to the strict positiveness of μ, the matrix $\hat{A}_\mu$ is always well-defined and its operator norm is bounded from above by $(n\mu)^{-1}$. By Theorem 4 and the chain rule, we find that $\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)$ is also differentiable at $(\boldsymbol{\epsilon}, \boldsymbol{G})$ when $\mu > 0$, with the derivatives given by

$$(\partial/\partial g_{ij})\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu) = -\operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G}\hat{\boldsymbol{A}}_\mu \boldsymbol{e}_j \psi(r_i) - \boldsymbol{V}_\mu \boldsymbol{e}_i \boldsymbol{e}_j^\top \hat{\boldsymbol{h}}_\mu \in \mathbb{R}^n,$$
$$\boldsymbol{V}_\mu = (\partial/\partial\boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu) = \operatorname{diag}\{\psi'(\boldsymbol{r})\} - \operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G}\hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\} \in \mathbb{R}^{n\times n}.$$

Note that the derivative formula in Theorem 4 is only guaranteed for the smoothed M-estimator $\hat{\boldsymbol{h}}_\mu$ with a positive ridge parameter μ, and not for the original M-estimator $\hat{\boldsymbol{h}}$ ($\mu = 0$). This is the benefit of considering the smoothed M-estimator $\hat{\boldsymbol{h}}_\mu$.

Next, we will use those derivative formulae to prove the interplay between $\|\hat{\boldsymbol{h}}_\mu\|^2$, $\operatorname{tr}[\boldsymbol{V}_\mu]$, and $\|\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)\|^2$.

5.2 Consistency of the risk estimate for the smoothed M-estimator

We have discussed the differentiable structure of the regularized M-estimator $\hat{\boldsymbol{h}}_\mu$. In this section, we derive the key relationship (31) below between $\|\hat{\boldsymbol{h}}_\mu\|^2$, $\mathrm{tr}[\boldsymbol{V}_\mu]$, and $\|\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)\|^2$, which can be seen as an extension of (8) to M-estimators with a Ridge penalty function that vanishes polynomially in n as $n \to +\infty$.

Lemma 1 *Let Assumption 1 and 2 be fulfilled. Let $\hat{\boldsymbol{h}}_\mu$ be the regularized M-estimator with a ridge penalty $\mu\|\boldsymbol{h}\|^2/2$ defined in (29). If $\mu = n^{-c}$ for some $c \in (0, 1/4]$, we have*

$$\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 - \frac{\mathrm{tr}[\boldsymbol{V}_\mu]^2}{n}\|\hat{\boldsymbol{h}}_\mu\|^2 = O_P(n^{1-c}), \tag{31}$$

where $\boldsymbol{\psi}_\mu$ is the vector $\psi(\boldsymbol{\epsilon}-\boldsymbol{G}\hat{\boldsymbol{h}}_\mu) \in \mathbb{R}^n$ and $\boldsymbol{V}_\mu$ is the Jacobian matrix $(\partial/\partial\boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon}-\boldsymbol{G}\hat{\boldsymbol{h}}_\mu) \in \mathbb{R}^{n\times n}$.

The proof is given in Appendix B.1. It is based on the following moment inequality.

Theorem 5 (Theorem 7.2 in Bellec (2023)) *Assume that $\boldsymbol{G} = (g_{ij})_{ij} \in \mathbb{R}^{n\times p}$ has i.i.d $\mathcal{N}(0,1)$ entries, and that $\boldsymbol{f} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ is a differentiable function with $\|\boldsymbol{f}\|^2 \le 1$. Then, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\Big[\Big|p\|\boldsymbol{f}\|^2 - \sum_{j=1}^p(\boldsymbol{f}^\top\boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n\frac{\partial f_i}{\partial g_{ij}})^2\Big|\Big] \le C\Big[\sqrt{p}\,\mathbb{E}[(1 + \sum_{ij}\|\frac{\partial\boldsymbol{f}}{\partial g_{ij}}\|^2)^{1/2}] + \mathbb{E}[\sum_{ij}\|\frac{\partial\boldsymbol{f}}{\partial g_{ij}}\|^2]\Big] \tag{32}$$

Above, we have used $\sum_{ij} = \sum_{i=1}^n\sum_{j=1}^p$ for brevity. The proof is based on the Gaussian Poincaré inequality (cf. (Boucheron et al., 2013, Theorem 3.20)) and the second order Stein's formula (see Bellec and Zhang (2021) or Theorem 6).

Below, we describe a sketch of proof of Lemma 1. To begin with, we take $\boldsymbol{f} = \boldsymbol{\psi}_\mu/(n^{1/2}\|\psi\|_\infty)$ in Theorem 5 so that $\|\boldsymbol{f}\|^2 \le 1$. By the derivative formula in Theorem 4 with $\|\hat{\boldsymbol{A}}_\mu\|_{op} = \|(\boldsymbol{G}^\top\mathrm{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G} + n\mu I_p)^{-1}\|_{op} \le (n\mu)^{-1} = n^{-1+c}$, we can show that the right-hand side of (32) is negligible compared to the left-hand side. As a consequence, we have

$$\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 \approx \frac{1}{n}\sum_{j=1}^p\Big(\boldsymbol{\psi}_\mu\boldsymbol{G}^\top\boldsymbol{e}_j - \sum_{i=1}^n\boldsymbol{e}_i^\top\frac{\partial\boldsymbol{\psi}_\mu}{\partial g_{ij}}\Big)^2.$$

Using Theorem 4 and the KKT condition $\boldsymbol{G}^\top\boldsymbol{\psi}_\mu = n\mu\hat{\boldsymbol{h}}_\mu$, the terms inside the square can be written explicitly by

$$\boldsymbol{\psi}_\mu\boldsymbol{G}^\top\boldsymbol{e}_j - \sum_{i=1}^n\boldsymbol{e}_i^\top\frac{\partial\boldsymbol{\psi}_\mu}{\partial g_{ij}} = \Big(n\mu\hat{\boldsymbol{h}}_\mu^\top + \boldsymbol{\psi}_\mu^\top\mathrm{diag}\,\psi'(\boldsymbol{r})\boldsymbol{G}\hat{\boldsymbol{A}}_\mu + \mathrm{tr}(\boldsymbol{V}_\mu)\hat{\boldsymbol{h}}_\mu^\top\Big)\boldsymbol{e}_j,$$

for all $j \in [p]$, where $\hat{\boldsymbol{A}}_\mu = (\boldsymbol{G}^\top\mathrm{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G}+n\mu I_p)^{-1}$. Combining the above two displays, we obtain

$$\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 \approx \frac{1}{n}\Big\|n\mu\hat{\boldsymbol{h}}_\mu + \hat{\boldsymbol{A}}_\mu^\top\boldsymbol{G}^\top\mathrm{diag}\,\psi'(\boldsymbol{r})\boldsymbol{\psi}_\mu + \mathrm{tr}(\boldsymbol{V}_\mu)\hat{\boldsymbol{h}}_\mu\Big\|^2.$$

Thanks to $\mu = n^{-c}$ and $\|\hat{\boldsymbol{A}}_\mu\|_{op} \le (n\mu)^{-1}$, $\|n\mu\hat{\boldsymbol{h}}_\mu\|$ and $\|\hat{\boldsymbol{A}}_\mu^\top\boldsymbol{G}^\top\mathrm{diag}\,\psi'(\boldsymbol{r})\boldsymbol{\psi}_\mu\|$ are negligible compared to $\mathrm{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu$, so that we obtain $p/n\cdot\|\boldsymbol{\psi}_\mu\|^2 \approx n^{-1}\|\mathrm{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2$ as desired.

5.3 Back to the original M-estimator

Finally, we derive the target (28) from Lemma 1. Toward that, we will show that the quantities of smoothed M-estimator, i.e., $(\hat{\boldsymbol{h}}_\mu, \boldsymbol{\psi}_\mu, \operatorname{tr}[\boldsymbol{V}_\mu])$, are close to $(\hat{\boldsymbol{h}}, \boldsymbol{\psi}, \operatorname{tr}[\boldsymbol{V}])$ under $\mu = \mu_n = n^{-c}$ for some $c \in (0, 1/4]$. More precisely, we prove the following:

$$
\begin{aligned}
&\|\hat{\boldsymbol{h}}_\mu - \hat{\boldsymbol{h}}\|^2 = o_P(1), && \|\hat{\boldsymbol{h}}\|^2 = O_P(1), && (33)\\
&\|\boldsymbol{\psi}_\mu\|^2 - \|\boldsymbol{\psi}\|^2 = o_P(n^{1-\frac{c}{2}}), && \|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|_2 = O_P(n^{\frac{1-c}{2}}), && (34)\\
&\operatorname{tr}[\boldsymbol{V}_\mu]^2 - \operatorname{tr}[\boldsymbol{V}]^2 = o_P(n^{2-\frac{c}{2}}). && && (35)
\end{aligned}
$$

Now we assume that the above displays are justified. Then, combined with Lemma 1, we have

$$
\begin{aligned}
\frac{p}{n^2}\|\boldsymbol{\psi}\|^2 - \frac{\operatorname{tr}[\boldsymbol{V}]^2}{n^2}\|\hat{\boldsymbol{h}}\|^2 &= \left(\frac{p}{n^2}\|\boldsymbol{\psi}_\mu\|^2 - \frac{\operatorname{tr}[\boldsymbol{V}_\mu]^2}{n^2}\|\hat{\boldsymbol{h}}\|^2\right) + \frac{p}{n^2}(\|\boldsymbol{\psi}|_2^2 - \|\boldsymbol{\psi}_\mu\|^2)\\
&\quad + \frac{\operatorname{tr}[\boldsymbol{V}_\mu]^2 - \operatorname{tr}[\boldsymbol{V}]^2}{n^2}\|\hat{\boldsymbol{h}}_\mu\|^2 - \frac{\operatorname{tr}[\boldsymbol{V}]^2}{n^2}(\|\hat{\boldsymbol{h}}\|^2 - \|\hat{\boldsymbol{h}}_\mu\|^2)\\
&= O_P(n^{-c}) + o_P(n^{-\frac{c}{2}}) + o_P(n^{-\frac{c}{2}}) - \frac{\operatorname{tr}[\boldsymbol{V}]^2}{n^2}(\|\hat{\boldsymbol{h}}\|^2 - \|\hat{\boldsymbol{h}}_\mu\|^2)\\
&= o_P(n^{-\frac{c}{2}}) - \frac{\operatorname{tr}[\boldsymbol{V}]^2}{n^2} o_P(1).
\end{aligned}
$$

Furthermore, we will argue that $n^2/\operatorname{tr}[\boldsymbol{V}]^2 = O_P(1)$, i.e., $\operatorname{tr}(\boldsymbol{V})/n$ is not too small (see Lemma 9). Thus, dividing by $\operatorname{tr}[\boldsymbol{V}]^2/n^2$ on the above display, we obtain the target (28).

Below, we describe a sketch of proof of (33), (34), and (35). For (33), we will show that $\|\hat{\boldsymbol{h}}\|^2$ and $\|\hat{\boldsymbol{h}}_\mu\|^2$ converge to the same finite limit by the main theorem in Thrampoulidis et al. (2018) and Theorem 7. (34) easily follows from the two KKT conditions and the Lipschitz property of ψ. The most technical part is (35). Thanks to Assumption 2, the noise vector $\boldsymbol{\epsilon}$ can be represented as

$$
\boldsymbol{\epsilon} = \boldsymbol{z} + \boldsymbol{\delta}, \quad \boldsymbol{z} \perp\!\!\!\perp \boldsymbol{\delta}, \quad (z_i)_{i=1}^n \overset{\text{iid}}{\sim} F, \quad (\delta_i)_{i=1}^n \overset{\text{iid}}{\sim} \tilde{F},
$$

where F has the density $z \mapsto \exp(-\phi(z))$. By the chain rule, $\boldsymbol{V} - \boldsymbol{V}_\mu$ can be written as

$$
\boldsymbol{V} - \boldsymbol{V}_\mu = \frac{\partial}{\partial \boldsymbol{\epsilon}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}} \cdot \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \boldsymbol{I}_n \cdot \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu). \tag{36}
$$

Here, by the second order Stein's formula extended to the density of the form $\exp(-\phi(z))$, we will show that $\operatorname{tr}[\boldsymbol{V}] - \operatorname{tr}[\boldsymbol{V}_\mu]$ concentrates around $\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu)$.

Theorem 6 (Equation (2.15) in Bellec and Zhang (2021)) *Assume that the random vector $\boldsymbol{z} \in \mathbb{R}^n$ has density $\boldsymbol{z} \mapsto \exp(-\phi(\boldsymbol{z}))$ where $\phi : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable function with Hessian $H(\boldsymbol{z}) = (\partial/\partial\boldsymbol{z})^2\phi(\boldsymbol{z}) \in \mathbb{R}^{n\times n}$. Then, for any differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, we have*

$$
\mathbb{E}\Big[\Big(\frac{\partial\phi(\boldsymbol{z})}{\partial\boldsymbol{z}}\boldsymbol{f}(\boldsymbol{z}) - \operatorname{tr}[\frac{\partial\boldsymbol{f}(\boldsymbol{z})}{\partial\boldsymbol{z}}]\Big)^2\Big] = \mathbb{E}\Big[\boldsymbol{f}(\boldsymbol{z})^\top\boldsymbol{H}(\boldsymbol{z})\boldsymbol{f}(\boldsymbol{z}) + \operatorname{tr}[(\frac{\partial\boldsymbol{f}(\boldsymbol{z})}{\partial\boldsymbol{z}})^2]\Big],
$$

provided that the expectation on the right-hand side exists.

Applying Theorem 6 with $\phi(\boldsymbol{z}) = \sum_{i=1}^n \phi(z_i)$ and $\boldsymbol{f} = \boldsymbol{\psi} - \boldsymbol{\psi}_\mu$, using the identity (36) and $\boldsymbol{H}(\boldsymbol{z}) = \operatorname{diag}\{\phi''(z_i) : i \in [n]\}$, we have

$$\mathbb{E}[(\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi}) - \operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu])^2] = \mathbb{E}\Big[\sum_{i=1}^n \phi''(z_i)(\psi_i - (\psi_\mu)_i)^2 + \operatorname{tr}[(\boldsymbol{V} - \boldsymbol{V}_\mu)^2]\Big],$$

and we can easily show that the right-hand side is $O(n)$. Hence, by the Markov inequality, we have

$$\operatorname{tr}[\boldsymbol{V}_\mu] - \operatorname{tr}[\boldsymbol{V}] = \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi}) + O_P(n^{1/2}).$$

By the Cauchy–Schwarz inequality, we bound $|\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi})|$ from above as

$$|\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi})| \le \|\phi'(\boldsymbol{z})\|\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\| \overset{(*)}{=} O_P(n^{1/2}) \cdot O_P(n^{\frac{1-c}{2}}) = O_P(n^{1-\frac{c}{2}}),$$

where $(*)$ follows from $\|\phi'(\boldsymbol{z})\| = O_P(n^{1/2})$ (see Lemma 8) and $\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\| = O_P(n^{1/2-c})$ by (34). As a consequence, we obtain

$$\operatorname{tr}[\boldsymbol{V}_\mu] - \operatorname{tr}[\boldsymbol{V}] = O_P(n^{1-\frac{c}{2}}) + O_P(n^{1/2}) = O_P(n^{1-\frac{c}{2}})$$

thanks to $c \le 1$. Finally, $\|\boldsymbol{V}\|_{op}$ and $\|\boldsymbol{V}_\mu\|_{op} \le 1$ lead to

$$|\operatorname{tr}[\boldsymbol{V}_\mu]^2 - \operatorname{tr}[\boldsymbol{V}]^2| \le |\operatorname{tr}[\boldsymbol{V}_\mu] + \operatorname{tr}[\boldsymbol{V}]||\operatorname{tr}[\boldsymbol{V}_\mu] - \operatorname{tr}[\boldsymbol{V}]| \le 2nO_P(n^{1-\frac{c}{2}}) = O_P(n^{2-\frac{c}{2}}),$$

which is exactly (35).

Acknowledgments and Disclosure of Funding

P. C. Bellec's research was partially supported by the National Science Foundation award DMS-1945428.

References

Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36): 14563–14568, 2013.

Pierre C Bellec. Out-of-sample error estimation for M-estimators with convex penalty. *Information and Inference: A Journal of the IMA*, 12(4):2782–2817, 2023.

Pierre C Bellec. Observable adjustments in single-index models for regularized M-estimators. *The Annals of Statistics, forthcoming*, 2025.

Pierre C. Bellec and Takuya Koriyama. Existence of solutions to the nonlinear equations characterizing the precise error of M-estimators. *arXiv preprint arXiv:2312.13254*, 2023.

Pierre C Bellec and Yiwei Shen. Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*, pages 1912–1947. PMLR, 2022.

Pierre C Bellec and Cun-Hui Zhang. Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903, 2021.

Pierre C Bellec and Cun-Hui Zhang. Debiasing convex regularized estimators and interval estimation in linear models. *The Annals of Statistics*, 51(2):391–436, 2023.

Pierre C Bellec, Yiwei Shen, and Cun-Hui Zhang. Asymptotic normality of robust M-estimators with convex penalty. *Electronic Journal of Statistics*, 16(2):5591–5622, 2022.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics*, 50(1):170–196, 2022.

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

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

David Donoho and Andrea Montanari. High Dimensional Robust M-estimation: Asymptotic Variance via Approximate Message Passing. *Probability Theory and Related Fields*, 166: 935–969, 2016.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 2018.

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

Noureddine El El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Sophie Lambert-Lacroix and Laurent Zwald. Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5(none):1015 – 1053, 2011. doi: 10.1214/11-EJS635. URL https://doi.org/10.1214/11-EJS635.

Po-Ling Loh. Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics*, 15(2):5933–5994, 2021.

Po-Ling Loh. A theoretical review of modern robust statistics. *Annual Review of Statistics and Its Application*, 12, 2024.

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

Bruno Loureiro, Cédric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*, pages 14283–14314. PMLR, 2022.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

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

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

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free Huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

SUPPLEMENT

Appendix A. Existence and uniqueness of solutions to (15)

The following theorem from our companion paper Bellec and Koriyama (2023) guarantees the uniqueness and the existence of the solution (α, κ) to (15), with an explicit upper bound of α. We include it here for convenience.

Theorem 7 (Section 2 Bellec and Koriyama (2023)) *Suppose that $\gamma \in (0,1)$ and (ρ, F_ϵ) satisfy the conditions*

1. *ρ is convex, Lipschitz, and $\{0\} = \arg\min_{x\in\mathbb{R}} \rho(x)$.*
2. *$\mathbb{P}_{W\sim F_\epsilon}(W \neq 0) > 0$ and $\inf_{\lambda>0} \mathbb{E}_{W\sim F_\epsilon}[\mathrm{dist}(G, \lambda\partial\rho(W))^2] > 1-\gamma$.*

Then, the nonlinear system of equations (15) has a unique solution $(\alpha,\kappa) \in \mathbb{R}^2_{>0}$, and this α is bounded from above as

$$\alpha \leq \frac{Q_{F_\epsilon}(r^2 c_\gamma)}{r c_\gamma} + \frac{b}{c_\gamma} \tag{37}$$

where c_γ is a positive constant depending on γ only, $Q_{F_\epsilon}(x) := \inf\{q > 0 : \mathbb{P}_{W\sim F_\epsilon}(|W| \geq q) \leq x\}$, and $r \in (0,1]$ and $b \geq 0$ are any constant such that the coercivity condition

$$\|\rho_\lambda\|_{\mathrm{lip}}^{-1}(\rho(x) - \rho(0)) \geq r(|x| - b) \quad \text{for all } x \in \mathbb{R}$$

is satisfied.

The upper bound (37) is helpful for Section 3, in which we will consider the scaled loss $\rho_\lambda : \mathbb{R} \to \mathbb{R}$, $x \mapsto \lambda^2\rho(x/\lambda)$ for some scale parameter $\lambda > 0$. Let $\alpha(\lambda)$ be the solution to the nonlinear system (15) with $\rho = \rho_\lambda$. Then, (37) with $\rho = \rho_\lambda$ gives the following explicit upper bound of $\alpha(\lambda)$ with respect to $\lambda > 0$:

$$\forall \lambda > 0, \quad \alpha^2(\lambda) \leq C_4(\rho, F_\epsilon, \gamma)(\lambda^2 + 1).$$

See Proposition 2 and its proof for its derivation. We observe that the upper bound explodes as $\lambda \to +\infty$, which reflects the fact that the out-of-sample error of the ordinary least square (OLS) estimator is unbounded when the noise distribution has no second moment.

Remark 3 *Suppose that the loss ρ is the Huber loss ρ_λ, parametrized by a scale parameter $\lambda > 0$. This loss function was introduced in (24) and used for numerical simulations in Section 4 and Appendix D. For convenience, we recall its definition:*

$$\rho_\lambda(x) := \begin{cases} x^2/2 & |x| \leq \lambda \\ \lambda|x| - \lambda^2/2 & |x| > \lambda \end{cases}$$

In this case, the proximal operator of $\rho = \rho_\lambda$ takes a closed form and satisfies

$$x - \mathrm{prox}[\kappa\rho_\lambda](x) = \kappa\lambda\zeta\Big(\frac{x}{\lambda(1+\kappa)}\Big) \quad \text{where} \quad \zeta(u) := \begin{cases} u & |u| \leq 1 \\ \mathrm{sign}(u) & |u| > 1 \end{cases}$$

for all $\kappa > 0$ and $x \in \mathbb{R}$. Thus, the nonlinear system (15) with $\rho = \rho_\lambda$ can be written as

$$\alpha^2\gamma - \kappa^2\lambda^2\,\mathbb{E}\Big[\zeta\Big(\frac{aG+W}{\lambda(1+\kappa)}\Big)^2\Big] = 0, \quad \alpha\gamma - \kappa\lambda\,\mathbb{E}\Big[\zeta\Big(\frac{aG+W}{\lambda(1+\kappa)}\Big)\cdot Z\Big] = 0$$

with positive unknown (α, κ). In the numerical simulations presented in Section 4 and Appendix D, this system was solved using the solver scipy.optimize.fsolve *from scipy (Virtanen et al., 2020).*

Appendix B. Proof of Theorem 1

B.1 Proof of Lemma 1

The claim of Lemma 1 is recalled here for convenience:

$$\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 - \frac{\mathrm{tr}[\boldsymbol{V}_\mu]^2}{n}\|\hat{\boldsymbol{h}}_\mu\|^2 = O_P(n^{1-c}) \text{ under } \mu = n^{-c} \text{ for all } c \in (0, 1/4].$$

To prove this, we use a variant of Theorem 5:

Lemma 2 (Theorem 7.4 in Bellec and Shen (2022)) *Let $h : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ and $\psi : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ be locally Lipschitz functions. Suppose $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has i.i.d $\mathcal{N}(0,1)$ entries, and we denote $h(\boldsymbol{G})$ by $\boldsymbol{h}$ and $\psi(\boldsymbol{G})$ by $\boldsymbol{\psi}$. Then, we have*

$$\mathbb{E}\Big[\frac{|\frac{p}{n}\|\boldsymbol{\psi}\|^2 - \frac{1}{n}\sum_{j=1}^p(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n\frac{\partial\psi_i}{\partial g_{ij}})^2|}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big] \le C(\sqrt{n+p}(1+\Xi^{1/2}) + \Xi), \tag{38}$$

where $\Xi = \mathbb{E}[\frac{1}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}\sum_{i=1}^n\sum_{j=1}^p(\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|^2)]$

Lemma 2 is obtained from Theorem 5 with $\boldsymbol{f} = n^{-1/2}\boldsymbol{\psi}/(\|\boldsymbol{\psi}\|/n + \|\boldsymbol{h}\|)$. Below, we prove Lemma 1 using Lemma 2 with $(\boldsymbol{\psi}, \boldsymbol{h}) = (\boldsymbol{\psi}_\mu, \boldsymbol{h}_\mu) = (\psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu), \hat{\boldsymbol{h}}_\mu)$. First, we bound Ξ in (38). By the derivative formula in Theorem 4, the derivatives inside the expectation can be written as

$$\sum_{i=1}^n\sum_{j=1}^p\|\frac{\partial\hat{\boldsymbol{h}}_\mu}{\partial g_{ij}}\|^2 = \sum_{i=1}^n\sum_{j=1}^p\|\hat{\boldsymbol{A}}_\mu\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu - \hat{\boldsymbol{A}}_\mu\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{e}_i\boldsymbol{e}_j^\top\hat{\boldsymbol{h}}_\mu\|^2,$$
$$\sum_{i=1}^n\sum_{j=1}^p\|\frac{\partial\hat{\boldsymbol{\psi}}_\mu}{\partial g_{ij}}\|^2 = \sum_{i=1}^n\sum_{j=1}^p\| - \boldsymbol{D}_\mu\boldsymbol{G}\hat{\boldsymbol{A}}_\mu\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu - \boldsymbol{V}_\mu\boldsymbol{e}_i\boldsymbol{e}_j^\top\hat{\boldsymbol{h}}_\mu\|^2,$$

where $\boldsymbol{D}_\mu = \mathrm{diag}\{\psi'(\boldsymbol{\epsilon}-\boldsymbol{G}\hat{\boldsymbol{h}}_\mu)\}$, $\hat{\boldsymbol{A}}_\mu = (\boldsymbol{G}^\top\mathrm{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G}+n\mu I_p)^{-1}$, and $\boldsymbol{V}_\mu = (\partial/\partial\boldsymbol{\epsilon})\psi(\boldsymbol{\epsilon}-\boldsymbol{G}\hat{\boldsymbol{h}}_\mu)$. Note that the operator norm of these matrices is bounded from above as

$\|\hat{\boldsymbol{A}}_\mu\|_{op} \le (n\mu)^{-1}$	by $\psi'(r_i) \ge 0$ for all $i \in [n]$
$\|\boldsymbol{D}_\mu\|_{op} \le 1$	by $\|\psi\|_{\mathrm{lip}} = 1$
$\|\boldsymbol{V}_\mu\|_{op} \le 1$	by (30)

Then, $\frac{1}{2}\sum_{ij}\|\frac{\partial \hat{\boldsymbol{h}}_\mu}{\partial g_{ij}}\|^2$ can be bounded from above as

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^p\|\frac{\partial \hat{\boldsymbol{h}}_\mu}{\partial g_{ij}}\|^2 &\le \sum_{ij}\|\hat{\boldsymbol{A}}_\mu \boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu\|^2 + \|\hat{\boldsymbol{A}}_\mu\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{e}_i\boldsymbol{e}_j^\top\hat{\boldsymbol{h}}_\mu\|^2 && \text{by } (a+b)^2\le 2(a^2+b^2)\\
&= \|\hat{\boldsymbol{A}}_\mu\|_F^2\|\boldsymbol{\psi}_\mu\|^2 + \|\hat{\boldsymbol{A}}_\mu\boldsymbol{G}^\top\boldsymbol{D}_\mu\|_F^2\|\hat{\boldsymbol{h}}_\mu\|^2 && \text{by } \textstyle\sum_i\|\boldsymbol{M}e_i\|^2=\|\boldsymbol{M}\|_F^2\\
&\le \|\hat{\boldsymbol{A}}_\mu\|_F^2(\|\boldsymbol{\psi}_\mu\|^2 + \|\boldsymbol{G}^\top\boldsymbol{D}_\mu\|_{op}^2\|\hat{\boldsymbol{h}}_\mu\|^2) && \text{by } \|\hat{\boldsymbol{A}}_\mu\boldsymbol{G}^\top\boldsymbol{D}_\mu\|_F^2\le\|\hat{\boldsymbol{A}}_\mu\|_F^2\|\boldsymbol{G}^\top\boldsymbol{D}_\mu\|_{op}^2\\
&\le \|\hat{\boldsymbol{A}}_\mu\|_F^2(\|\boldsymbol{\psi}_\mu\|^2 + \|\boldsymbol{G}\|_{op}^2\|\hat{\boldsymbol{h}}_\mu\|^2) && \text{by } \|\boldsymbol{D}_\mu\|_{op}\le 1\\
&\le \|\hat{\boldsymbol{A}}_\mu\|_F^2(n+\|\boldsymbol{G}\|_{op}^2)(\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2)\\
&\le p(n\mu)^{-2}(n+\|\boldsymbol{G}\|_{op}^2)(\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2) && \text{by } \|\hat{\boldsymbol{A}}_\mu\|_F^2\le p\|\hat{\boldsymbol{A}}_\mu\|_{op}^2\le p(n\mu)^{-2}.
\end{aligned}$$

By the same algebra, we bound $\frac{1}{2n}\sum_{ij}\|\frac{\partial\hat{\boldsymbol{\psi}}_\mu}{\partial g_{ij}}\|^2$ from above as

$$\begin{aligned}
\frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^p\|\frac{\partial\hat{\boldsymbol{\psi}}_\mu}{\partial g_{ij}}\|^2 &\le \frac{1}{n}\sum_{ij}\Big(\|\boldsymbol{D}_\mu\boldsymbol{G}\hat{\boldsymbol{A}}_\mu\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu\|^2 + \|\boldsymbol{V}_\mu\boldsymbol{e}_i\boldsymbol{e}_j^\top\hat{\boldsymbol{h}}_\mu\|^2\Big)\\
&= n^{-1}\|\boldsymbol{D}_\mu\boldsymbol{G}\hat{\boldsymbol{A}}_\mu\|_F^2\|\boldsymbol{\psi}_\mu\|^2 + n^{-1}\|\boldsymbol{V}_\mu\|_F^2\|\hat{\boldsymbol{h}}_\mu\|^2\\
&\le \|\hat{\boldsymbol{A}}_\mu\|_F^2\|\boldsymbol{G}\|_{op}^2\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2\\
&\le (\|\hat{\boldsymbol{A}}_\mu\|_F^2\|\boldsymbol{G}\|_{op}^2+1)(\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2)\\
&\le (p(n\mu)^{-2}\|\boldsymbol{G}\|_{op}^2+1)(\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2)
\end{aligned}$$

By the above displays, we obtain the upper bound of $\Xi = \mathbb{E}[\frac{1}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}\sum_{ij}(\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|^2)]$:

$$\Xi \le \mathbb{E}\left[2p(n\mu)^{-2}(n+\|\boldsymbol{G}\|_{op}^2) + 2(p(n\mu)^{-2}\|\boldsymbol{G}\|_{op}^2+1)\right] \overset{(*)}{=} O(n^{2c}),$$

where $(*)$ follows from $\mu = n^{-c}$ with $c>0$ and $\mathbb{E}[\|\boldsymbol{G}\|_{op}^2] = O(n)$ for the matrix $\boldsymbol{G}\in\mathbb{R}^{n\times p}$ with iid standard normal entries as $p/n\to\gamma\in(0,\infty)$. Substituting this bound to the right-hand side of Lemma 2, we have

$$\mathbb{E}\Big[\frac{|\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 - \frac{1}{n}\sum_{j=1}^p(\boldsymbol{\psi}_\mu^\top\boldsymbol{G}e_j - \sum_{i=1}^n\frac{\partial\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu}{\partial g_{ij}})^2|}{\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2}\Big] \le C(\sqrt{n+p}(1+\Xi^{1/2})+\Xi) \overset{(**)}{=} O(n^{c+1/2}), \tag{39}$$

where we have used $0<c<1/2$ and $\Xi = O(n^{2c})$ for $(**)$. It remains to bound the error $\sum_{j=1}^p(\boldsymbol{\psi}_\mu^\top\boldsymbol{G}e_j - \sum_{i=1}^n(\partial/\partial g_{ij})\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu)^2 - \|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2$ inside the expectation on the left-hand side. Now the derivatives formula and the KKT condition $\boldsymbol{G}^\top\boldsymbol{\psi}_\mu = n\mu\hat{\boldsymbol{h}}_\mu$ yield

$$\sum_{j=1}^p(\boldsymbol{\psi}_\mu^\top\boldsymbol{G}e_j - \sum_{i=1}^n\frac{\partial\boldsymbol{e}_i^\top\boldsymbol{\psi}_\mu}{\partial g_{ij}})^2 = \|n\mu\hat{\boldsymbol{h}}_\mu + \hat{\boldsymbol{A}}_\mu^\top\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{\psi}_\mu + \operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2,$$

so that we have

$$\begin{aligned}
&\frac{1}{2}\Big|\sum_{j=1}^{p}(\boldsymbol{\psi}_\mu^\top \boldsymbol{G} e_j - \sum_{i=1}^{n}\frac{\partial \boldsymbol{e}_i^\top \boldsymbol{\psi}_\mu}{\partial g_{ij}})^2 - \|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2\Big| \\
&= 2^{-1}\|n\mu\hat{\boldsymbol{h}}_\mu + \hat{\boldsymbol{A}}_\mu^\top \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu + \operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2 - \|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2| \\
&\le 2^{-1}\|n\mu\hat{\boldsymbol{h}}_\mu + \hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|^2 + \|n\mu\hat{\boldsymbol{h}}_\mu + \boldsymbol{A}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|\|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\| && \text{by } |(a+b)^2-b^2| \le a^2 + 2|ab| \\
&\le n^2\mu^2\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\boldsymbol{A}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|^2 + |\operatorname{tr}[\boldsymbol{V}_\mu]|\|n\mu\hat{\boldsymbol{h}}_\mu + \boldsymbol{A}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|\|\hat{\boldsymbol{h}}_\mu\| && \text{by } 2^{-1}(a+b)^2 \le a^2+b^2 \\
&\le n^2\mu^2\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\boldsymbol{A}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|^2 + n\|n\mu\hat{\boldsymbol{h}}_\mu + \boldsymbol{A}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|\|\hat{\boldsymbol{h}}_\mu\| && \text{by } |\operatorname{tr}[\boldsymbol{V}_\mu]| \le n\|\boldsymbol{V}_\mu\|_{op} \le n \\
&\le n^2\mu^2\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|^2 + n^2\mu\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|\|\hat{\boldsymbol{h}}_\mu\| && \text{by triangle inequality} \\
&\le (n^2\mu^2 + n^2\mu + 2^{-1})\|\hat{\boldsymbol{h}}_\mu\|^2 + (1+2^{-1})\|\hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\boldsymbol{\psi}_\mu\|^2 && \text{by } ab \le 2^{-1}(a^2+b^2) \\
&\le (\|\hat{\boldsymbol{h}}_\mu\|^2 + n^{-1}\|\boldsymbol{\psi}_\mu\|^2)(n^2\mu^2 + n^2\mu + 2^{-1} + 2^{-1}3n\|\hat{\boldsymbol{A}}_\mu \boldsymbol{G}^\top \boldsymbol{D}_\mu\|_{op}^2) \\
&\le (\|\hat{\boldsymbol{h}}_\mu\|^2 + n^{-1}\|\boldsymbol{\psi}_\mu\|^2)(n^2\mu^2 + n^2\mu + 2^{-1} + 2^{-1}3n(n\mu)^{-2}\|\boldsymbol{G}\|_{op}^2) && \text{by } \|\hat{\boldsymbol{A}}_\mu\|_{op} \le (n\mu)^{-1} \\
& && \text{and } \|\boldsymbol{D}_\mu\|_{op} \le 1.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\frac{1}{n}\left|\sum_{j=1}^{p}(\boldsymbol{\psi}_\mu^\top \boldsymbol{G} e_j - \sum_{i=1}^{n}\frac{\partial \boldsymbol{e}_i^\top \boldsymbol{\psi}_\mu}{\partial g_{ij}})^2 - \|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2\right|}{\|\boldsymbol{\psi}_\mu\|^2/n + \|\hat{\boldsymbol{h}}_\mu\|^2}\Big] \\
&\le C_5\,\mathbb{E}\big[n\mu^2 + n\mu + n^{-1} + (n\mu)^{-2}\|\boldsymbol{G}\|_{op}^2\big] \\
&= O(n^{1-2c} + n^{1-c} + n^{-1} + n^{-1+2c}) && \text{by } \mu = n^{-c} \text{ and } \mathbb{E}[\|\boldsymbol{G}\|_{op}^2] = O(n), \\
&= O(n^{1-c}) && \text{by } 0 < c \le 1/4 < 2/3
\end{aligned}$$

Combined with (39), by the triangle inequality,

$$\mathbb{E}\left[\frac{|\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 - \frac{1}{n}\|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2|}{\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\boldsymbol{\psi}_\mu\|^2/n}\right] = O(n^{c+1/2}) + O(n^{1-c}) = O(n^{1-c}) \tag{40}$$

by $c \in (0, 1/4]$. Thus, we obtain

$$\Big|\frac{p}{n}\|\boldsymbol{\psi}_\mu\|^2 - \frac{1}{n}\|\operatorname{tr}[\boldsymbol{V}_\mu]\hat{\boldsymbol{h}}_\mu\|^2\Big| = \Big(\|\hat{\boldsymbol{h}}_\mu\|^2 + \frac{\|\boldsymbol{\psi}_\mu\|^2}{n}\Big)O_P(n^{1-c}) \le (\|\hat{\boldsymbol{h}}_\mu\|^2 + \|\psi\|_\infty^2)O_P(n^{1-c}) = O_P(n^{1-c}),$$

where we have used $\|\hat{\boldsymbol{h}}_\mu\|_2^2 = O_P(1)$ (see Lemma 3). This finishes the proof.

B.2 Convergence of smoothed quantities

The lemmas below are the key to relating Lemma 1 to Theorem 1.

Lemma 3 *Suppose that (ρ, F_ϵ) satisfies Assumption 1 and Assumption 2. Then, under $\mu = n^{-c}$ for any $c > 0$, we have*

$$\|\hat{\boldsymbol{h}}\|^2 \to^p \alpha^2, \qquad \|\hat{\boldsymbol{h}}_\mu\|^2 \to^p \alpha^2, \qquad \alpha^2 < +\infty, \qquad \|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\| = o_p(1)$$

where α is the unique solution to the nonlinear system (15).

Proof The definition of $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{h}}_\mu$ are recalled here for convenience:

$$\hat{\boldsymbol{h}} \in \arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - \boldsymbol{g}_i^\top \boldsymbol{h}), \quad \hat{\boldsymbol{h}}_\mu \in \arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(\epsilon_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \frac{\mu\|\boldsymbol{h}\|^2}{2},$$

where $(\epsilon_i)_{i=1}^n \overset{\text{iid}}{\sim} F_\epsilon$, $(\boldsymbol{g}_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, and $(\epsilon_i)_{i=1}^n \perp\!\!\!\perp (\boldsymbol{g}_i)_{i=1}^n$. Since Assumption 1-2 imply the assumption in Theorem 7 (see Remark 1), the nonlinear system of equations (15) has a unique solution α and $\|\hat{\boldsymbol{h}}\|^2 \to^p \alpha^2$. It remains to show $\|\hat{\boldsymbol{h}}_\mu\|^2 \to^p \alpha^2$ under the diminishing regularization parameter $\mu = n^{-c}$. Define

$$f: \qquad \mathbb{R}^p \to \mathbb{R}, \qquad \boldsymbol{x} \mapsto n^{-c}\|\boldsymbol{x}\|^2/2$$

so that $\hat{\boldsymbol{h}}_\mu$ is the regularized M-estimator with the Lipschitz convex loss ρ and penalty f. Suppose the following condition is satisfied:

$$\forall a \in \mathbb{R}, \forall \tau > 0, \quad n^{-1}(e_f(a\boldsymbol{g}; \tau) - f(\boldsymbol{0}_p)) \to^p 0, \text{ where } \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p). \tag{41}$$

where $e_f(x;\tau) = \arg\min_u \frac{1}{2\tau}(x-u)^2 + f(u)$ is the Moreau envelope of f. Let $L(c,\tau) := \mathbb{E}[e_\rho(cZ + W;\tau) - \rho(W)]$ and assume $Z \sim N(0,1)$, $W \sim F_\epsilon$, $Z \perp\!\!\!\perp W$. Then, by (Thrampoulidis et al., 2018, Theorem 3.1) (with $F = 0$ and taking for instance the signal distribution to be a single point mass at 0), if the convex-concave optimization

$$\inf_{\alpha\geq 0, \tau_g > 0} \sup_{\beta \geq 0, \tau_h > 0} \frac{\beta\tau_g}{2} + \frac{1}{\gamma}\cdot L(\alpha, \frac{\tau_g}{\beta}) - \frac{\alpha\tau_h}{2} - \frac{\alpha^2\beta^2}{2\tau_h}, \tag{42}$$

admits a unique solution α_*, we have $\|\hat{\boldsymbol{h}}_\mu\|^2 \to^p \alpha_*^2$. Note that the stationary condition of the above convex-concave optimization is the nonlinear system of equations (15) (cf. (Thrampoulidis et al., 2018, Section 5.1)), while Theorem 7 guarantees the uniqueness and existence of the solution to (15). Therefore, we obtain $\|\hat{\boldsymbol{h}}_\mu\|_2^2 \to^p \alpha_*^2$ where α_* is the unique solution to (15) (we add a star to denote the unique solution here, to avoid confusion with the variable α of the minimization in (42)). It remains to show (41). Note that the Moreau envelope $e_f(\boldsymbol{x};\tau)$ of the convex function $\boldsymbol{x} \in \mathbb{R}^p \mapsto n^{-c}\|\boldsymbol{x}\|^2/2$ is given by

$$\forall \boldsymbol{x} \in \mathbb{R}^p, \forall \tau > 0, \quad e_f(\boldsymbol{x};\tau) = \min_{\boldsymbol{u}\in\mathbb{R}^p} \frac{\|\boldsymbol{x}-\boldsymbol{u}\|^2}{2\tau} + \frac{n^{-c}}{2}\|\boldsymbol{u}\|^2 = \frac{n^{-c}}{2(1+n^{-c}\tau)}\|\boldsymbol{x}\|^2,$$

so that for all $a \in \mathbb{R}$ and $\tau > 0$, we have

$$\frac{e_{f_\mu}(a\boldsymbol{g};\tau) - f_\mu(\boldsymbol{0}_p)}{n} = \frac{n^{-c}}{1+n^{-c}\tau}\frac{a^2\|\boldsymbol{g}\|^2}{n} = \frac{n^{-c}}{1+n^{-c}\tau}\frac{a^2}{n}\sum_{i=1}^n g_i^2 \to^p 0\cdot a^2 = 0,$$

where we have used the weak law of large number to the iid sum $n^{-1}\sum_{i=1}^n g_i^2$ with $\mathbb{E}[g_i^2] = 1$. This finishes the proof of (41).

It remains to show $\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}\| = o_p(1)$. Now we verify $\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\|^2 \leq \|\hat{\boldsymbol{h}}\|^2 - \|\hat{\boldsymbol{h}}_\mu\|^2$. Letting $\mathcal{L}: \mathbb{R}^p \to \mathbb{R}$ be the convex function $\mathcal{L}(\boldsymbol{h}) := \sum_{i=1}^n \rho(\epsilon - \boldsymbol{g}_i^\top \boldsymbol{h})$ then $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{h}}_\mu$ solve

$$\hat{\boldsymbol{h}} \in \arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{h}), \qquad \hat{\boldsymbol{h}}_\mu \in \arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{h}) + \frac{\mu}{2}\|\boldsymbol{h}\|^2$$

Note in passing that $\hat{\boldsymbol{h}}_\mu$ is also a minimizer of the convex function $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$

$$\hat{\boldsymbol{h}}_\mu \in \operatorname*{arg\,min}_{\boldsymbol{h}\in\mathbb{R}^p} \mathcal{F}(\boldsymbol{h}) \quad \text{where} \quad \mathcal{F}(\boldsymbol{h}) := \mathcal{L}(\boldsymbol{h}) + \frac{\mu}{2}\Big(\|\boldsymbol{h}\|^2 - \|\boldsymbol{h} - \hat{\boldsymbol{h}}_\mu\|^2\Big),$$

since the gradient of $\boldsymbol{h} \mapsto \|\boldsymbol{h} - \hat{\boldsymbol{h}}_\mu\|^2$ is $\boldsymbol{0}_p$ at $\hat{\boldsymbol{h}}_\mu$ and $\hat{\boldsymbol{h}}_\mu$ satisfies the new KKT condition $\nabla\mathcal{L}(\hat{\boldsymbol{h}}_\mu) = \boldsymbol{0}_p$. Then, $\mathcal{F}(\hat{\boldsymbol{h}}_\mu) \leq \mathcal{F}(\hat{\boldsymbol{h}})$ and $\mathcal{L}(\hat{\boldsymbol{h}}) \leq \mathcal{L}(\hat{\boldsymbol{h}}_\mu)$ yield

$$\begin{aligned} 0 \geq \mathcal{F}(\hat{\boldsymbol{h}}_\mu) - \mathcal{F}(\hat{\boldsymbol{h}}) &= \frac{\mu}{2}\big(\|\hat{\boldsymbol{h}}_\mu\|^2 - \|\hat{\boldsymbol{h}}\|^2 + \|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\|^2\big) + \mathcal{L}(\hat{\boldsymbol{h}}_\mu) - \mathcal{L}(\hat{\boldsymbol{h}}) \\ &\geq \frac{\mu}{2}\big(\|\hat{\boldsymbol{h}}_\mu\|^2 - \|\hat{\boldsymbol{h}}\|^2 + \|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\|^2\big) + 0, \end{aligned}$$

so that $\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\|^2 \leq \|\hat{\boldsymbol{h}}\|^2 - \|\hat{\boldsymbol{h}}_\mu\|^2$. Combined with $\|\hat{\boldsymbol{h}}\|^2, \|\hat{\boldsymbol{h}}_\mu\|^2 \to^p \alpha^2 < +\infty$, we have $\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\|^2 = o_p(1)$ and conclude the proof. ■

Lemma 4 *Suppose that (ρ, F_ϵ) satisfies Assumption 1-2. Then, under $\mu = n^{-c}$ for some $c > 0$, we have*

$$\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|_2 = o_P(n^{\frac{1-c}{2}}) \qquad \textit{and} \qquad \|\boldsymbol{\psi}\|^2 - \|\boldsymbol{\psi}_\mu\|^2 = o_P(n^{1-\frac{c}{2}}),$$

where $\boldsymbol{\psi} = \psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}})$ and $\boldsymbol{\psi}_\mu = \psi(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)$.

Proof Note that $\boldsymbol{G}^\top\boldsymbol{\psi}_\mu = n\mu\hat{\boldsymbol{h}}_\mu$ and $\boldsymbol{G}^\top\boldsymbol{\psi} = \boldsymbol{0}_p$ by the KKT conditions. Then, using Lemma 7 that is introduced later, $\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|^2$ can be bounded from above as

$$\begin{aligned} \|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|^2 &\leq (\boldsymbol{\psi}_\mu - \boldsymbol{\psi})^\top(\boldsymbol{G}\hat{\boldsymbol{h}} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu) && \text{by Lemma 7 with } \boldsymbol{u} = \boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu \text{ and } \boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}} \\ &\leq (n\mu\hat{\boldsymbol{h}}_\mu - \boldsymbol{0}_p)^\top(\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu) && \text{by } \boldsymbol{G}^\top\boldsymbol{\psi}_\mu = n\mu\hat{\boldsymbol{h}}_\mu \text{ and } \boldsymbol{G}^\top\boldsymbol{\psi} = \boldsymbol{0}_p \\ &\leq n\mu\|\hat{\boldsymbol{h}}_\mu\|\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{h}}_\mu\| && \text{by the Cauchy–Schwarz inequality} \\ &= o_P(n^{1-c}) && \text{by Lemma 3 and } \mu = n^{-c}, \end{aligned} \tag{43}$$

which finishes the proof for $\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|^2$. For the bound of $\|\boldsymbol{\psi}_\mu\|^2 - \|\boldsymbol{\psi}\|^2$, the Cauchy–Schwarz inequality implies

$$|\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{\psi}_\mu\|^2| = |(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu)^\top(\boldsymbol{\psi} + \boldsymbol{\psi}_\mu)| \leq \|\boldsymbol{\psi} + \boldsymbol{\psi}_\mu\|\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\| \leq 2\sqrt{n}\|\psi\|_\infty\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\|.$$

Since $\|\psi\|_\infty < +\infty$ by Assumption 1 and we have shown that $\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\|^2 = o_P(n^{1-c})$, we obtain $|\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{\psi}_\mu\|^2| = o_P(n^{1-\frac{c}{2}})$. This finishes the proof. ■

Lemma 5 *Suppose that (ρ, F_ϵ) satisfies Assumption 1-2. Then, under $\mu = n^{-c}$ for some $c \in (0, 1)$, we have*

$$\operatorname{tr}[\boldsymbol{V}_\mu]^2 - \operatorname{tr}[\boldsymbol{V}]^2 = o_P(n^{2-\frac{c}{2}}),$$

where $\boldsymbol{V} = (\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}$ and $\boldsymbol{V}_\mu = (\partial/\partial\boldsymbol{\epsilon})\boldsymbol{\psi}_\mu$.

Proof By Assumption 2, we can write $\boldsymbol{\epsilon} = \boldsymbol{z} + \boldsymbol{\delta}$ where $\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{\delta}$ and $(z_i)_{i=1}^n$ has i.i.d. density $\exp(-\phi(z))$. By the chain rule, $\boldsymbol{V} - \boldsymbol{V}_\mu$ can be written as

$$\boldsymbol{V} - \boldsymbol{V}_\mu = \frac{\partial}{\partial \boldsymbol{\epsilon}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}} \cdot \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \boldsymbol{I}_n \cdot \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu) = \frac{\partial}{\partial \boldsymbol{z}}(\boldsymbol{\psi} - \boldsymbol{\psi}_\mu).$$

Then, Theorem 6 applied with $\boldsymbol{f} = \boldsymbol{\psi} - \boldsymbol{\psi}_\mu$ yields

$$\begin{aligned}
&\mathbb{E}\big[(\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi}) - \operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu])^2\big] && (44)\\
&= \mathbb{E}\Big[\sum_{i=1}^n \phi''(z_i)(\psi_i - (\psi_\mu)_i)^2 + \operatorname{tr}[(\boldsymbol{V} - \boldsymbol{V}_\mu)^2]\Big] && \text{by Theorem 6}\\
&\le \|\phi''\|_\infty^2 \,\mathbb{E}[\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\|^2] + n\,\mathbb{E}[\|\boldsymbol{V} - \boldsymbol{V}_\mu\|_{op}^2] && \text{using } \operatorname{tr}[\boldsymbol{M}] \le n\|\boldsymbol{M}\|_{op} \text{ for } \boldsymbol{M} \in \mathbb{R}^{n\times n}\\
&\le 4\|\phi''\|_\infty^2 \|\psi\|_\infty^2 n + 4n && \text{using } \|\boldsymbol{V}\|_{op}, \|\boldsymbol{V}_\mu\|_{op} \le 1 \text{ from (30)}\\
&= C_6(\rho, F_\epsilon) n, && \text{since } \|\phi''\|_\infty, \|\psi\|_\infty < +\infty \text{ by Assumption 1-2,}
\end{aligned}$$

so that $\big|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu] - \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi})\big| = O_P(n^{1/2})$. From this bound and the triangle inequality, it follows that

$$\begin{aligned}
\big|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]\big| &\le \big|\phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi})\big| + \big|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu] - \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\mu - \boldsymbol{\psi})\big| && (45)\\
&\le \|\phi'(\boldsymbol{z})\|\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mu\| + O_P(n^{1/2}) && \text{by the Cauchy–Schwarz}\\
&= O_P(n^{1/2}) o_P(n^{\frac{1-c}{2}}) + O_P(n^{1/2}) && \text{by Lemma 4 and 8}\\
&= o_P(n^{1-\frac{c}{2}}) && \text{by } c{<}1.
\end{aligned}$$

Finally, using $\|\boldsymbol{V}\|_{op}, \|\boldsymbol{V}_\mu\|_{op} \le 1$ from (30), we have

$$|\operatorname{tr}[\boldsymbol{V}]^2 - \operatorname{tr}[\boldsymbol{V}_\mu]^2| = |\operatorname{tr}[\boldsymbol{V}] - \operatorname{tr}[\boldsymbol{V}_\mu]| \cdot |\operatorname{tr}[\boldsymbol{V}] + \operatorname{tr}[\boldsymbol{V}_\mu]| \le o_P(n^{1-\frac{c}{2}}) \cdot 2n = o_P(n^{2-\frac{c}{2}}),$$

which concludes the proof. ■

In order to relax the assumption on the noise and prove Proposition 1, we now provide here a modification of this argument to allow for a vanishing smooth noise component z_i: the conclusion $|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]|/n \to^P 0$ still holds if the linear model noise is $\epsilon_i = \sigma_n z_i + \delta_i$ with vanishing σ_n (depending on n) provided that $\sigma_n \ge \sqrt{\mu}$ and $\sigma_n\sqrt{n} \to +\infty$.

Lemma 6 (Vanishing noise component) *Assume that the noise $\boldsymbol{\epsilon}$ in the linear model has iid coordinates of the form $\epsilon_i = \sigma_n z_i + \delta_i$ where δ_i and z_i are independent and z_i has density $z \mapsto \exp(-\phi(z))$ with twice-continuously differentiable ϕ and $\sup_{x\in\mathbb{R}} |\phi''(z)| < +\infty$. If $\sigma_n \ge \sqrt{\mu}$, we have*

$$\mathbb{E}\Big[\frac{|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]|}{n}\Big] \le \mathbb{E}\Big[\min\Big(2, \frac{\|\phi'(\boldsymbol{z})\|}{\sqrt{n}}\|\boldsymbol{h}_\mu\|\|\boldsymbol{h} - \boldsymbol{h}_\mu\|\Big)\Big] + \frac{2\|\phi''\|_\infty\|\psi\|_\infty}{\sigma_n\sqrt{n}} + \frac{2}{\sqrt{n}}. \tag{46}$$

If we take $\mu = n^{-c}$ with $c \in (0,1)$ as in Lemma 3, then $\|\phi'(\boldsymbol{z})\| n^{-\frac{1}{2}} \|\boldsymbol{h}_\mu\|\|\boldsymbol{h} - \boldsymbol{h}_\mu\| \to^P 0$ by Lemma 3 and Lemma 8 for the first term, while $\sigma_n\sqrt{n} \ge \sqrt{n^{1-c}} \to +\infty$ for the second term. This implies that $|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]|/n$ converges to 0 in L1 and in probability for all $\sigma_n \ge n^{-c/2}$.

Proof This is a modification of the argument of Lemma 5 to obtain a condition on how small the amplitude σ_n of the smooth noise z_i is allowed. Similarly to (44), by Theorem 6 we have

$$\begin{aligned}\mathbb{E}\big[(\phi'(\boldsymbol{z})^\top(\frac{\boldsymbol{\psi}_\mu - \boldsymbol{\psi}}{\sigma_n}) - \operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu])^2\big] &= \mathbb{E}\Big[\sum_{i=1}^n \phi''(z_i)\frac{(\psi_i - (\psi_\mu)_i)^2}{\sigma_n^2} + \operatorname{tr}[(\boldsymbol{V} - \boldsymbol{V}_\mu)^2]\Big] \\ &\le 4n(\|\phi''\|_\infty^2\|\psi\|_\infty^2/\sigma_n^2 + 1)\end{aligned}$$

using $\|\boldsymbol{V}\|_{op}, \|\boldsymbol{V}_\mu\|_{op} \le 1$ from (30). By the triangle inequality combined with $0 \le \operatorname{tr}[\boldsymbol{V}] \le n$ and $0 \le \operatorname{tr}[\boldsymbol{V}_\mu] \le n$,

$$|\operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]| \le \min\Big(2n, \frac{\|\phi'(\boldsymbol{z})\|\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|}{\sigma_n}\Big) + \Big|\phi'(\boldsymbol{z})^\top(\frac{\boldsymbol{\psi}_\mu - \boldsymbol{\psi}}{\sigma_n}) - \operatorname{tr}[\boldsymbol{V} - \boldsymbol{V}_\mu]\Big|$$

Recalling (43) we have $\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|^2 \le n\mu\|\boldsymbol{h}_\mu\|\|\boldsymbol{h} - \boldsymbol{h}_\mu\|$ if $\sigma_n \ge \sqrt{\mu}$ we obtain

$$\mathbb{E}\big[|\operatorname{tr}[\boldsymbol{V}-\boldsymbol{V}_\mu]|/n\big] \le \mathbb{E}\big[\min(2, (\|\phi'(\boldsymbol{z})\|n^{-1/2})\|\boldsymbol{h}_\mu\|\|\boldsymbol{h}-\boldsymbol{h}_\mu\|)\big]+2(\|\phi''\|_\infty^2\|\psi\|_\infty^2/\sigma_n^2+1)^{1/2}/\sqrt{n}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain (46). ■

Lemma 7 *If $\psi : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz and nondecreasing, it holds that*

$$\|\psi(\boldsymbol{u}) - \psi(\boldsymbol{v})\|^2 \le (\psi(\boldsymbol{u}) - \psi(\boldsymbol{v}))^\top(\boldsymbol{u} - \boldsymbol{v})$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$.

Proof We have $0 \le \psi(x) - \psi(y) \le (x - y)$ for all $x, y \in \mathbb{R}$ such that $x \le y$. Switching the role of x and y in this inequality, we have $(\psi(x) - \psi(y))^2 \le (\psi(x) - \psi(y))(x - y)$ for all $x, y \in \mathbb{R}$. Applying this inequality with $(x, y) = (u_i, v_i)$ for each $i \in [n]$, we conclude the proof. ■

Lemma 8 *Suppose $(z_i)_{i=1}^n$ has iid density $\exp(-\phi(z))$ where $\phi : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and $\|\phi''\|_\infty < +\infty$. Then $\mathbb{E}[\phi'(z_i)^2] = \mathbb{E}[\phi''(z_i)]$ and $\lim_{n\to\infty} \mathbb{P}(\|\phi'(\boldsymbol{z})\|^2 \le n(\|\phi''\|_\infty + 1)) = 1$.*

Proof Note $\mathbb{E}[\phi'(z_i)^2] = \mathbb{E}[\phi''(z_i)] \le \|\phi''\|_\infty < +\infty$ by integration part. By the weak law of large number, we have $n^{-1}\|\phi'(\boldsymbol{z})\|^2 = n^{-1}\sum_{i=1}^n \phi'(z_i)^2 \to^p \mathbb{E}[\phi'(z)^2] = \mathbb{E}[\phi''(z_i)]$. Thus, we obtain

$$\mathbb{P}(n^{-1}\|\phi'(\boldsymbol{z})\|^2 > \|\phi''\|_\infty + 1) \le \mathbb{P}(n^{-1}\|\phi'(\boldsymbol{z})\|^2 > \mathbb{E}[\phi''(z_i)] + 1) \to 0,$$

which completes the proof. ■

B.3 Lower bound on the trace of V

In this section, we derive a lower bound of $\mathrm{tr}[\boldsymbol{V}]/n$.

Lemma 9 *Suppose that (ρ, F_ϵ) satisfy Assumption 1-2. Let $\alpha(\rho, F_\epsilon, \gamma)$ be the unique solution to the nonlinear system of equations (15) for $(\rho, F_\epsilon, \gamma)$ and $\eta = \eta(\rho)$ be some positive constant such that $\psi(x)^2/\|\psi\|_\infty^2 + \psi'(x) \ge \eta^2$ for almost every $x \in \mathbb{R}$ (we can always take such $\eta > 0$ thanks to Assumption 1(3)). Then, we have*

$$\mathbb{P}\left(\frac{\mathrm{tr}[\boldsymbol{V}]^2}{n^2} \ge B(\gamma, \rho, F_\epsilon)\right) \to 1, \text{ where } B(\gamma,\rho,F_\epsilon) = C(\gamma)\cdot\min(\eta(\rho)^4, \frac{\|\psi\|_\infty^2\cdot\eta(\rho)^2}{\alpha^2(\rho,F_\epsilon,\gamma)}) > 0.$$

Proof Recall that we have shown in Lemma 5 that

$$\mathrm{tr}(\boldsymbol{V})^2/n^2 - \mathrm{tr}(\boldsymbol{V}_\mu)^2/n^2 = o_P(n^{-\frac{c}{2}}) = o_P(1) \text{ under } \mu = n^{-c} \text{ for all } c \in (0,1).$$

Thus, it suffices to show the lower bound for $\mathrm{tr}[\boldsymbol{V}_\mu]^2/n^2$ for some $c \in (0,1)$, say $c = 1/4$. Define the matrices $\boldsymbol{D}$ and $\tilde{\boldsymbol{D}}_\mu$ as

$$\boldsymbol{D}_\mu := \mathrm{diag}\{\psi'(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)\}, \quad \tilde{\boldsymbol{D}}_{\boldsymbol{\mu}} := \|\psi\|_\infty^{-2}\,\mathrm{diag}\{\psi^2(\boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\mu)\}.$$

Thanks to $\|\psi\|_{\mathrm{lip}} \le 1$ and $\psi(x)^2/\|\psi\|_\infty^2 + \psi'(x) \ge \eta^2$, we have

$$\mathrm{tr}(\boldsymbol{D}_\mu) \le n, \quad \tilde{\boldsymbol{D}}_\mu + \boldsymbol{D}_\mu \succeq \eta^2 \boldsymbol{I}_n$$

where $\succeq$ is the positive semi-definite order. By the derivative formula (Theorem 4), we have

$$\begin{aligned}
\boldsymbol{V}_\mu &= \boldsymbol{D}_\mu - \boldsymbol{D}_\mu\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{G} + n\mu\boldsymbol{I}_p)^{-1}\boldsymbol{G}^\top\boldsymbol{D}_\mu\\
&= \boldsymbol{D}_\mu^{1/2}(\boldsymbol{I}_n - \boldsymbol{D}_\mu^{1/2}\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{G} + n\mu\boldsymbol{I}_p)^{-1}\boldsymbol{G}^\top\boldsymbol{D}_\mu^{1/2})\boldsymbol{D}_\mu\\
&= \boldsymbol{D}_\mu^{1/2}\boldsymbol{H}_\mu\boldsymbol{D}_\mu^{1/2},
\end{aligned}$$

where $\boldsymbol{H}_\mu := \boldsymbol{I}_n - \boldsymbol{D}_\mu^{1/2}\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{D}_\mu\boldsymbol{G} + n\mu\boldsymbol{I}_p)^{-1}\boldsymbol{G}^\top\boldsymbol{D}_\mu^{1/2}$ is positive semi-definite, and satisfies $\|\boldsymbol{H}_\mu\|_{op} \le 1$ and $\mathrm{tr}(\boldsymbol{H}_\mu) \ge n - p$. Then, simple algebra yields

$$\begin{aligned}
\mathrm{tr}[\boldsymbol{V}_\mu] &= \mathrm{tr}(\boldsymbol{D}_\mu^{1/2}\boldsymbol{H}_\mu\boldsymbol{D}_\mu^{1/2})\\
&= \mathrm{tr}(\boldsymbol{H}_\mu^{1/2}\boldsymbol{D}_\mu\boldsymbol{H}_\mu^{1/2}) && \text{by } \mathrm{tr}(\boldsymbol{D}_\mu^{1/2}\boldsymbol{H}_\mu\boldsymbol{D}_\mu^{1/2}) = \mathrm{tr}(\boldsymbol{H}_\mu\boldsymbol{D}_\mu) = \mathrm{tr}(\boldsymbol{H}_\mu^{1/2}\boldsymbol{D}_\mu\boldsymbol{H}_\mu^{1/2})\\
&\ge \mathrm{tr}(\boldsymbol{H}_\mu^{1/2}(\eta^2\boldsymbol{I}_n - \tilde{\boldsymbol{D}}_\mu)\boldsymbol{H}_\mu^{1/2}) && \text{by } \tilde{\boldsymbol{D}}_\mu + \boldsymbol{D}_\mu \succeq \eta^2\boldsymbol{I}_n\\
&= \eta^2\,\mathrm{tr}(\boldsymbol{H}_\mu) - \mathrm{tr}(\boldsymbol{H}_\mu^{1/2}\tilde{\boldsymbol{D}}_\mu\boldsymbol{H}_\mu^{1/2})\\
&\ge \eta^2(n-p) - \mathrm{tr}(\tilde{\boldsymbol{D}}_\mu)\|\boldsymbol{H}_\mu\|_{op} && \text{by } \mathrm{tr}[\boldsymbol{H}_\mu] \ge n-p \text{ and } \mathrm{tr}[\boldsymbol{A}\boldsymbol{B}] \le \mathrm{tr}[\boldsymbol{A}]\cdot\|\boldsymbol{B}\|_{op}\\
&\ge \eta^2(n-p) - \|\psi\|_\infty^{-2}\|\boldsymbol{\psi}_\mu\|^2\cdot 1.
\end{aligned}$$

Thus, under the event $\Omega := \{\|\boldsymbol{\psi}_\mu\|^2 \le n\|\psi\|_\infty^2\eta^2(1-\gamma)/2\}$, we have

$$\mathbf{1}_\Omega\cdot n^{-2}\,\mathrm{tr}[\boldsymbol{V}_\mu]^2 \ge \mathbf{1}_\Omega\cdot 4^{-1}\eta^4(1-\gamma)^2 \tag{47}$$

It remains to bound $n^{-2}\,\mathrm{tr}[\boldsymbol{V}_\mu]^2$ from below under the complement Ω^c. Lemma 1 implies

$$n^{-2}\mathrm{tr}(\boldsymbol{V}_\mu)^2\|\hat{\boldsymbol{h}}_\mu\|^2 - n^{-2}p\|\boldsymbol{\psi}_\mu\|^2 = O_P(n^{-c}) = o_P(1),$$

where $n^{-2}\mathrm{tr}(\boldsymbol{V}_\mu)^2\|\hat{\boldsymbol{h}}_\mu\|^2 = n^{-2}\mathrm{tr}(\boldsymbol{V}_\mu)^2\alpha^2 + o_P(1)$ from $\mathrm{tr}[\boldsymbol{V}_\mu]^2/n^2 \le \|\boldsymbol{V}_\mu\|_{op}^2 \le 1$ and $\|\hat{\boldsymbol{h}}_\mu\|^2 \to^p \alpha^2 > 0$ by Lemma 3. Thus, we have $n^{-2}\mathrm{tr}[\boldsymbol{V}_\mu]^2\alpha^2 - \gamma n^{-1}\|\boldsymbol{\psi}_\mu\|^2 = o_P(1)$, so that

$$\mathbb{P}\Big(\frac{\mathrm{tr}[\boldsymbol{V}_\mu]^2}{n^2} - \frac{\gamma}{n\alpha^2}\|\boldsymbol{\psi}_\mu\|^2 \ge -\frac{\|\psi\|_\infty^2\eta^2(1-\gamma)\gamma}{4\alpha^2}\Big) \to 1.$$

Thereby, under $\Omega^c = \{\|\boldsymbol{\psi}_\mu\|^2 \ge n\|\psi\|_\infty^2\eta^2(1-\gamma)/2\}$, we have

$$\mathbb{P}\Big(\mathbf{1}_{\Omega^c}\frac{\mathrm{tr}(\boldsymbol{V}_\mu)^2}{n^2} \ge \mathbf{1}_{\Omega^c}\Big(\frac{\|\psi\|_\infty^2\eta^2(1-\gamma)\gamma}{2\alpha^2} - \frac{\|\psi\|_\infty^2\eta^2(1-\gamma)\gamma}{4\alpha^2}\Big) = \mathbf{1}_{\Omega^c}\frac{\|\psi\|_\infty^2\eta^2(1-\gamma)\gamma}{4\alpha^2}\Big) \to 1 \tag{48}$$

Consequently, (47) and (47) yield

$$\mathbb{P}\Big(\frac{\mathrm{tr}(\boldsymbol{V}_\mu)^2}{n^2} \ge \min\{\frac{\eta^4(1-\gamma)^2}{4}, \frac{\|\psi\|_\infty^2\eta^2(1-\gamma)\gamma}{4\alpha^2}\}\Big) \to 1,$$

which completes the proof. ■

B.4 Proof of Theorem 1

Recall that the goal is to show (28). Putting Lemma 1, Lemma 3, and Lemma 4, Lemma 5 together, we have

$$\begin{aligned}
&\frac{p}{n^2}\|\boldsymbol{\psi}\|^2 - \frac{\mathrm{tr}[\boldsymbol{V}]^2}{n^2}\|\hat{\boldsymbol{h}}\|^2 \qquad (49)\\
&= \Big(\frac{p}{n^2}\|\boldsymbol{\psi}_\mu\|^2 - \frac{\mathrm{tr}[\boldsymbol{V}_\mu]^2}{n^2}\|\hat{\boldsymbol{h}}\|^2\Big) + \frac{p}{n^2}(\|\boldsymbol{\psi}\|_2^2 - \|\boldsymbol{\psi}_\mu\|^2) + \frac{\mathrm{tr}[\boldsymbol{V}_\mu]^2 - \mathrm{tr}[\boldsymbol{V}]^2}{n^2}\|\hat{\boldsymbol{h}}_\mu\|^2 - \frac{\mathrm{tr}[\boldsymbol{V}]^2}{n^2}(\|\hat{\boldsymbol{h}}\|^2 - \|\hat{\boldsymbol{h}}_\mu\|^2)\\
&= O_P(n^{-c}) + o_P(n^{-c/2}) + o_P(n^{-c/2}) - \frac{\mathrm{tr}[\boldsymbol{V}]^2}{n^2}o_P(1)\\
&= o_P(n^{-c/2}) - \frac{\mathrm{tr}[\boldsymbol{V}]^2}{n^2}o_P(1).
\end{aligned}$$

Multiplying $(\mathrm{tr}(\boldsymbol{V})^2/n^2)^{-1}$, which is $O_P(1)$ by Lemma 9, we conclude the proof.

B.5 Proof of Proposition 1

The strategy is exactly the same as for Theorem 1, with Lemma 5 replaced by Lemma 6 to allow for a vanishing smooth component in the noise. More precisely, set $c = 1/4$ so that $\mu = n^{-1/4}$. The conclusions of Lemma 1, Lemma 3 and Lemma 4 still hold for $\epsilon_i = \sigma_n z_i + \delta_i$ with $\sigma_n \to 0$, where α is now the solution to the system (7) with $W \sim \tilde{F}$, since the contribution of terms involving $\sigma_n z_i$ in (42) is 0 due to $\sigma_n \to 0$ and the fact that the loss is Lipschitz by Assumption 1. The condition $\sigma_n \ge \sqrt{\mu}$ of Lemma 6 is satisfied thanks to the assumption $\sigma_n \ge n^{-1/8}$ in Proposition 1. The argument and conclusion of Lemma 9 still hold as well, with $B(\gamma, \rho, F_\epsilon)$ replaced by $B(\gamma, \rho, \tilde{F})$, again because α is now the solution to the system (7) with $W \sim \tilde{F}$. The decomposition (49), and the same argument as for Theorem 1, thus yield the conclusion of Proposition 1.

B.6 Proof of Corollary 1

Fix $\epsilon > 0$. Letting $R_k = \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^\star)\|^2$ for each $k \in [K] := \{1, \dots, K\}$, Theorem 1 or Proposition 1 imply $\mathbb{P}(|R_k - \hat{R}_k| \geq \epsilon/2) \to 0$ for all k. Note that $\hat{R}_{\hat{k}} < \hat{R}_k$ for all k by the definition of $\hat{k}$. Then, we have

$$
\begin{aligned}
\mathbb{P}\big(R_{\hat{k}} > \min_{k\in[K]} R_k + \epsilon\big) &\leq \mathbb{P}(\exists k \in [K], R_{\hat{k}} > R_k + \epsilon) \\
&\leq \sum_{k=1}^K \mathbb{P}(R_{\hat{k}} > R_k + \epsilon) && \text{by the union bound} \\
&= \sum_{k=1}^K \mathbb{P}(R_{\hat{k}} > R_k + \epsilon, \hat{R}_k \geq \hat{R}_{\hat{k}}) && \text{by definition of } \hat{k} \\
&\leq \sum_{k=1}^K \sum_{l=1}^K \mathbb{P}(R_l > R_k + \epsilon, \hat{R}_k > \hat{R}_l) && \text{by the union bound.}
\end{aligned}
$$

Here, $\mathbb{P}(R_l > R_k + \epsilon, \hat{R}_k > \hat{R}_l) \to 0$ as $n \to \infty$ by the following argument:

$$
\begin{aligned}
\mathbb{P}(R_l > R_k + \epsilon, \hat{R}_k > \hat{R}_l) &\leq \mathbb{P}(R_l - \hat{R}_l - (R_k - \hat{R}_k) > \hat{R}_k - \hat{R}_l + \epsilon > \epsilon) \\
&\leq \mathbb{P}(|R_l - \hat{R}_l| > \epsilon/2) + \mathbb{P}(|R_k - \hat{R}_k| > \epsilon/2) \to 0.
\end{aligned}
$$

Since K is finite, we conclude $\mathbb{P}(R_{\hat{k}} > \min_{k\in[K]} R_k + \epsilon) \to 0$. Since $\epsilon > 0$ is taken arbitrarily, this finishes the proof.

Appendix C. Proofs for Section 3

Throughout of this section, we fix (ρ, F_ϵ) that satisfy Assumption 1, Assumption 2 and Assumption 3. For all $\lambda > 0$, define ρ_λ and ψ_λ as

$$
\forall \lambda > 0, \quad \rho_\lambda(\cdot) := \lambda^2 \rho(\cdot/\lambda), \quad \psi_\lambda(\cdot) := \rho'_\lambda(\cdot) = \lambda \psi(\cdot/\lambda).
$$

Note in passing that $\|\psi_\lambda\|_{\text{lip}} = \|\psi\|_{\text{lip}} = 1$ and $\|\psi_\lambda\|_\infty = \lambda\|\psi\|_\infty < +\infty$. Define $R(\lambda)$, $\hat{R}(\lambda)$, and $\alpha(\lambda)$ as

$$
\begin{aligned}
R(\lambda) &:= \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star)\|^2 && \text{with } \hat{\boldsymbol{\beta}}_\lambda \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \\
\hat{R}(\lambda) &:= p\frac{\|\boldsymbol{\psi}_\lambda\|^2}{\operatorname{tr}[\boldsymbol{V}_\lambda]^2} && \text{with } \boldsymbol{V}_\lambda = \frac{\partial \boldsymbol{\psi}_\lambda}{\partial \boldsymbol{y}} \in \mathbb{R}^{n\times n},\ \boldsymbol{\psi}_\lambda := \psi_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda) \\
\alpha(\lambda) &:= \alpha(\rho_\lambda, F_\epsilon, \gamma) && \text{(the solution to (15) with } \rho = \rho_\lambda).
\end{aligned}
$$

C.1 Proof of Proposition 2

$\hat{R}(\lambda) = R(\lambda) + o_P(1)$ by Theorem 1, while $R(\lambda) \to^p \alpha^2(\lambda)$ by Thrampoulidis et al. (2018) and Theorem 7. It remains to show the upper bound of $\alpha^2(\lambda)$. Inequality (37) in Theorem 7 with $\rho = \rho_\lambda$ implies that $\alpha(\lambda)$ is bounded from above as

$$
\alpha(\lambda) \leq \frac{Q_{F_\epsilon}(r^2 c_\gamma)}{r c_\gamma} + \frac{b}{c_\gamma}
$$

where $r \in (0,1]$ and $b \geq 0$ are any constant such that the coercivity condition

$$\|\rho_\lambda\|_{\mathrm{lip}}^{-1}(\rho_\lambda(x) - \rho_\lambda(0)) \geq r(|x| - b) \quad \text{for all } x \in \mathbb{R}$$

is satisfied. Now we verify that

$$r = \frac{\min\{|\psi(\pm 1)|\}}{\|\rho\|_{\mathrm{lip}}} \quad \text{and} \quad b = \lambda \cdot \frac{\max\{|(\mp 1)\psi(\pm 1) - \rho(\pm 1) + \rho(0)|\}}{\min\{|\psi(\pm 1)|\}}$$

satisfy the coercivity condition. By the convexity, evaluating the derivative at λ, using $\rho_\lambda(\cdot) = \lambda^2 \rho(\cdot/\lambda)$ and $\psi_\lambda(\cdot) = \lambda\psi(\cdot/\lambda)$, we have

$$\rho_\lambda(x) \geq \rho_\lambda(\lambda) + (x - \lambda)\psi_\lambda(\lambda) = \lambda\psi(1)x - \lambda^2\psi(1) + \lambda^2\rho(1)$$

for all $x \in \mathbb{R}$. By the same argument, evaluating the derivative at $-\lambda$ gives

$$\rho_\lambda(x) \geq \rho_\lambda(-\lambda) + (x + \lambda)\psi_\lambda(-\lambda) = \lambda\psi(-1)x + \lambda^2\psi(-1) + \lambda^2\rho(-1).$$

Thanks to $\{0\} = \arg\min_x \rho(x)$ by Assumption 1, $\psi(-1) < 0 < \psi(1)$ holds, and hence,

$$\rho_\lambda(x) - \rho_\lambda(0) \geq \lambda \min\Big(|\psi(1)|, |\psi(1)|\Big)|x| - \lambda^2 \max\Big(|\psi(1) - \rho(1) + \rho(0)|, | - \psi(-1) - \rho(-1) + \rho(0)|\Big).$$

Dividing the both sides by $\|\rho_\lambda\|_{\mathrm{lip}} = \lambda\|\rho\|_{\mathrm{lip}}$, we obtain

$$\frac{\rho_\lambda(x) - \rho_\lambda(0)}{\|\rho_\lambda\|_{\mathrm{lip}}} \geq \frac{\min(|\psi(\pm 1)|)}{\|\rho\|_{\mathrm{lip}}}|x| - \lambda\frac{\max(|(\mp)\psi(\pm) - \rho(\pm) + \rho(0)|)}{\|\rho\|_{\mathrm{lip}}}.$$

This means that the pair of (r, b) specified above satisfies the coercivity condition. Substituting this to the upper bound of $\alpha(\lambda)$, since the dependence on λ only comes from b, we obtain

$$\alpha(\lambda) \leq C_7(\rho, F_\epsilon, \gamma)(1 + \lambda),$$

which finishes the proof.

C.2 Proof of Theorem 2

Thanks to Proposition 2, we have

$$\alpha^2(\lambda) - \alpha^2(\tilde{\lambda}) = \hat{R}(\lambda) - \hat{R}(\tilde{\lambda}) + o_P(1),$$

so it suffices to show a suitable smoothness of the map $\lambda \mapsto \hat{R}(\lambda)$.

Lemma 10 *We have*

$$\forall \lambda > 0, \quad |\operatorname{tr}[\boldsymbol{V}_\lambda]| \leq n, \quad \mathbb{P}\left(n^2 \cdot [\operatorname{tr}\boldsymbol{V}_\lambda]^{-2} \leq C_8(\rho, F_\epsilon, \gamma)(1 + \lambda^{-2})\right) \to 1.$$

Proof $\operatorname{tr}[\boldsymbol{V}_\lambda]/n \leq 1$ immediately follows from Proposition 3 and $\|\psi_\lambda\|_{\mathrm{lip}} = \|\psi\|_{\mathrm{lip}} = 1$. Note in passing that $\psi_\lambda(\cdot) = \lambda\psi(\cdot/\lambda)$ satisfies $\|\psi_\lambda\|_\infty = \lambda\|\psi\|_\infty < +\infty$, and for almost every x,

$$\begin{aligned}\psi_\lambda^2(x)/\|\psi_\lambda\|_\infty^2 + \psi_\lambda'(x) &= \psi(x/\lambda)^2/\|\psi\|_\infty + \psi'(x/\lambda) \\ &\geq \eta^2 > 0\end{aligned}$$

thanks to Assumption 1(3) where η is a constant that only depends on ρ. Thus, Lemma 9 with $\rho = \rho_\lambda$ implies

$$\mathbb{P}\Big(\frac{\mathrm{tr}[\boldsymbol{V}_\lambda]^2}{n^2} \geq C(\gamma)\min\Big(\eta^4, \frac{\lambda^2\|\psi\|_\infty^2\eta^2}{\alpha^2(\lambda)}\Big)\Big) \to 1.$$

Putting this lower bound and $\alpha^2(\lambda) \leq C'(\gamma, \rho, F_\epsilon)(\lambda^2+1)$ by Proposition 2 together, we have

$$\frac{\mathrm{tr}[\boldsymbol{V}_\lambda]^2}{n^2} \geq C(\gamma)\min\Big(\eta^4, \frac{\lambda^2\|\psi\|_\infty^2\eta^2}{C'(\rho, F_\epsilon, \gamma)(1+\lambda^2)}\Big) \geq C(\gamma)\min\Big(\eta^4, \frac{\eta^2\|\psi\|_\infty^2}{C'(\rho, F_\epsilon, \gamma)}\Big)\frac{\lambda^2}{1+\lambda^2}$$

with high probability. This finishes the proof. ■

Lemma 11 *We have the followings for all $\lambda, \tilde\lambda > 0$:*

$$\mathbb{P}\Big(\|\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde\lambda}\| \leq \sqrt{n}C_9(\gamma, \rho, F_\epsilon)|\lambda - \tilde\lambda|^{1/2}(1 + \lambda^{1/2} + \tilde\lambda^{1/2})\Big) \to 1,$$
$$\mathbb{P}\Big(|\|\boldsymbol{\psi}_\lambda\|^2 - \|\boldsymbol{\psi}_{\tilde\lambda}\|^2| \leq nC_{10}(\gamma, \rho, F_\epsilon)|\lambda - \tilde\lambda|^{1/2}(1 + \lambda^{3/2} + \tilde\lambda^{3/2})\Big) \to 1.$$

Proof Note that the second display immediately follows from the first display since

$$\begin{aligned}
|\|\boldsymbol{\psi}_\lambda\|^2 - \|\boldsymbol{\psi}_{\tilde\lambda}\|^2| &= |\|\boldsymbol{\psi}_\lambda\| - \|\boldsymbol{\psi}_{\tilde\lambda}\||(\|\boldsymbol{\psi}_\lambda\| + \|\boldsymbol{\psi}_{\tilde\lambda}\|)\\
&\leq |\|\boldsymbol{\psi}_\lambda\| - \|\boldsymbol{\psi}_{\tilde\lambda}\||\sqrt{n}(\|\psi_\lambda\|_\infty + \|\psi_{\tilde\lambda}\|_\infty)\\
&= |\|\boldsymbol{\psi}_\lambda\| - \|\boldsymbol{\psi}_{\tilde\lambda}\||\sqrt{n}\|\psi\|_\infty(\lambda + \tilde\lambda) && \text{by } \|\psi_\lambda\|_\infty = \lambda\|\psi\|_\infty
\end{aligned}$$

It remains to bound $|\|\boldsymbol{\psi}_\lambda\| - \|\boldsymbol{\psi}_{\tilde\lambda}\||$. Below, we write $\boldsymbol{r}_\lambda = \boldsymbol{\epsilon} - \boldsymbol{G}\hat{\boldsymbol{h}}_\lambda$ with $\hat{\boldsymbol{h}}_\lambda = \boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_*)$ so that $\boldsymbol{\psi}_\lambda = \psi_\lambda(\boldsymbol{r}_\lambda)$ and $\|\hat{\boldsymbol{h}}_\lambda\|^2 = R(\lambda)$. By Assumption 3, there exists some constant $c(\rho) > 0$ that only depends on ρ such that for all $\lambda, \tilde\lambda > 0$,

$$\sup_{\boldsymbol{u}\in\mathbb{R}^n} \|\psi_\lambda(\boldsymbol{u}) - \psi_{\tilde\lambda}(\boldsymbol{u})\|_2 \leq \sqrt{n}\cdot\sup_{x\in\mathbb{R}}|\psi_\lambda(x) - \psi_{\tilde\lambda}(x)| \leq \sqrt{n}c(\rho)|\lambda - \tilde\lambda|,$$

so that $\|\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde\lambda}\|$ can be upper bounded from above as

$$\|\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde\lambda}\| \leq \|\psi_\lambda(\boldsymbol{r}_\lambda) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda})\| + \|\psi_\lambda(\boldsymbol{r}_{\tilde\lambda}) - \psi_{\tilde\lambda}(\boldsymbol{r}_{\tilde\lambda})\| \leq \|\psi_\lambda(\boldsymbol{r}_\lambda) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda})\| + \sqrt{n}c(\rho)|\lambda - \tilde\lambda|,$$

Here, using the KKT conditions $\boldsymbol{G}^\top\psi_\lambda(\boldsymbol{r}_\lambda) = \boldsymbol{G}^\top\psi_{\tilde\lambda}(\boldsymbol{r}_{\tilde\lambda}) = \boldsymbol{0}_p$, we bound the first from above as

$$\begin{aligned}
\|\psi_\lambda(\boldsymbol{r}_\lambda) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda})\|^2 &\leq (\psi_\lambda(\boldsymbol{r}_\lambda) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda}))^\top(\boldsymbol{r}_\lambda - \boldsymbol{r}_{\tilde\lambda}) && \text{by Lemma 7 and } \|\psi_\lambda\|_{\mathrm{lip}} = 1\\
&= (\psi_\lambda(\boldsymbol{r}_\lambda) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda}))^\top\boldsymbol{G}(\hat{\boldsymbol{h}}_{\tilde\lambda} - \hat{\boldsymbol{h}}_\lambda)\\
&= (\psi_{\tilde\lambda}(\boldsymbol{r}_{\tilde\lambda}) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda}))^\top\boldsymbol{G}(\hat{\boldsymbol{h}}_{\tilde\lambda} - \hat{\boldsymbol{h}}_\lambda) && \text{by } \boldsymbol{G}^\top\psi_\lambda(\boldsymbol{r}_\lambda) = \boldsymbol{G}^\top\psi_{\tilde\lambda}(\boldsymbol{r}_{\tilde\lambda}) = \boldsymbol{0}_p\\
&\leq \|\psi_{\tilde\lambda}(\boldsymbol{r}_{\tilde\lambda}) - \psi_\lambda(\boldsymbol{r}_{\tilde\lambda})\|\|\boldsymbol{G}\|_{op}(\|\hat{\boldsymbol{h}}_\lambda\| + \|\hat{\boldsymbol{h}}_{\tilde\lambda}\|)\\
&\leq \sqrt{n}c(\rho)|\lambda - \tilde\lambda|\|\boldsymbol{G}\|_{op}(\|\hat{\boldsymbol{h}}_\lambda\| + \|\hat{\boldsymbol{h}}_{\tilde\lambda}\|)
\end{aligned}$$

Therefore, using $\|\hat{\boldsymbol{h}}_\lambda\| \to^p \alpha(\lambda)$ and $\alpha(\lambda) \le C_{11}(\rho, \gamma F_\epsilon)(1+\lambda)$ by Proposition 2, we have

$$
\begin{aligned}
\|\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde{\lambda}}\|^2 &\le 2\sqrt{n} c(\rho)|\lambda - \tilde{\lambda}| \|\boldsymbol{G}\|_{op}(\|\hat{\boldsymbol{h}}_\lambda\| + \|\hat{\boldsymbol{h}}_{\tilde{\lambda}}\|) + 2nc(\rho)^2|\lambda - \tilde{\lambda}|^2 \\
&\le C_{12}(\rho) \cdot n \cdot (n^{-1/2}\|\boldsymbol{G}\|_{op} + 1)\big(|\lambda - \tilde{\lambda}|(\|\hat{\boldsymbol{h}}_\lambda\| + \|\hat{\boldsymbol{h}}_{\tilde{\lambda}}\|) + |\lambda - \tilde{\lambda}|^2\big) \\
&\le C_{13}(\rho, \gamma, F_\epsilon) \cdot n\big(|\lambda - \tilde{\lambda}|(1 + \lambda + \tilde{\lambda}) + |\lambda - \tilde{\lambda}|^2\big) \\
&\le C_{14}(\rho, \gamma, \epsilon) \cdot n|\lambda - \tilde{\lambda}| \cdot (1 + \lambda + \tilde{\lambda})
\end{aligned}
$$

with high probability. This finishes the proof. ■

Lemma 12 *Suppose that (ρ, F_ϵ) satisfy Assumption 1, Assumption 2 and Assumption 3. Then, we have for all $\lambda, \tilde{\lambda} > 0$ that*

$$
\mathbb{P}\Big(n^{-2}|\operatorname{tr}[\boldsymbol{V}_\lambda]^2 - \operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2| \le C_{15}(\rho, F_\epsilon, \gamma)|\lambda - \tilde{\lambda}|^{1/2}(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2})\Big) \to 1,
$$

Proof By the same argument in the proof of Lemma 5, the variant of second order Stein's formula leads to

$$
\operatorname{tr}[\boldsymbol{V}(\lambda)] - \operatorname{tr}[\boldsymbol{V}(\tilde{\lambda})] = \partial_{\boldsymbol{z}}(\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde{\lambda}}) = \phi'(\boldsymbol{z})^\top(\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde{\lambda}}) + O_P(n^{1/2}),
$$

where ϕ is the density function in Assumption 2 such that $\|\phi''\|_\infty < +\infty$. By Lemma 11, Lemma 8, and Cauchy–Schwarz, we have

$$
|\operatorname{tr}[\boldsymbol{V}_\lambda] - \operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]| \le \|\phi'(\boldsymbol{z})\|\|\boldsymbol{\psi}_\lambda - \boldsymbol{\psi}_{\tilde{\lambda}}\| + O_P(\sqrt{n}) \le nC_{16}(\rho, F_\epsilon, \gamma)|\lambda - \tilde{\lambda}|^{1/2}(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2})
$$

with high probability. Combined with $\operatorname{tr}[\boldsymbol{V}_\lambda] \in [0, n]$ from Lemma 10, we conclude the proof. ■

Proof [Proof of Theorem 2] Lemma 10, Lemma 11, and Lemma 12 imply that the followings events hold simultaneously with high probability:

$$
\begin{aligned}
n^2 \operatorname{tr}[\boldsymbol{V}_{\lambda'}]^{-2} &\le C_{17}(\rho, F_\epsilon, \gamma)(1 + \lambda'^{-2}) \qquad \text{for } \lambda' \in \{\lambda, \tilde{\lambda}\}, \\
n^{-1}|\|\boldsymbol{\psi}_\lambda\|^2 - \|\boldsymbol{\psi}_{\tilde{\lambda}}\|^2| &\le C_{18}(\rho, F_\epsilon, \gamma)|\lambda - \tilde{\lambda}|^{1/2}(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2}), \\
n^{-2}\|\operatorname{tr}[\boldsymbol{V}_\lambda]^2 - \operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2| &\le C_{19}(\rho, F_\epsilon, \gamma)|\lambda - \tilde{\lambda}|^{1/2}(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2}).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
|\hat{R}(\lambda) - \hat{R}(\tilde{\lambda})| &= \Big|\frac{p\|\boldsymbol{\psi}_\lambda\|^2}{\operatorname{tr}[\boldsymbol{V}_\lambda]^2} - \frac{p\|\boldsymbol{\psi}_{\tilde{\lambda}}\|^2}{\operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2}\Big| \\
&= \frac{n^2}{\operatorname{tr}[\boldsymbol{V}_\lambda]^2}\frac{n}{p}\Big|\frac{\|\boldsymbol{\psi}_\lambda\|^2}{n} - \frac{\operatorname{tr}[\boldsymbol{V}_\lambda]^2}{\operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2}\frac{\|\boldsymbol{\psi}_{\tilde{\lambda}}\|^2}{n}\Big| \\
&\le \frac{n^2}{\operatorname{tr}[\boldsymbol{V}_\lambda]^2}\frac{n}{p}\Big(\frac{|\|\boldsymbol{\psi}_\lambda\|^2 - \|\boldsymbol{\psi}_{\tilde{\lambda}}\|^2|}{n} + \frac{n^2}{\operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2}\frac{|\operatorname{tr}[\boldsymbol{V}_\lambda]^2 - \operatorname{tr}[\boldsymbol{V}_{\tilde{\lambda}}]^2|}{n^2}\frac{\|\boldsymbol{\psi}_{\tilde{\lambda}}\|^2}{n}\Big) \\
&\le C(\rho, \gamma, F_\epsilon)|\lambda - \tilde{\lambda}|^{1/2}(1 + \lambda^{-2})\Big(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2} + (1 + \tilde{\lambda}^{-2})(1 + \lambda^{1/2} + \tilde{\lambda}^{1/2})\tilde{\lambda}\Big)
\end{aligned}
$$

Combined with $\hat{R}(\lambda) \to^p \alpha^2(\lambda)$ for each $\lambda \in (0, \infty)$, we have the following for all $\lambda, \tilde{\lambda} > 0$:

$$|\alpha^2(\lambda) - \alpha^2(\tilde{\lambda})| \leq C(\rho, \gamma, F_\epsilon)|\lambda - \tilde{\lambda}|^{1/2} L(\lambda, \tilde{\lambda}),$$

where $L(\lambda, \tilde{\lambda}) := (1+\lambda^{-2})\big(1+\lambda^{1/2}+\tilde{\lambda}^{1/2}+(1+\tilde{\lambda}^{-2})(1+\lambda^{1/2}+\tilde{\lambda}^{1/2})\tilde{\lambda}\big)$. Since $\sup_{\lambda,\tilde{\lambda}\in[\lambda_{\min},\lambda_{\max}]} L(\lambda, \tilde{\lambda})$ is bounded from above by some positive constant $C(\lambda_{\min}, \lambda_{\max})$, we conclude the proof. ■

C.3 Proof of Theorem 3

Lemma 13 *Fix $I = [\lambda_{\min}, \lambda_{\max}]$ for some $0 < \lambda_{\min} < \lambda_{\max} < +\infty$. For all $N \in \mathbb{N}$, let $I_N = \{\lambda_{\min}^{i/N} \cdot \lambda_{\max}^{1-i/N} : i = 0, 1, \ldots N\}$ be the finite grid over I and take $\hat{\lambda}_N \in \arg\min_{\lambda \in I_N} \hat{R}(\lambda)$. Then, for all $\delta > 0$, there exists $N_\delta = N(\rho, F_\epsilon, \gamma, \lambda_{\min}, \lambda_{\max}, \delta) \in \mathbb{N}$ such that*

$$\lim_{n\to\infty} \mathbb{P}(|\alpha^2(\hat{\lambda}_{N_\delta}) - \alpha^2(\lambda_{\text{opt}})| < \delta) = \lim_{n\to\infty} \mathbb{P}(|R(\hat{\lambda}_{N_\delta}) - \alpha^2(\lambda_{\text{opt}})| < \delta) \to 1,$$

where $\hat{\lambda}_{N_\delta} \in \arg\min_{\lambda\in I_{N_\delta}} \hat{R}(\lambda)$ and $\lambda_{\text{opt}} \in \arg\min_{\lambda\in I} \alpha^2(\lambda)$.

Proof [Proof of Lemma 13] First, we prove $\mathbb{P}(|\alpha^2(\hat{\lambda}_{N_\delta}) - \alpha^2(\lambda_{\text{opt}})| < \delta) \to 1$. Let us take some $N \in \mathbb{N}$ to be specified later. Consider the decomposition

$$0 < \alpha^2(\hat{\lambda}_N) - \alpha^2(\lambda_{\text{opt}}) = \alpha^2(\hat{\lambda}_N) - \min_{\lambda\in I_N} \alpha^2(\lambda) + \min_{\lambda\in I_N} \alpha^2(\lambda) - \alpha^2(\lambda_{\text{opt}}).$$

Now, the cardinality of I_N is finite and $\hat{R}(\lambda) = \alpha^2(\lambda) + o_P(1)$ for all $\lambda > 0$ by Proposition 2. Then, by the same argument of Appendix B.6, where the role of R is replaced by α^2, we have

$$\alpha^2(\hat{\lambda}_N) - \min_{\lambda\in I_N} \alpha^2(\lambda) = o_P(1),$$

so in particular $\lim_{n\to\infty} \mathbb{P}(\alpha^2(\hat{\lambda}_N) - \min_{\lambda\in I_N} \alpha^2(\lambda) < \delta/2) = 1$. Below, we show $\min_{\lambda\in I_N} \alpha^2(\lambda) - \alpha^2(\lambda_{\text{opt}}) \leq \delta/2$ for some $N = N_\delta$. By the definition of I_N, we can take $\lambda_N^\star \in I_N$ such that $\lambda_N^\star \leq \lambda_{\text{opt}} < (\lambda_{\max}/\lambda_{\min})^{1/N} \lambda_N^\star$, so that

$$|\lambda_{\text{opt}} - \lambda_N^\star| \leq [(\lambda_{\max}/\lambda_{\min})^{1/N} - 1]\lambda_N^\star \leq [(\lambda_{\max}/\lambda_{\min})^{1/N} - 1]\lambda_{\max}.$$

On the other hand, Theorem 2 implies that

$$\text{for all } \lambda, \tilde{\lambda} \in I = [\lambda_{\min}, \lambda_{\max}], \quad |\alpha^2(\lambda) - \alpha^2(\tilde{\lambda})| \leq L|\lambda - \tilde{\lambda}|^{1/2},$$

where $L = L(\rho, F_\epsilon, \gamma, \lambda_{\min}, \lambda_{\max}) > 0$ is some positive constant. From the above displays, it follows that

$$\begin{aligned}
0 &< \min_{\lambda\in I_N} \alpha^2(\lambda) - \alpha^2(\lambda_{\text{opt}}) && \text{by } \lambda_{\text{opt}} \in \arg\min_{\lambda\in I} \alpha^2(\lambda) \text{ and } I_N \subset I \\
&\leq \alpha^2(\lambda_N^\star) - \alpha^2(\lambda_{\text{opt}}) && \text{by } \lambda_N^* \in I_N \\
&\leq L|\lambda_N^\star - \lambda_{\text{opt}}|^{1/2} && \text{by } \lambda_N^\star, \lambda_{\text{opt}} \in I \\
&\leq L\big\{\lambda_{\max}[(\lambda_{\max}/\lambda_{\min})^{1/N} - 1]\big\}^{1/2} && \text{by } |\lambda_{\text{opt}} - \lambda_N^\star| \leq [(\lambda_{\max}/\lambda_{\min})^{1/N} - 1]\lambda_{\max} \\
&\leq \delta/2 && \text{if } (\lambda_{\max}/\lambda_{\min})^{1/N} \leq 1 + \delta^2/(4L\lambda_{\max}).
\end{aligned}$$

Therefore, if we take $N = N_\delta := \lceil \log(\lambda_{\max}/\lambda_{\min})/\log\{1+\delta^2/(4L\lambda_{\max})\}\rceil$, we obtain $\mathbb{P}(|\alpha^2(\hat\lambda_{N_\delta}) - \alpha^2(\lambda_{\text{opt}})| < \delta) \to 1$.

Next, we prove $\mathbb{P}(|R(\hat\lambda_N) - \alpha^2(\lambda_{\text{opt}})| \le \epsilon) \to 1$. By the triangle inequality, it follows that

$$|R(\hat\lambda_N) - \alpha^2(\lambda_{\text{opt}})| \le |R(\hat\lambda_N) - \min_{\lambda\in I_N} R(\lambda)| + |\min_{\lambda\in I_N} R(\lambda) - \min_{\lambda\in I_N} \alpha^2(\lambda)| + |\min_{\lambda\in I_N} \alpha^2(\lambda) - \alpha^2(\lambda_{\text{opt}})|.$$

The first term is $o_P(1)$ by Corollary 1, while the second term is $o_P(1)$ as $R(\lambda) \to^p \alpha^2(\lambda)$ for all $\lambda > 0$ and the cardinality of I_N is finite. For the third term, we have shown that it is less than $\delta/2$ for $N = N_\delta$. This completes the proof of $\mathbb{P}(|R(\hat\lambda_{N_\delta}) - \alpha^2(\lambda_{\text{opt}})| \le \delta) \to 1$. ■

Proof [Proof of Theorem 3] Our goal is to construct an array $(N_n)_{n=1}^\infty$ of integers such that

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(|\alpha^2(\hat\lambda_{N_n}) - \alpha^2(\lambda_{\text{opt}})| > \epsilon) + \mathbb{P}(|R(\hat\lambda_{N_n}) - \alpha^2(\lambda_{\text{opt}})| > \epsilon) = 0. \tag{50}$$

By Lemma 13 with $\delta = 1/k$, there exists a function $N : \mathbb{N} \to \mathbb{N}$ such that

$$\forall k \in \mathbb{N}, \quad \lim_{n\to\infty} \underbrace{\mathbb{P}(|\alpha^2(\hat\lambda_{N(k)}) - \alpha^2(\lambda_{\text{opt}})| \ge 1/k) + \mathbb{P}(|R(\hat\lambda_{N(k)}) - \alpha^2(\lambda_{\text{opt}})| \ge 1/k)}_{:=p_{n,k}} = 0. \tag{51}$$

Thus, if we can find some function $\varphi : \mathbb{N} \to \mathbb{N}$ such that $\lim_{n\to\infty} \varphi(n) \to \infty$ and $\lim_{n\to\infty} p_{n,\varphi(n)} = 0$, the array $\{N_n\}_{n=1}^\infty := \{N \circ \varphi(n)\}_{n=1}^\infty$ satisfies (50), thereby completing the proof. Below we construct such $n \mapsto \varphi(n)$.

By (51), there exists an array of integers $\{n_k\}_{k=1}^\infty$ such that

$$\forall k \in \mathbb{N}, \quad \forall n \ge n_k, \quad |p_{n,k}| \le 1/k.$$

Here, we assume without of loss of generality that $\{n_k\}_k$ is strictly increasing; otherwise redefine recursively as $\tilde n_k := \max(\tilde n_{k-1}, n_k) + 1$. For this $\{n_k\}_{k=1}^\infty$, define the step function $\varphi : \mathbb{N} \to \mathbb{N}$ as $\varphi(n) := \sum_{k=1}^\infty k\mathbf{1}\{n_k \le n < n_{k+1}\}$. Then, we have $\lim_{n\to\infty} \varphi(n) = +\infty$, and $|p_{n,k}| \le 1/k$ holds for $k = \varphi(n)$ since $n \ge n_k$ by construction. Therefore, we obtain

$$\forall n \ge n_1, \quad |p_{n,\varphi(n)}| \le 1/\varphi(n).$$

Since $\varphi(n) \to +\infty$ as $n \to \infty$, we have $\lim_{n\to\infty} p_{n,\varphi(n)} = 0$. This finishes the proof. ■

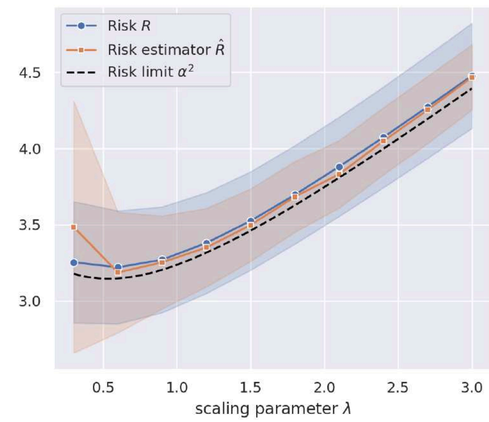
Appendix D. Additional numerical simulation

D.1 Relaxing Assumption 2

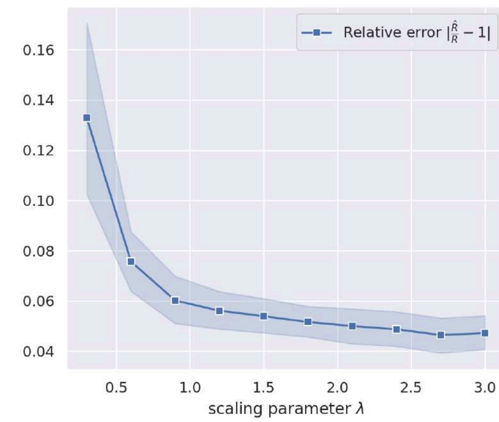
In the same setting as Section 4, we change the noise distribution to be the following discrete distribution so that Assumption 2 is not satisfied:

$$\epsilon_i \overset{\text{iid}}{\sim} 3 \cdot \lceil \text{t-dist}(\text{df} = 2) \rceil$$

where $\lceil x \rceil = \max\{n \in \mathbb{Z} : n \le x\}$. Figure 5 implies that our result still holds in this setting.



(a) $R(\lambda)$, $\hat{R}(\lambda)$ and $\alpha^2(\lambda)$.

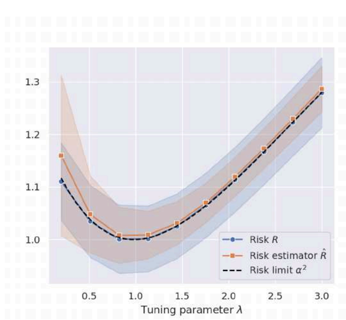


(b) Relative error $|\hat{R}(\lambda)/R(\lambda) - 1|$

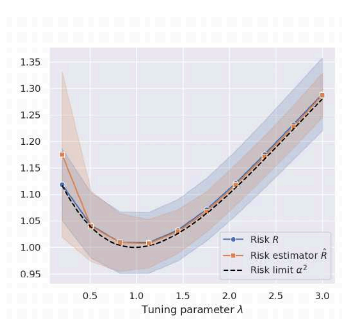
Figure 5: Plot of the out-of-sample error $R(\lambda)$ and estimator $\hat{R}(\lambda)$ over 100 repetitions, with $n = 4000$, $p = 1200$, the Huber loss for different values of the scale parameters λ. The noise distribution is $3\lceil \text{t-dist}(\text{df} = 2) \rceil$. $\alpha^2(\lambda)$ is the theoretical limit given by the nonlinear system (7).

D.2 Robustness of risk estimation across covariate distributions

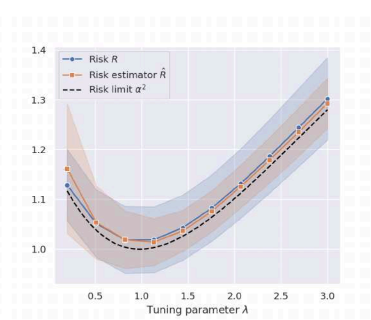
We vary the distribution of the covariate $\boldsymbol{X}$ from Gaussian to Rademacher, uniform and Student's t-distribution in Figure 6. In each of these settings, we confirm the effectiveness of the risk estimator, suggesting that our results extend to a broader class of covariate distributions.



(a) Rademacher



(b) Uniform $[-1, 1]$



(c) T-dist with $\text{df} = 4$

Figure 6: We change the distribution of the design matrix $\boldsymbol{X}$ to Uniform, Student's T, and Rademacher distributions with proper normalization so that the variance is 1. Simulation parameter: the noise distribution $F_\epsilon = \text{t-dist}(\text{df} = 2)$, sample size $n = 4000$, feature $p = 1200$, scaling parameter $\lambda = 1$, 100 repetition.

D.3 Risk behavior as dimensionality ratio $\gamma = p/n$ approaches 1

We plot the risk and its estimator as $\gamma = p/n$ varies in Figure 7. The results show that the risk estimation becomes increasingly unstable as $\gamma \to 1^-$.

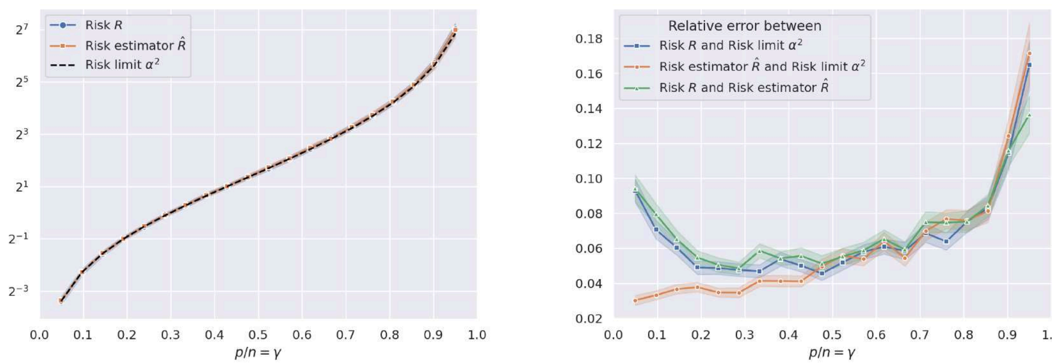


Figure 7: We plot the risk and its estimator for different ratio $\gamma = p/n$. Simulation parameters are as follows: sample size $n = 4000$, noise distribution F_ϵ = t-dist(df = 2), scaling parameter $\lambda = 1$, with 100 repetition. The relative error between A and B is $|A/B - 1|$.

D.4 Convergence of risk estimation variance with increasing sample size

We plot the risk and its estimator as n varies for different values of $\gamma = p/n$: $\gamma = 0.25$ in Figure 8 and $\gamma = 0.5$ in Figure 9. In both cases, the variances of the risk and risk estimator decrease as $n \to +\infty$. Interestingly, these figures suggest that the variances decays in $\sqrt{n}$-rate.

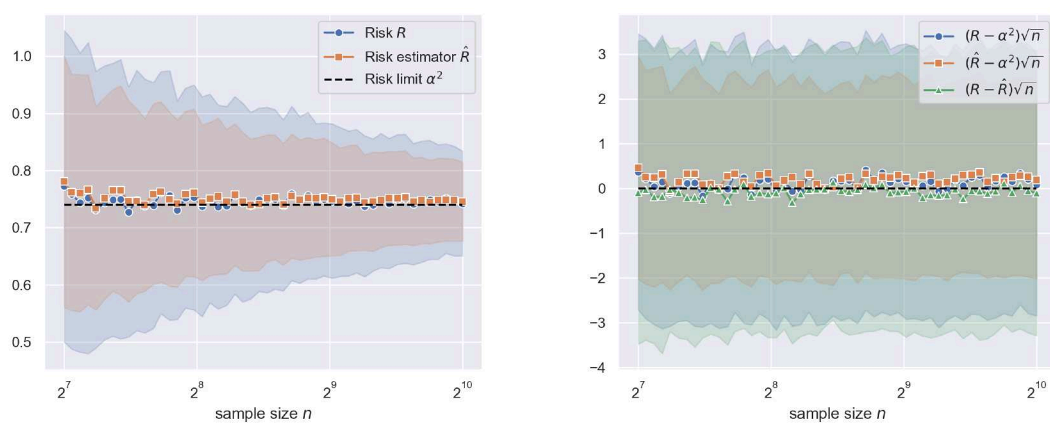


Figure 8: We plot the risk and its estimator for different n. Simulation parameters are as follows: $p/n = \gamma = 0.25$, F_ϵ = t-dist(df = 2), scaling parameter $\lambda = 1$, with 1000 repetition.

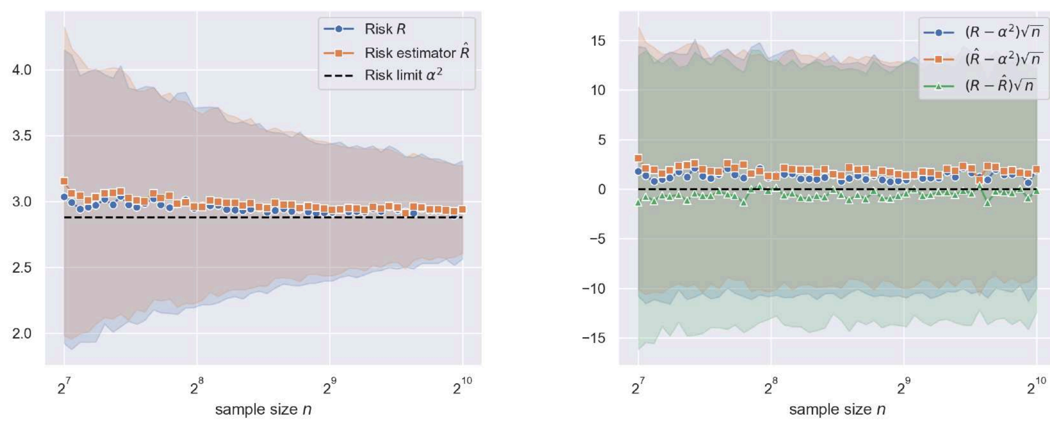


Figure 9: We plot the risk and its estimator for different n. Simulation parameters are as follows: $p/n = \gamma = 0.5$, $F_\epsilon =$ t-dist(df $= 2$), scaling parameter $\lambda = 1$, with 1000 repetition.