# Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning

**Pierre C. Bellec** PIERRE.BELLEC@RUTGERS.EDU and **Yiwei Shen** YS573@STAT.RUTGERS.EDU
*Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.*

**Editors:** Po-Ling Loh and Maxim Raginsky

## Abstract

This paper studies M-estimators with gradient-Lipschitz loss function regularized with convex penalty in linear models with Gaussian design matrix and arbitrary noise distribution. A practical example is the robust M-estimator constructed with the Huber loss and the Elastic-Net penalty and the noise distribution has heavy-tails. Our main contributions are three-fold. (i) We provide general formulae for the derivatives of regularized M-estimators $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ where differentiation is taken with respect to both $\boldsymbol{y}$ and $\boldsymbol{X}$; this reveals a simple differentiability structure shared by all convex regularized M-estimators. (ii) Using these derivatives, we characterize the distribution of the residual $r_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$ in the intermediate high-dimensional regime where dimension and sample size are of the same order. (iii) Motivated by the distribution of the residuals, we propose a novel adaptive criterion to select tuning parameters of regularized M-estimators. The criterion approximates the out-of-sample error up to an additive constant independent of the estimator, so that minimizing the criterion provides a proxy for minimizing the out-of-sample error. The proposed adaptive criterion does not require the knowledge of the noise distribution or of the covariance of the design. Simulated data confirms the theoretical findings, regarding both the distribution of the residuals and the success of the criterion as a proxy of the out-of-sample error. Finally our results reveal new relationships between the derivatives of $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ and the effective degrees of freedom of the M-estimator, which are of independent interest.

**Keywords:** Robust estimation, M-estimator, Adaptive tuning, High-dimensional statistics, Residual distribution.

## 1. Introduction

This paper studies properties of robust estimators in linear models $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with response $\boldsymbol{y} \in \mathbb{R}^n$, unknown regression vector $\boldsymbol{\beta}^*$ and $\boldsymbol{X}$ is a design matrix with $n$ rows $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Each row $\boldsymbol{x}_i$ being a high-dimensional feature vector in $\mathbb{R}^p$, centered and normally distributed with covariance $\boldsymbol{\Sigma}$, and each $\varepsilon_i$ is independent of $\boldsymbol{X}$ with continuous distribution. Throughout, let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ be a regularized $M$-estimator given as a solution of the convex minimization problem

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + g(\boldsymbol{b}) \tag{1.1}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a convex data-fitting loss function and $g : \mathbb{R}^p \to \mathbb{R}$ a convex penalty. We may write $\widehat{\boldsymbol{\beta}}_{\rho,g}(\boldsymbol{y}, \boldsymbol{X})$ for (1.1) to emphasize the dependence on the loss-penalty pair $(\rho, g)$; if the argument $(\boldsymbol{y}, \boldsymbol{X})$ is dropped then $\widehat{\boldsymbol{\beta}}$ is implicitly understood at the observed that $(\boldsymbol{y}, \boldsymbol{X})$. Typical examples of losses include the square loss $\rho(u) = u^2/2$, the Huber loss $H(u) = \int_0^{|u|} \min(1, t) dt$ or its scaled version $\rho = \Lambda^2 H(u/\Lambda)$ for some tuning parameter $\Lambda > 0$, while typical examples of penalty functions include the Elastic-Net $g(\boldsymbol{b}) = \lambda \|\boldsymbol{b}\|_1 + \mu \|\boldsymbol{b}\|^2/2$ for tuning parameters $\lambda, \mu \geq 0$.

The paper introduces the following criterion to select a loss-penalty pair $(\rho, g)$ with small out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2$: for a given set of candidate loss-penalty pairs $\{(\rho, g)\}$ and the corresponding $M$-estimator $\widehat{\boldsymbol{\beta}}_{\rho,g}$ in (1.1), select the pair $(\rho, g)$ that minimizes the criterion

$$\mathrm{Crit}(\rho, g) = \Big\| \boldsymbol{r} + \frac{\hat{\mathsf{df}}}{\mathrm{tr}[\boldsymbol{V}]} \psi(\boldsymbol{r}) \Big\|^2 \quad \text{with} \quad \begin{cases} \boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\rho,g} & \in \mathbb{R}^n, \\ \hat{\mathsf{df}} = \mathrm{tr}[\boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}_{\rho,g}] & \in \mathbb{R}, \\ \boldsymbol{V} = \mathrm{diag}\{\psi'(\boldsymbol{r})\}(\boldsymbol{I}_n - \boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}_{\rho,g}) & \in \mathbb{R}^{n\times n} \end{cases} \tag{1.2}$$

where $\mathrm{tr}[\cdot]$ is the trace, $\psi : \mathbb{R} \to \mathbb{R}$ is the derivative of $\rho$, $\psi'$ the derivative of $\psi$ and we extend $\psi$ and $\psi'$ to functions $\mathbb{R}^n \to \mathbb{R}^n$ by componentwise application of the univariate function of the same symbol. Above, $(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}_{\rho,g} \in \mathbb{R}^{p\times n}$ denotes the Jacobian of (1.1) with respect to $\boldsymbol{y}$ for $\boldsymbol{X}$ fixed, at the observed data $(\boldsymbol{y}, \boldsymbol{X})$. As we will see while studying particular examples, for pairs $(\rho, g)$ commonly used in robust high-dimensional statistics such as the square loss, Huber loss with the $\ell_1$-penalty or Elastic-Net penalty, the ratio $\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}]$ in (1.2) admits simple, closed-form expressions. The criterion (1.2) has an appealing adaptivity property: it does not require any knowledge of the noise $\boldsymbol{\varepsilon}$ or its distribution, nor any knowledge of the covariance $\boldsymbol{\Sigma}$ of the design.

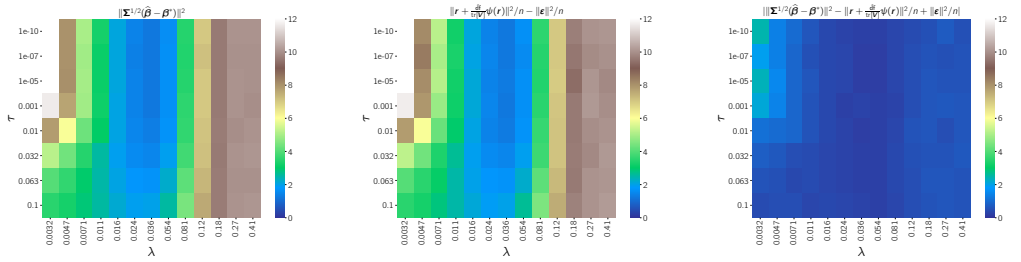


Figure 1: Heatmaps for $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2$, its approximation $\|\boldsymbol{r} + (\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}])\psi(\boldsymbol{r})\|^2/n - \|\boldsymbol{\varepsilon}\|^2/n$ and the approximation error $|\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 - \|\boldsymbol{r} + (\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}])\psi(\boldsymbol{r})\|^2/n - \|\boldsymbol{\varepsilon}\|^2/n|$ for the Huber loss and Elastic-Net penalty on a grid of tuning parameters $(\lambda, \tau)$ where $\lambda \in [0.0032, 0.41]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is the average over 100 repetitions. See Section 6 for more details.

### 1.1. Contributions

1. The end goal of this paper is to provide theoretical justification and theoretical guarantees for the criterion (1.2) in the high-dimensional regime where the ratio $p/n$ has a finite limit and $\boldsymbol{X}$ has anisotropic Gaussian distribution. The theoretical results will justify the approximation

$$\big\| \boldsymbol{r} + \big(\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}]\big)\psi(\boldsymbol{r}) \big\|^2/n \approx \|\boldsymbol{\varepsilon}\|^2/n + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2. \tag{1.3}$$

Figure 1 illustrates the accuracy of (1.3) on simulated data. To study the criterion (1.2) and derive the approximation (1.3), we develop novel results of independent interest regarding $M$-estimators in (1.1):

2. The paper derives general formula for the derivatives $(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}$ and $(\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}$. This sheds light on the differentiability structure of $M$-estimators for general loss-penalty pairs: for any $\rho, g$

with $g$ strongly convex, there exists $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ depending on $(\boldsymbol{y}, \boldsymbol{X})$ such that for almost every $(\boldsymbol{y}, \boldsymbol{X})$,

$$(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i), \quad (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = \widehat{\boldsymbol{A}}\boldsymbol{e}_j \psi(r_i) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i)\widehat{\beta}_j,$$

for $r_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$, $\forall i \in [n]$,$j \in [p]$ where $\boldsymbol{e}_j \in \mathbb{R}^p$ and $\boldsymbol{e}_i \in \mathbb{R}^n$ are canonical basis vectors.

3. The paper obtains a stochastic representation for the residual $y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$ for some fixed $i = 1, ..., n$, extending some results of El Karoui et al. (2013) on unregularized $M$-estimators to penalized ones as in (1.1). In short, for each $i = 1, ..., n$ the $i$-th residual satisfies $r_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$

$$r_i + \frac{\hat{\mathrm{df}}}{\mathrm{tr}[\boldsymbol{V}]}\psi(r_i) \approx \varepsilon_i + Z_i \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| \tag{1.4}$$

where $Z_i \sim N(0, 1)$ is independent of $\varepsilon_i$. This stochastic representation is the motivation for the criterion (1.2) as the amplitude of the normal part in the right-hand side is proportional to the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|$ that we wish to minimize, while the variance of the noise $\varepsilon_i$ does not depend on the choice of $(\rho, g)$.

Simulated data in Figure 2 confirms that the stochastic representation for the $i$-th residual $r_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$ is accurate. Our working assumption throughout the paper is the following.

**Assumption A** *For constants $\gamma, \mu > 0$ independent of $n, p$ we have $p/n \le \gamma$, the loss $\rho : \mathbb{R} \to \mathbb{R}$ is convex with a unique minimizer at 0, continuously differentiable and its derivative $\psi = \rho'$ is 1-Lipschitz. The design matrix $\boldsymbol{X}$ has iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows for some invertible covariance $\boldsymbol{\Sigma}$ and the noise $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ with continuous distribution. The penalty $g : \mathbb{R}^p \to \mathbb{R}$ is $\mu$-strongly convex w.r.t. $\boldsymbol{\Sigma}$ in the sense that $\boldsymbol{b} \mapsto g(\boldsymbol{b}) - (\mu/2)\boldsymbol{b}^\top \boldsymbol{\Sigma}\boldsymbol{b}$ is convex in $\boldsymbol{b} \in \mathbb{R}^p$.*

Throughout the paper, we consider a sequence (say, indexed by $n$) of regression problems with $p$, $\boldsymbol{\beta}^*$, $\boldsymbol{\Sigma}$ and the loss-penalty pair $(\rho, g)$ depending implicitly on $n$. For some deterministic sequence $(a_n)$, the stochastically bounded notation $O_P(a_n)$ in this context may hide constants depending on $\gamma, \mu$ only, that is, $O_P(a_n)$ denotes a sequence of random variables $W_n$ such that for any $\epsilon > 0$ there exists $K$ depending on $(\epsilon, \gamma, \mu)$ satisfying $\mathbb{P}(|W_n| \ge K a_n) \le \epsilon$.

Since Assumption A requires $p/n \le \gamma$, the Bolzano-Weierstrass theorem lets us extract a subsequence of regression problems such that $p/n \to \gamma'$ along this subsequence, for some constant $\gamma'$. This is the asymptotic regime we have in mind throughout the paper, although our results do not require a specific limit for the ratio $p/n$. For some results, we will require the following additional assumption which is satisfied by robust loss functions and penalty that shrink towards 0.

**Assumption B** *The penalty is minimized at $\boldsymbol{0}$, that is, $g(\boldsymbol{0}) = \min_{\boldsymbol{b}\in\mathbb{R}^p} g(\boldsymbol{b})$; the loss is Lipschitz as in $|\psi| \le M$ for some constant $M$ independent of $n, p$; the signal is bounded as in $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*\|^2 \le M$.*

### 1.2. Related works

The context of the present work is the study of $M$-estimators in the regime $\frac{p}{n}$ has a finite limit. This literature pioneered in Bayati and Montanari (2012); El Karoui et al. (2013); Donoho and Montanari (2016); Stojnic (2013) typically describes the subtle behavior of $\widehat{\boldsymbol{\beta}}$ in this regime by solving a system of nonlinear equations. This system depends on a prior distribution for the components of $\boldsymbol{\beta}^*$, and

either depends on the covariance $\boldsymbol{\Sigma}$ (Celentano et al., 2020; Dobriban and Wager, 2018) or assume $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ (Bayati and Montanari, 2012; Thrampoulidis et al., 2018; Celentano and Montanari, 2019, among many others). Solutions to the nonlinear system are a powerful tool to understand $\widehat{\boldsymbol{\beta}}$ in theory, e.g., to characterize the deterministic limit of $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|$, see e.g., the general results in Celentano and Montanari (2019) for the square loss and Thrampoulidis et al. (2018) for general loss-penalty pairs. However, since the system and its solution depend on unobservable quantities ($\boldsymbol{\Sigma}$ and prior on $\boldsymbol{\beta}^*$), the system solution is not directly usable for practical purposes such as parameter tuning.

The present work distinguishes itself from most of this literature as the goal is to describe the behavior of $\widehat{\boldsymbol{\beta}}$ using observable quantities that only depend on the data $(\boldsymbol{y}, \boldsymbol{X})$ (and not unobservable ones such as $\boldsymbol{\Sigma}$ or a prior distribution on $\boldsymbol{\beta}^*$ that appear in the aforementioned nonlinear system of equations). As we will see this view lets us perform adaptive tuning of parameters in a fully adaptive manner using the criterion (1.2). The criterion (1.2) appeared in previous works for the square loss only: Bayati et al. (2013); Miolane and Montanari (2018) studied (1.2) for the Lasso with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and (Bellec, 2020, Section 3) for the square loss and general penalty (note that for the square loss $\rho(u) = u^2/2$, (1.2) reduces to $n^2\|\boldsymbol{r}\|^2/(n-\hat{\mathsf{df}})^2$ due to $\psi(u) = u$ and $\mathrm{tr}[\boldsymbol{V}] = n - \hat{\mathsf{df}}$. The property $\psi(u) = u$ of the square loss hides the subtle interplay between $\boldsymbol{r}, \psi(\boldsymbol{r}), \hat{\mathsf{df}}$ and $\mathrm{tr}[\boldsymbol{V}]$ in (1.2) for $\rho$ different than the square loss).

A criterion different from (1.2) is studied in Bayati et al. (2013); Miolane and Montanari (2018) (for the Lasso and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$) and Bellec (2020) (for general loss-penalty pairs), with the purpose of estimating the out-of-sample error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2$. In the case of an M-estimator and with the notation in (1.2), this criterion is the right-hand side of the approximation

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 \approx \mathrm{tr}[\boldsymbol{V}]^{-2}\big(\|\psi(\boldsymbol{r})\|^2(2\hat{\mathsf{df}} - p) + \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}^\top\psi(\boldsymbol{r})\|^2\big).$$

That criterion from Bayati et al. (2013); Miolane and Montanari (2018); Bellec (2020) has the drawback to require the knowledge of the covariance $\boldsymbol{\Sigma}$, and is thus not readily usable unless $\boldsymbol{\Sigma}$ is known or can be consistently estimated. On the other hand, the criterion (1.2) is fully adaptive: it does not depend on $\boldsymbol{\Sigma}$, and can thus be used even if $\boldsymbol{\Sigma}$ cannot itself be consistently estimated. Another line of work (Rad et al., 2020; Xu et al., 2021; Rad and Maleki, 2020) proposed the ALO criterion

$$\sum_{i=1}^n \Big(r_i + \frac{H_{ii}}{V_{ii}}\psi(r_i)\Big)^2 \tag{1.5}$$

(when specialized to linear models), where $\boldsymbol{V}$ is the matrix defined in (1.2) and $H_{ii} = \boldsymbol{x}_i^\top \frac{\partial}{\partial y_i}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ in the notation of the present paper. This criterion differs from (1.2) proposed in the present paper, in that (1.2) replaces $H_{ii}$ and $V_{ii}$ by their respective averages, $\hat{\mathsf{df}}/n = \frac{1}{n}\sum_{i=1}^n H_{ii}$ and $\mathrm{tr}[\boldsymbol{V}]/n = \frac{1}{n}\sum_{i=1}^n V_{ii}$. This extra averaging step lets us prove, for non-smooth penalty functions, theoretical guarantees for selecting a loss-penalty pair by minimizing the criterion (1.2) (cf. Sections 4 and 5 below). On the other hand, we are not aware of similar theoretical guarantees for selecting a loss-penalty pair based on (1.5), since the theoretical analysis of (1.5) is so far restricted to twice continuously differentiable loss and penalty functions, with a uniform upper bound on the Lipschitz constant of the second derivatives (Rad and Maleki, 2020, Assumption 6 required in Theorem 3 and Corollary 1). This rules out the Elastic-Net and other non-smooth penalty functions typically used for high-dimensional data, as well as the Huber loss which is not twice continuously

differentiable. The criterion (1.2) of the present paper thus improves upon (1.5) since it enjoys theoretical guarantees for non-smooth penalty functions and the Huber loss.

This work leverages probabilistic results on functions of standard normal random variables Bellec and Zhang (2019)(Bellec, 2020, §6, §7) which are consequences of Stein's formula Stein (1981). Consequently, the main limitation of our work is that it currently requires Gaussian design for the probabilistic results, although simulations in Appendix G suggest that the results hold for more general distributions, including design with Rademacher entries. On the other hand, the differentiability result (2.1) is deterministic and does not rely on any probabilistic assumption.

## 2. Differentiability of regularized M-estimators

The first step towards the study of the criterion (1.2) is to justify the almost sure existence of the derivatives of $\widehat{\boldsymbol{\beta}}$ that appear in (1.2) through the scalar scalar $\hat{\mathsf{df}}$ and the matrix $\boldsymbol{V}$ in (1.2). Although the criterion (1.2) only involves the derivatives of $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ with respect to $\boldsymbol{y}$ for a fixed $\boldsymbol{X}$, the proof of our results rely on the interplay between the derivatives with respect to $\boldsymbol{y}$ and with respect to $\boldsymbol{X}$: this *differentiability structure* of $M$-estimators is the content of the following result.

**Theorem 1** *Let Assumption A be fulfilled. For almost every $(\boldsymbol{y}, \boldsymbol{X})$ the map $(\boldsymbol{y}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is differentiable at $(\boldsymbol{y}, \boldsymbol{X})$ and there exists a matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ depending on $(\boldsymbol{X}, \boldsymbol{y})$ with $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le (n\mu)^{-1}$ s.t.*

$$\begin{aligned}(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i),\\ (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{e}_j \psi(r_i) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i)\widehat{\beta}_j,\end{aligned} \qquad \text{where } r_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}, \tag{2.1}$$

*$\boldsymbol{e}_i \in \mathbb{R}^n, \boldsymbol{e}_j \in \mathbb{R}^p$ are canonical basis vectors , $\psi := \rho'$ and $\psi'$ denote the derivatives. Furthermore,*

$$\hat{\mathsf{df}} = \mathrm{tr}[\boldsymbol{X}(\partial/\partial \boldsymbol{y})\hat{\boldsymbol{\beta}}] = \mathrm{tr}[\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top \mathrm{diag}\{\psi'(\boldsymbol{r})\}], \tag{2.2}$$

$$\boldsymbol{V} = \mathrm{diag}\{\psi'(\boldsymbol{r})\}(\boldsymbol{I}_n - \boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}) = \mathrm{diag}\{\psi'(\boldsymbol{r})\} - \mathrm{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top \mathrm{diag}\{\psi'(\boldsymbol{r})\} \tag{2.3}$$

*satisfy $0 \le \hat{\mathsf{df}} \le n$ and $0 \le \mathrm{tr}[\boldsymbol{V}] \le n$.*

Since the same matrix $\widehat{\boldsymbol{A}}$ appears in both the derivatives with respect to $y_i$ and to $x_{ij}$, (2.1) provides relationship between $(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}$ and $(\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}$, for instance $(\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{A}}\boldsymbol{e}_j\psi(r_i) - \widehat{\beta}_j(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}$. Although the matrix $\widehat{\boldsymbol{A}}$ is not explicit for arbitrary loss-penalty pair, closed-form expressions are available for particular examples such as the Elastic-Net penalty as discussed in Section 6.

**Remark 2** *For the square loss $\rho(u) = u^2/2$, the differentiability formulae (2.1) reduce to*

$$\begin{aligned}(\partial/\partial y_l)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_l,\\ (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{e}_j(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \widehat{\beta}_j\end{aligned} \tag{2.4}$$

*for almost every $(\boldsymbol{y}, \boldsymbol{X})$ and some matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ depending on $(\boldsymbol{y}, \boldsymbol{X})$, since in this case $\psi' = 1$.*

In the simple case where $g$ is twice continuously differentiable, (2.1) follows (Bellec and Zhang, 2019) with

$$\widehat{\boldsymbol{A}} = \big(\boldsymbol{X}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{X} + n\nabla^2 g(\widehat{\boldsymbol{\beta}})\big)^{-1} \tag{2.5}$$

by differentiating the KKT conditions $\boldsymbol{X}^\top\psi(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = n\nabla g(\widehat{\boldsymbol{\beta}})$. To illustrate why this is true, provided that $\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X})$ is differentiable, if $(\boldsymbol{y}(t),\boldsymbol{X}(t))$ are smooth perturbations of $(\boldsymbol{y},\boldsymbol{X})$ with $(\boldsymbol{y}(0),\boldsymbol{X}(0)) = (\boldsymbol{y},\boldsymbol{X})$ and $\frac{d}{dt}(\boldsymbol{y}(t),\boldsymbol{X}(t))|_{t=0} = (\dot{\boldsymbol{y}},\dot{\boldsymbol{X}})$, differentiation of $\boldsymbol{X}(t)^\top\psi(\boldsymbol{y}(t) - \boldsymbol{X}(t)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}(t),\boldsymbol{X}(t))) = n\nabla g(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}(t),\boldsymbol{X}(t)))$ at $t=0$ and the chain rule yields

$$\dot{\boldsymbol{X}}^\top\psi(\boldsymbol{r}) - \boldsymbol{X}^\top\operatorname{diag}\{\psi'(\boldsymbol{r})\}(\dot{\boldsymbol{y}} - \dot{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X})) = \widehat{\boldsymbol{A}}^{-1}\tfrac{d}{dt}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}(t),\boldsymbol{X}(t))\big|_{t=0}$$

with $\widehat{\boldsymbol{A}}$ in (2.5). This gives (2.1) if the penalty $g$ is twice-differentiable. Theorem 1 reveals that for *arbitrary* convex penalty functions including non-differentiable ones, the differentiability structure (2.1) always holds, as in the case of twice differentiable penalty $g$, even for penalty functions such as $g(\boldsymbol{b}) = \mu\|\boldsymbol{b}\|^2/2 + \lambda\|\mathrm{mat}(\boldsymbol{b})\|_{\mathrm{nuc}}$ where $\mathrm{mat}:\mathbb{R}^p\to\mathbb{R}^{d_1\times d_2}$ is a linear isomorphism to the space of $d_1\times d_2$ matrices and $\|\cdot\|_{\mathrm{nuc}}$ is the nuclear norm: in this case by Theorem 1 there exists a matrix $\widehat{\boldsymbol{A}}\in\mathbb{R}^{p\times p}$ such that (2.1) holds although no closed-form expression for $\widehat{\boldsymbol{A}}$ is known.

The representation (2.1) is a powerful tool as it provides explicit derivatives of quantities of interest such as $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, $\|\psi(\boldsymbol{r})\|^2$ or $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2$. These explicit derivatives can then be used in probabilistic identities and inequalities that involve derivatives, for instance Stein's formulae (Stein, 1981), the Gaussian Poincaré inequalty (Boucheron et al., 2013, Theorem 3.20), or normal approximations (Chatterjee, 2009; Bellec and Zhang, 2019).

**Remark 3** *Similar derivative formulae hold if an intercept is included in the minimization, as in*

$$\big(\widehat{\beta}_0(\boldsymbol{y},\boldsymbol{X}),\ \widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X})\big) = \operatorname*{argmin}_{b_0\in\mathbb{R},\boldsymbol{b}\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^n \rho(y_i - b_0 - \boldsymbol{x}_i^\top\boldsymbol{b}) + g(\boldsymbol{b}) \tag{2.6}$$

*Let Assumption A be fulfilled, and assume further $\|\psi'(\boldsymbol{r})\|_2 > 0$ with $\boldsymbol{r} := \boldsymbol{y} - \mathbf{1}_n\hat{\beta}_0 - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}$ where $\mathbf{1}_n = (1,...,1)^\top\in\mathbb{R}^n$. For almost every $(\boldsymbol{y},\boldsymbol{X})$ the map $(\boldsymbol{y},\boldsymbol{X})\mapsto\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X})$ is differentiable at $(\boldsymbol{y},\boldsymbol{X})$, and there exists $\widehat{\boldsymbol{A}}\in\mathbb{R}^{p\times p}$ depending on $(\boldsymbol{y},\boldsymbol{X})$ with $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op}\le(n\mu)^{-1}$ such that*

$$(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X}) = \widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{\Psi}'\boldsymbol{e}_i,\quad (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X}) = \widehat{\boldsymbol{A}}\boldsymbol{e}_j\psi(r_i) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{\Psi}'\boldsymbol{e}_i\widehat{\beta}_j, \tag{2.7}$$

*where $\boldsymbol{e}_i\in\mathbb{R}^n, \boldsymbol{e}_j\in\mathbb{R}^p$ are canonical basis vectors, $\psi = \rho'$ and $\boldsymbol{\Psi}' := \operatorname{diag}\{\psi'(\boldsymbol{r})\} - \psi'(\boldsymbol{r})\psi'(\boldsymbol{r})^\top/\sum_{i\in[n]}\psi'(r_i)$.*

## 3. Distribution of individual residuals

We now turn to the distribution of a single residual $r_i = y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}$ for some fixed observation $i\in\{1,...,n\}$ (for instance, fix $i=1$). By leveraging the differentiability structure (2.1) and the normal approximation from Bellec and Zhang (2019), the following result provides a clear picture of the distribution of $r_i$.

**Theorem 4** *Let Assumption A be fulfilled and let* $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ *be given by Theorem 1. Then for every* $i = 1, ..., n$ *there exists* $Z_i \sim N(0,1)$ *independent of* $\varepsilon_i$ *such that*

$$\left|\Big(r_i+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i)\Big)-\Big(\varepsilon_i+\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|Z_i\Big)\right| \le O_P(n^{-1/4})(|\psi(\varepsilon_i)|+\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|) \quad (3.1)$$

*Furthermore, if* $\varepsilon_i$ *has a fixed distribution* $F$*, there exists a bivariate variable* $(\tilde{\varepsilon}_i^n, \tilde{Z}_i^n)$ *converging in distribution to the product measure* $F \otimes N(0,1)$ *such that*

$$r_i + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i) = \tilde{\varepsilon}_i^n + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|\tilde{Z}_i^n. \quad (3.2)$$

*If* $\varepsilon_i$ *has a fixed distribution* $F$ *and Assumption B holds then* $|\psi(\varepsilon_i)| + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| = O_P(1)$.

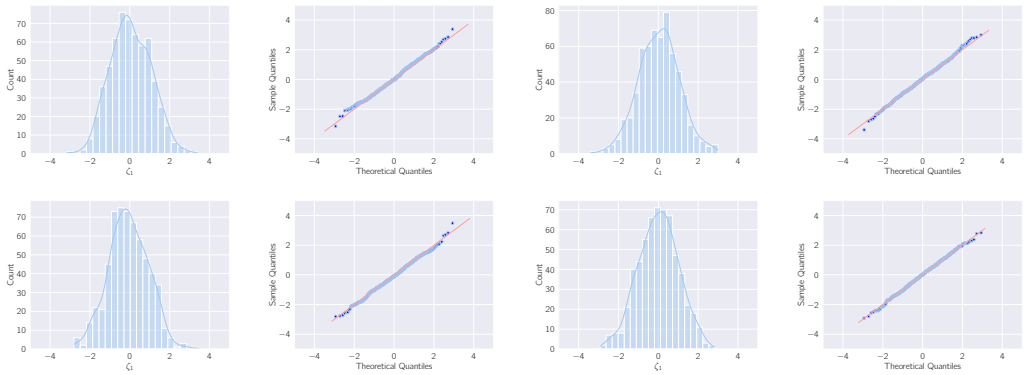


Figure 2: Histogram and QQ-plot for $\zeta_1$ in (3.3) under Huber Elastic-Net regression for different choices of tuning parameters $(\lambda, \tau)$. Left Top: $(0.036, 10^{-10})$, Right Top: $(0.054, 0.01)$, Left Bottom: $(0.036, 0.01)$, Right Bottom: $(0.024, 0.1)$. Each figure contains 600 data points generated with anisotropic design matrix and iid $\varepsilon_i$ from the $t$-distribution with 2 degrees of freedom. A detailed setup is provided in Section 6.

Theorem 4 is a formal statement regarding the informal normal approximation

$$\zeta_i := \frac{r_i + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i) - \varepsilon_i}{\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|} \approx N(0,1). \quad (3.3)$$

Simulations in Figure 2 confirm the normality of $\zeta_i$ for the Huber loss with Elastic-Net penalty and four combinations of tuning parameters. For the square loss $\rho(u) = u^2/2$, because $\psi(u) = u$, asymptotic normality of the residuals hold in the following form.

**Theorem 5** *Let Assumption A hold with* $\rho(u) = u^2/2$ *and* $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$*. Then for* $i = 1$*,*

$$\frac{(1 + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}])(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})}{(\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2)^{1/2}} \to^d N(0,1) \qquad \text{as } n \to +\infty. \quad (3.4)$$

It is informative to provide a sketch of the proof of Theorem 4 to explain the appearance of $\psi(r_i)$ and $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ in the normal approximation results (3.1) and (3.3). A variant of the normal approximation of Bellec and Zhang (2019) proved in the supplement states that for a differentiable function $\mathbf{f}:\mathbb{R}^q\to\mathbb{R}^q\setminus\{\mathbf{0}\}$ and $\boldsymbol{z}\sim N(\mathbf{0},\boldsymbol{I}_q)$, there exists $Z\sim N(0,1)$ such hat

$$\mathbb{E}\Big[\Big|\frac{\mathbf{f}(\boldsymbol{z})^\top\boldsymbol{z}-\sum_{k=1}^q(\partial/\partial z_k)f_k(\boldsymbol{z})}{\|\mathbf{f}(\boldsymbol{z})\|}-Z\Big|^2\Big]\le C_1\mathbb{E}\Big[\frac{\sum_{k=1}^q\|(\partial/\partial z_k)\mathbf{f}(\boldsymbol{z})\|^2}{\|\mathbf{f}(\boldsymbol{z})\|^2}\Big]. \tag{3.5}$$

Some technical hurdles aside, the proof sketch is the following: Apply the previous display to $q=p$, $\boldsymbol{z}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ conditionally on $(\boldsymbol{\varepsilon},(\boldsymbol{x}_l)_{l\in[n]\setminus\{i\}})$ and to $\mathbf{f}(\boldsymbol{z})=\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ in the simple case where $\boldsymbol{\beta}^*=0$ (this amounts to performing a change of variable by translation of $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$). Then the right-hand side of the previous display is negligible in probability compared to $Z$, and in the left-hand side $\mathbf{f}(\boldsymbol{z})^\top\boldsymbol{z}=\boldsymbol{x}_i^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ and $\sum_{k=1}^q(\partial/\partial z_k)f_k(\boldsymbol{z})\approx\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i)$ as the second term in (2.1) is negligible. This completes the sketch of the proof of (3.3).

**Proximal operator representation.** From the above asymptotic normality results, a stochastic representation for the $i$-th residual $r_i=y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}$ can be obtained as follows: With $\mathrm{prox}[t\rho](u)$ the proximal operator of $x\mapsto t\rho(x)$ defined as the unique solution $z\in\mathbb{R}$ of equation $z+t\psi(z)=u$,

$$r_i=y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}=\mathrm{prox}[\hat{t}\rho]\big(\tilde{\varepsilon}_i^n+\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|\tilde{Z}_i^n\big)\qquad\text{with }\hat{t}=\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$$

where $(\tilde{\varepsilon}_i^n,\tilde{Z}_i^n)$ converges in distribution to the product measure $F\otimes N(0,1)$ where $F$ is the law of $\varepsilon_i$.

## 4. A proxy of the out-of-sample error if $\boldsymbol{\Sigma}$ is known

The approximations of the previous sections for $r_i+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i)$ and the fact that $\varepsilon_i$ is independent of $Z_i\sim N(0,1)$ in (3.1) suggest that $(r_i+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(r_i))^2\approx\varepsilon_i^2+\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2Z_i^2$; and averaging over $\{1,...,n\}$ one can hope for the approximation $\|\boldsymbol{r}+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\|^2/n\approx\|\boldsymbol{\varepsilon}\|^2/n+\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2$. The following result makes this heuristic precise.

**Theorem 6** *Let Assumption A be fulfilled and $\widehat{\boldsymbol{A}}$ be given by Theorem 1. Then*

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2+\|\boldsymbol{\varepsilon}\|^2/n=\big\|\boldsymbol{r}+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\big\|^2/n+O_P(n^{-1/2})\,\mathrm{Rem},$$

*where* $\mathrm{Rem}:=\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2+\frac{1}{n}\|\psi(\boldsymbol{r})\|^2+(\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2+\frac{1}{n}\|\psi(\boldsymbol{r})\|^2)^{1/2}\|\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}\|$. *Thus*

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2+\|\boldsymbol{\varepsilon}\|^2/n=(1+O_P(n^{-1/2}))\big\|\boldsymbol{r}+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\big\|^2/n.$$

Theorem 6 provides a first candidate, $\big\|\boldsymbol{r}+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\big\|^2/n$ to estimate

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|^2+\|\boldsymbol{\varepsilon}\|^2/n. \tag{4.1}$$

Estimation of (4.1) is useful as $\|\boldsymbol{\varepsilon}\|^2/n$ is independent of the choice of the estimator $\widehat{\boldsymbol{\beta}}$ and in particular independent of the chosen loss-penalty pair in (1.1). Given two or more estimators (1.1), choosing the one with smallest $\big\|\boldsymbol{r}+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\big\|^2$ is thus a good proxy for minimizing the out-of-sample error.

**Corollary 7** *Let $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}$ be two $M$-estimators (1.1) Assumption A with loss-penalty pair $(\rho, g)$ and $(\tilde{\rho}, \tilde{g})$ respectively. Assume that both satisfy Assumption A and let $\psi = \rho'$ and $\tilde{\psi} = \tilde{\rho}'$. Let $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ be the residuals, $\widehat{\boldsymbol{A}}, \widetilde{\boldsymbol{A}}$ be the corresponding matrices of size $p \times p$ given by Theorem 1. Further assume that both estimators satisfy Assumption B and that $\boldsymbol{\varepsilon}$ has iid coordinates independent with $\mathbb{E}[|\varepsilon_i|^{1+q}] \le M$ for constants $q \in (0,1), M > 0$ independent of $n, p$. Let $\Omega = \{\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{op} \le 2\sqrt{n} + \sqrt{p}\} \cap \{\|\boldsymbol{\varepsilon}\|^2 \le n^{2/(1+q)}\}$. Then for any $\eta > 0$ independent of $n, p$ there exists $C(\gamma, \mu, \eta, q, M) > 0$ depending only on $\{\gamma, \mu, \eta, q, M\}$ such that*

$$\mathbb{P}\Big(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 - \|\boldsymbol{\Sigma}^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 > \eta, \;\; \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\|^2 \le \|\widetilde{\boldsymbol{r}} + \operatorname{tr}[\boldsymbol{\Sigma}\widetilde{\boldsymbol{A}}]\tilde{\psi}(\widetilde{\boldsymbol{r}})\|^2\Big) \le C(\gamma, \mu, \eta, q, M) n^{-q/(1+q)} + \mathbb{P}(\Omega^c) \to 0.$$

Provided that the noise random variables $\varepsilon_i$ have at least $1 + q$ moments, Theorem 7 implies that with probability approaching one given two $M$-estimators $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, choosing the estimator corresponding to the smallest criteria among $\|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\boldsymbol{r}\|^2$ and $\|\widetilde{\boldsymbol{r}} + \operatorname{tr}[\boldsymbol{\Sigma}\widetilde{\boldsymbol{A}}]\widetilde{\boldsymbol{r}}\|^2$ leads to the smallest out-of-sample error, up to any small constant $\eta > 0$. This allows noise random variables $\varepsilon_i$ with infinite variance. A similar result can be obtained to select among $K$ different $M$-estimators (1.1).

**Corollary 8** *As in Theorem 7, assume $\mathbb{E}[|\varepsilon_i|^{1+q}] \le M$ and let $\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_K$ be $M$-estimators of the form (1.1) with loss-penalty pair $(\rho_k, g_k)$ satisfying Assumptions A and B. For each $k = 1, ..., K$, let $\boldsymbol{r}_k = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k$ be the residuals and $\widehat{\boldsymbol{A}}_k$ be the corresponding matrix of size $p \times p$ from Theorem 1. Let $\hat{k} \in \operatorname{argmin}_{k=1,...,K} \|\boldsymbol{r}_k + \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}_k]\psi_k(\boldsymbol{r}_k)\|$ where $\psi_k = \rho_k'$. Then if $(\gamma, \mu, \eta, q, M)$ are constants independent of $n, p$*

$$\mathbb{P}\big(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_{\hat{k}} - \boldsymbol{\beta}^*)\|^2 > \min_{k=1,...,K} \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^*)\|^2 + \eta\big) \to 0 \qquad \text{if } K = o(n^{q/(1+q)}).$$

*In other words, if $K = o(n^{q/(1+q)})$, the selector $\hat{k}$ picks an optimal M-estimator in the sense*

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_{\hat{k}} - \boldsymbol{\beta}^*)\|^2 - \min_{k=1,...,K} \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^*)\|^2 \to^P 0.$$

Given $K$ different loss-penalty pairs and the corresponding $M$-estimators in (1.1), minimizing the criterion $\|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\boldsymbol{r}\|$ thus provably selects a loss-penalty pair that leads to an optimal out-of-sample error, up to an arbitrary small constant $\eta > 0$ independent of $n, p$. The requirement $K = o(n^{q/(1+q)})$ means that the cardinality of the collection of $M$-estimators to select from should grow more slowly than a power of $n$. This is typically satisfied for default tuning parameter grids in popular libraries (e.g., `sklearn.linear_model.Lasso` from Pedregosa et al. (2011)) with tuning parameters evenly spaced in a log-scale that consequently have cardinality logarithmic in the parameter range. The major drawback of the criterion $\|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\boldsymbol{r}\|$ is the dependence through $\operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ on the covariance $\boldsymbol{\Sigma}$ of the design, which is typically unknown. The next section introduces an estimator of $\operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ that does not require the knowledge of $\boldsymbol{\Sigma}$.

## 5. Degrees of freedom and estimating $\operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ without the knowledge of $\boldsymbol{\Sigma}$

This section focuses on estimating $\operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$. The matrix $\widehat{\boldsymbol{A}}$ from Theorem 1 can be estimated from the data $(\boldsymbol{y}, \boldsymbol{X})$ in the sense that $\widehat{\boldsymbol{A}}$ is a measurable function of $(\boldsymbol{y}, \boldsymbol{X})$ (thanks to the observation

that derivatives are limits, and limits of measurable functions are again measurable). The difficulty is thus to estimate $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ without the knowledge of $\boldsymbol{\Sigma}$. To illustrate this difficulty, consider Ridge regression with square loss $\rho(u) = u^2/2$ and penalty $g(\boldsymbol{b}) = \tau\|\boldsymbol{b}\|^2/2$. Then $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) = (\boldsymbol{X}^\top\boldsymbol{X} + \tau n\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ and $\widehat{\boldsymbol{A}}$ in Theorem 1 is given explicitly by $\widehat{\boldsymbol{A}} = (\boldsymbol{X}^\top\boldsymbol{X} + \tau n\boldsymbol{I}_p)^{-1}$ and

$$\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}] = \mathrm{tr}[(\boldsymbol{G}^\top\boldsymbol{G} + n\tau\boldsymbol{\Sigma}^{-1})^{-1}], \qquad \text{where } \boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}.$$

Above, $\boldsymbol{G}$ is a random matrix with iid $N(0,1)$ entries the value of $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ is highly dependent on the spectrum of $\boldsymbol{\Sigma}^{-1}$. In this particular case, the limit of $\mathrm{tr}[(\boldsymbol{G}^\top\boldsymbol{G} + n\tau\boldsymbol{\Sigma}^{-1})^{-1}]$ can be obtained using random matrix theory (Marčenko and Pastur, 1967) as the limiting behavior of the Stieltjes transform of $\boldsymbol{G}^\top\boldsymbol{G}/n + \tau\boldsymbol{\Sigma}^{-1}$ and its spectral distribution is known; however the limit of the spetral distribution depends on the spectrum of $\tau\boldsymbol{\Sigma}^{-1}$. This is not desirable here as we wish to construct estimators that require no knowledge on $\boldsymbol{\Sigma}$. For more involved loss-penalty pairs such as the Elastic-Net in Example 1, such random matrix theory results do not apply as $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ depends on the random support of $\widehat{\boldsymbol{\beta}}$.

Instead, we do not rely on known random matrix theory results. With the matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ given by Theorem 1, our proposal to estimate $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ is the ratio $\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}]$ with $\hat{\mathsf{df}}$ and $\boldsymbol{V}$ in (2.2)-(2.3). Both the scalar $\hat{\mathsf{df}}$ and the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are observable; in particular they do not depend on $\boldsymbol{\Sigma}$.

**Theorem 9** *Let Assumption A be fulfilled and $\widehat{\boldsymbol{A}}$ be given by Theorem 1. Then*

$$\mathbb{E}[|\,\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\,\mathrm{tr}[\boldsymbol{V}]/n - \hat{\mathsf{df}}/n|] \le C_2(\gamma,\mu)n^{-1/2}. \tag{5.1}$$

Simulations in Figure 3 and table 1 confirm that the approximation $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}] \approx \hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}]$ is accurate for the Huber loss with Elastic-Net penalty. For the square loss, $\psi' = 1$ and $\mathrm{tr}[\boldsymbol{V}] = n - \hat{\mathsf{df}}$ so that (5.1) becomes $\mathbb{E}|(1-\hat{\mathsf{df}}/n)(1+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]) - 1| \le C_3(\gamma,\mu)n^{-1/2}$ and the following result holds.

**Corollary 10** *Let Assumption A be fulfilled with $\rho(u) = u^2/2$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Then $(1 - \hat{\mathsf{df}}/n)(1+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]) \to^{\mathbb{P}} 1$ and the normality (3.4) holds with $1+\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ replaced by $(1-\hat{\mathsf{df}}/n)^{-1}$.*

For general loss $\rho$, the criterion (1.2) replaces $\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ by $\hat{\mathsf{df}}/\mathrm{tr}[\boldsymbol{V}]$ in the proxy of the out-of-sample error $\|\boldsymbol{r} + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\|^2$ studied in the previous section. Thanks to (5.1), this replacement preserves the good properties of $\|\boldsymbol{r} + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\|^2$ proved in Theorems 7 and 8.

**Theorem 11** *For $k = 1, ..., K$, let $(\rho_k, g_k)$ be a loss-penalty pair satisfying Assumptions A and B with $\psi_k = \rho_k'$, let $\widehat{\boldsymbol{\beta}}_k, \boldsymbol{r}_k, \widehat{\boldsymbol{A}}_k$ be the corresponding $M$-estimator residual vector and matrix of size $p \times p$ given by Theorem 1 as in Theorem 8 and let $\hat{\mathsf{df}}_k = \mathrm{tr}[\boldsymbol{X}\widehat{\boldsymbol{A}}_k\boldsymbol{X}^\top\,\mathrm{diag}\{\psi_k'(\boldsymbol{r}_k)\}]$ and $\boldsymbol{V}_k = \mathrm{diag}\{\psi_k'(\boldsymbol{r}_k)\}(\boldsymbol{I}_n - \boldsymbol{X}\widehat{\boldsymbol{A}}_k\boldsymbol{X}^\top\,\mathrm{diag}\{\psi_k'(\boldsymbol{r}_k)\})$. For a small constant $\eta > 0$ independent of $n, p$, say $\eta = 0.05$, define*

$$\hat{k} \in \operatorname*{argmin}_{k=1,\dots,K} \left\|\boldsymbol{r}_k + \frac{\hat{\mathsf{df}}_k}{\mathrm{tr}[\boldsymbol{V}_k]}\psi_k(\boldsymbol{r}_k)\right\|^2 \qquad \textit{subject to} \qquad \frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki}) \ge \eta.$$

*If $\varepsilon_i$ has $1+q$ moments in the sense that $\mathbb{E}[|\varepsilon_i|^{1+q}] \le M$ for constants $q \in (0,1), M > 0$. If $(M, q, \eta, \mu, \gamma)$ and $\tilde{\eta} > 0$ are independent of $n, p$ then*

$$\mathbb{P}\Big(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_{\hat{k}} - \boldsymbol{\beta}^*)\| > \min_{k=1,\dots,K:\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki})\ge\eta} \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^*)\| + \tilde{\eta}\Big) \to 0 \qquad \textit{if } K = o(n^{q/(1+q)}).$$

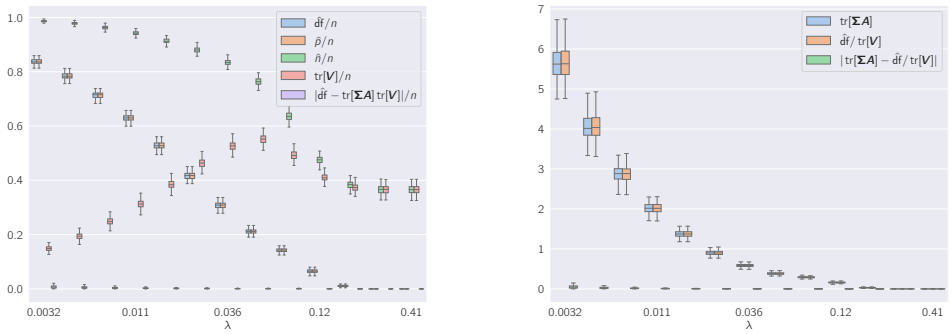


Figure 3: Above: Boxplots for $\hat{\mathrm{df}}, \hat{p}, \hat{n}, \mathrm{tr}[\boldsymbol{V}], \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ and $|\,\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}] - \hat{\mathrm{df}}/\,\mathrm{tr}[\boldsymbol{V}]|$ in Huber Elastic-Net regression with $\tau = 10^{-10}$ and $\lambda \in [0.0032, 0.41]$. Each box contains 200 data points. Below: heatmaps for $\hat{\mathrm{df}}/n$, $\mathrm{tr}[\boldsymbol{V}]/n$ and $\hat{n}/n = \sum_{i=1}^n \psi'(r_i)/n$ under the simulation setup in Figure 1. The detailed simulation setup is given in Section 6.

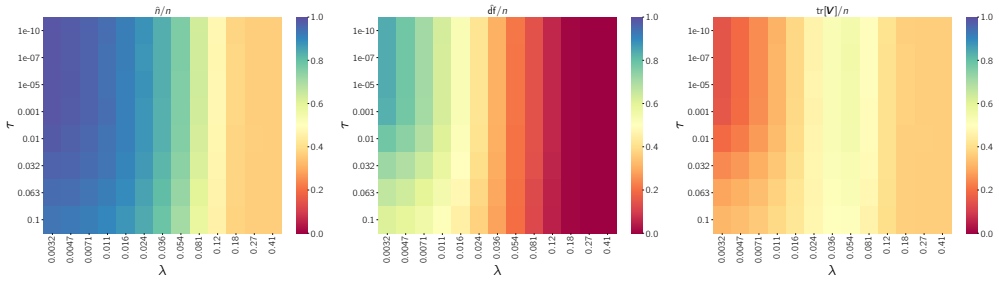


Figure 1 illustrates on simulations the success of the criterion (1.2) over a grid of tuning parameters for $M$-estimators with the Huber loss and Elastic-Net penalty. The criterion (1.2) is thus successful at selecting a $M$-estimator with smallest out-of-sample error up to an additive constant $\tilde{\eta}$, among those $M$-estimators indexed in $\{1, ..., K\}$ that are such that $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki}) \geq \eta$. On the one hand it is unclear to us whether the restriction $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki}) \geq \eta$ can be omitted. On the other hand there is a practical meaning in excluding $M$-estimators with small $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki})$: For the Huber loss $H(u) := u^2/2$ for $|u| \leq 1$ and $|u| - 1/2$ for $|u| \geq 1$ the quantity $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki})$ is the number of of data points in $\{1, ..., n\}$ such that the residual $y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k$ fall within the quadratic regime of the loss function. Observations $i \in \{1, ..., n\}$ that fall in the linear regime of the loss are excluded from the fit, in the sense that for some $i$ with $r_{ki} = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k > 1$, replacing $y_i$ by $\tilde{y}_i = y_i + 1000$ (or any positive value) does not change the $M$-estimator solution $\widehat{\boldsymbol{\beta}}_k$ (this can be seen from the KKT conditions directly, or by integration the derivative with respect to $y_i$ in (2.1)). Thus the constraint $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki}) \geq \eta$ requires that at most a constant fraction of the observations are excluded from the fit (or equivalently, at least a constant fraction of the $n$ observations participate in the fit). For scaled versions of the Huber loss, $\rho_k(u) = a^2 H(a^{-1}u)$ for some $a > 0$, the value $\hat{n} = \frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki})$ again counts the number of residuals falling in the quadratic regime of the loss, i.e., the number of observations participating in the fit. The heatmaps of Figure 3 illustrate $\hat{n}$ in a simulation for a wide range of parameters. Similarly, for smooth robust loss functions such as

$\rho_k(u) = \sqrt{1+u^2}$, the constraint $\frac{1}{n}\sum_{i=1}^n \psi_k'(r_{ki}) \geq \eta$ requires that at most a constant fraction of the $n$ observations are such that $\psi_k'(r_{ki}) < \eta/2$, i.e., such that the second derivative $\psi_k'$ is too small (and the loss $\rho_k$ too flat).

Theorems 1, 5, 6 and 9 provide our general results applicable to a single regularized $M$-estimator (1.1) while corollaries such as Theorem 11 are obtained using the union bound. The next section specializes our results and notation to the Huber loss with Elastic-Net penalty and details the simulation setup used in the figures.

## 6. Example and simulation setting: Huber loss with Elastic-Net penalty

In simulations and in the example below, we focus on the loss-penalty pair

$$\rho(u;\Lambda) = \Lambda^2 H(\Lambda^{-1}u), \qquad g(\boldsymbol{b};\lambda,\tau) = \lambda\|\boldsymbol{b}\|_1 + (\tau/2)\|\boldsymbol{b}\|_2^2 \tag{6.1}$$

for tuning parameters $\Lambda, \lambda, \tau \geq 0$ where $H(u) := u^2/2$ for $|u| \leq 1$ and $|u| - 1/2$ for $|u| \geq 1$.

**Example 1** *With $(\rho, g)$ in (6.1), matrix $\widehat{\boldsymbol{A}}$ in (2.1) matrix $\boldsymbol{V}$ in (2.3) and $\hat{\mathsf{df}}$ in (2.2) we have*

$$\begin{aligned}
&\widehat{\boldsymbol{A}}_{\hat{S},\hat{S}} = (\boldsymbol{X}_{\hat{S}}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{X}_{\hat{S}} + n\tau\boldsymbol{I}_{\hat{p}})^{-1}, \quad A_{i,j} = 0 \text{ if } i \notin \hat{S} \text{ or } j \notin \hat{S},\\
&\boldsymbol{V} = \operatorname{diag}\{\psi'(\boldsymbol{r})\} - \operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{X}_{\hat{S}}\widehat{\boldsymbol{A}}_{\hat{S},\hat{S}}\boldsymbol{X}_{\hat{S}}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\},\\
&\hat{\mathsf{df}} = \operatorname{tr}[\boldsymbol{X}_{\hat{S}}\widehat{\boldsymbol{A}}_{\hat{S},\hat{S}}\boldsymbol{X}_{\hat{S}}^\top \operatorname{diag}\{\psi'(\boldsymbol{r})\}],
\end{aligned} \tag{6.2}$$

*where $\hat{S}$ is the active set $\{j \in [p] : \hat{\beta}_j \neq 0\}$ and $\hat{p}$ is the size of $\hat{S}$; $\boldsymbol{X}_{\hat{S}}$ is the submatrix of $\boldsymbol{X}$ selecting columns with index in $\hat{S}$ and $\widehat{\boldsymbol{A}}_{\hat{S},\hat{S}}$ is the submatrix of $\widehat{\boldsymbol{A}}$ with entries indexed in $\hat{S} \times \hat{S}$.*

| $(\lambda,\tau)$ | $(0.036, 10^{-10})$ | $(0.054, 0.01)$ | $(0.036, 0.01)$ | $(0.024, 0.1)$ |
|---|---|---|---|---|
| $\hat{\mathsf{df}}/n$ | $0.31 \pm 0.012$ | $0.21 \pm 0.0095$ | $0.3 \pm 0.011$ | $0.37 \pm 0.0093$ |
| $\hat{p}/n$ | $0.31 \pm 0.012$ | $0.22 \pm 0.0098$ | $0.31 \pm 0.012$ | $0.47 \pm 0.014$ |
| $\hat{n}/n$ | $0.83 \pm 0.011$ | $0.76 \pm 0.014$ | $0.83 \pm 0.012$ | $0.84 \pm 0.012$ |
| $\operatorname{tr}[\boldsymbol{\Sigma}\boldsymbol{A}]$ | $0.58 \pm 0.039$ | $0.39 \pm 0.027$ | $0.58 \pm 0.038$ | $0.8 \pm 0.038$ |
| $\lvert\operatorname{tr}[\boldsymbol{\Sigma}\boldsymbol{A}] - \hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}]\rvert$ | $0.0019 \pm 0.0015$ | $0.0015 \pm 0.0012$ | $0.0021 \pm 0.0016$ | $0.0023 \pm 0.0017$ |
| $\lVert\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\rVert^2$ | $1.3 \pm 0.18$ | $1.7 \pm 0.25$ | $1.3 \pm 0.19$ | $1.9 \pm 0.21$ |
| $\zeta_1$ | $0.056 \pm 1$ | $0.021 \pm 1$ | $0.0044 \pm 1$ | $0.042 \pm 0.97$ |

Table 1: Simulation for Huber Elastic-Net regression under different choices of $(\lambda,\tau)$. $(n,p) = (1001, 1000)$. For each choice of $(\lambda,\tau)$, 600 data points are simulated with anisotropic design matrix and i.i.d. $t$-distributed noises with 2 degrees of freedom. A detailed setup is provided in Section 6.

The identities (6.2) are proved in (Bellec, 2020, §2.6). Simulations in Figures 1 to 3 and table 1 illustrate typical values for $\hat{\mathsf{df}}, \operatorname{tr}[\boldsymbol{V}], \operatorname{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$, the out-of-sample error and the criterion (1.2), $\hat{n} = \sum_{i=1}^n \psi'(r_i)$ and $\hat{p} = |\hat{S}|$ under anisotropic Gaussian design and heavy-tailed $\varepsilon_i$. The simulation setup is as follows.

**Data Generation Process.** Simulation data are generated from a linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with anisotropic Gaussian design $\boldsymbol{\Sigma}$ and heavy-tail noise vector $\boldsymbol{\varepsilon}$. The design matrix $\boldsymbol{X}$ has $n = 1001$ rows and $p = 1000$ columns. Each row of $\boldsymbol{X}$ is i.i.d. $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, with the same $\boldsymbol{\Sigma}$ across all repetitions, generated once by $\boldsymbol{\Sigma} = \boldsymbol{R}^\top \boldsymbol{R}/(2p)$ with $\boldsymbol{R} \in \mathbb{R}^{2p\times p}$ being a Rademacher matrix with i.i.d. entries $\mathbb{P}(\boldsymbol{R}_{ij} = \pm 1) = \frac{1}{2}$. The true signal vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ has its first 100 coordinates set to $p^{1/2}/100 = \sqrt{10}/10$ and the rest 900 coordinates set to 0. The noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ has i.i.d. entries from the t-distribution with 2 degrees of freedom (so that $\mathrm{Var}[\varepsilon_i] = \infty$, i.e., $\varepsilon_i$ is heavy-tailed).

**Estimation Process.** Each dataset $(\boldsymbol{y}, \boldsymbol{X})$ is fitted by a Huber Elastic-Net estimator with loss-penalty pair in (6.1). We focus on 2d heatmaps with respect to the two penalty parameters $(\lambda, \tau)$ of the penalty; to this end the Huber loss parameter $\Lambda$ is set to $\Lambda = 0.054n^{1/2}$ and a grid for $(\lambda, \tau)$ in then set so that $\hat{\mathrm{df}}/n$ varies on the grid from 0 to 1 (cf. the middle heatmap in Figure 3). The Elastic-Net penalty $g(\boldsymbol{b}; \lambda, \tau) = \lambda\|\boldsymbol{b}\|_1 + (\tau/2)\|\boldsymbol{b}\|_2^2$ is used with $(\lambda, \tau) \in \{(0.036, 10^{-10}), (0.054, 0.01), (0.036, 0.01), (0.024, 0.1)\}$ in Figure 2 and table 1, $(\lambda, \tau) \in [0.0032, 0.41]\times\{10^{-10}\}$ in Figure 3, and $(\lambda, \tau) \in [0.0032, 0.041] \times [10^{-10}, 0.1]$ in Figure 1. More simulation results are provided in the supplementary materials. In these simulations, the criterion (1.5) from Rad and Maleki (2020) was also computed and was not noticeably different from (1.2), cf. the lower half of Figures 6 and 7.

## 7. Relaxing the strong convexity assumption

While previous results rely heavily on the $\mu$-strong convexity assumption (with respcet to $\boldsymbol{\Sigma}$, as stated in the last part of Assumption A), the proof of the following proposition presents a device that lets us generalize the results under the following condition: For any $\boldsymbol{\varepsilon}$, there exists an open set $U_{\boldsymbol{\varepsilon}} \subset \mathbb{R}^{n\times p}$ such that the mapping

$$\Phi_{\boldsymbol{\varepsilon}} : \begin{cases} U_{\boldsymbol{\varepsilon}} \to \mathbb{R}^{n+p}, \\ \boldsymbol{X} \mapsto \dfrac{\big(n^{-1/2}\psi(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})),\ \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta})\big)}{(\frac{1}{n}\|\psi(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}))\|^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta})\|^2)^{1/2}} \end{cases} \text{ is } \frac{L}{\sqrt{n}}\text{-Lipschitz.} \quad (7.1)$$

In this definition, $\boldsymbol{\varepsilon}$ is held fixed and $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ is the composition of $\widehat{\boldsymbol{\beta}}$ with the function $\boldsymbol{X} \mapsto (\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ so that $\Phi_{\boldsymbol{\varepsilon}}$ is a function of $\boldsymbol{X}$ only. The following proposition shows that strong convexity on the penalty function can be relaxed, provided that the above Lipschitz condition holds and the expectations are restricted to the event $\{\boldsymbol{X} \in U_{\boldsymbol{\varepsilon}}\}$.

**Proposition 12** *Let $L, \gamma > 0$ be constants and assume $p/n \le \gamma$. Consider a convex differentiable loss $\rho : \mathbb{R} \to \mathbb{R}$ such that $\psi = \rho'$ is 1-Lipschitz and a convex penalty $g$, assume $\boldsymbol{X}$ has iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows with invertible $\boldsymbol{\Sigma}$ and the noise $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ with continuous distribution. Assume that for some open $U_{\boldsymbol{\varepsilon}} \subset \mathbb{R}^{n\times p}$, the Lipschitz condition (7.1) holds and that almost everywhere in $U_{\boldsymbol{\varepsilon}}$, the derivative formulae (2.1) hold for some matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ satisfying $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le L/n$. Then*

$$\mathbb{E}\Big[I\{\boldsymbol{X} \in U_{\boldsymbol{\varepsilon}}\}\Big|\frac{\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\,\mathrm{tr}[\boldsymbol{V}]}{n} - \frac{\hat{\mathrm{df}}}{n}\Big|\Big] \le \frac{C_4(\gamma, L)}{\sqrt{n}} \quad (7.2)$$

$$\mathbb{E}\Big[\frac{I\{\boldsymbol{X} \in U_{\boldsymbol{\varepsilon}}\}}{\mathrm{Rem}}\Big|\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2 + \frac{\|\boldsymbol{\varepsilon}\|^2}{n} - \frac{\big\|\boldsymbol{r} + \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]\psi(\boldsymbol{r})\big\|^2}{n}\Big|\Big] \le \frac{C_5(\gamma, L)}{\sqrt{n}} \quad (7.3)$$

*where* $I\{\boldsymbol{X} \in U_\varepsilon\}$ *is the indicator function of the event* $\{\boldsymbol{X} \in U_\varepsilon\}$ *and* Rem *is defined in Theorem 6.*

Proposition 12 is proved in Appendix E. Consequently, if the event $\{\boldsymbol{X} \in U_\varepsilon\}$ has high probability and the Lipschitz condition (7.1) holds in this event, the main results Theorems 6 and 9 still hold, with no strong convexity assumption on the penalty. The proof relies on an application of Kirszbraun's theorem already presented in (Bellec, 2020).

The Lipschitz condition (7.1) and inequality $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le L/n$ have been proved to hold in the regime $n \asymp p$ for covariance $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}_{jj} = 1, \forall j \in [p]$ in the following two cases:

- The Lasso (i.e., square loss and L1 penalty) under the assumption that $\|\boldsymbol{\beta}^*|_0 \le s_* n$ for some enough small constant $s_*$;
- The Huber Lasso (i.e., Huber Loss and L1 penalty) under the assumption that $\|\boldsymbol{\beta}^*|_0 \le s_* n$ for some small enough constant $s_*$, and that at least $(1 - s_*)n$ components of the noise (the "inliers") are iid standard normal;

cf. (Bellec, 2020, Assumption 2.3 and Proposition 12.1). In both cases, the constant $s_*$ only depends on $\gamma$, the condition number of $\boldsymbol{\Sigma}$ and the multiplicative constant of the tuning parameters.

Our results can thus be extended on a case-by-case basis for loss-penalty pairs such that the Lipschitz condition (7.1) holds, for instance for the two above examples.

## Acknowledgments

P.C.B.'s research was partially supported by the NSF Grant DMS-1811976 and DMS-1945428 and by Ecole Des Ponts ParisTech.

## References

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pages 944–952, 2013.

Pierre C Bellec. Out-of-sample error estimate for robust m-estimators with convex penalty. *arXiv:2008.11840*, 2020. URL https://arxiv.org/pdf/2008.11840.pdf.

Pierre C Bellec and Cun-Hui Zhang. Second order stein: Sure for sure and other applications in high-dimensional inference. *Annals of Statistics, accepted, to appear*, 2018. URL https://arxiv.org/pdf/1811.04121.pdf.

Pierre C Bellec and Cun-Hui Zhang. De-biasing convex regularized estimators and interval estimation in linear models. *arXiv:1912.11943*, 2019. URL https://arxiv.org/pdf/1912.11943.pdf.

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. *arXiv preprint arXiv:1903.10603*, 2019.

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

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

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

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

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

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.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Iosif Pinelis. Large deviations: Growth of empirical average of iid non-negative random varialbes with infinite expectations? MathOverflow, 2021. URL https://mathoverflow.net/q/390939. URL:https://mathoverflow.net/q/390939 (version: 2021-05-24).

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

Kamiar Rahnama Rad, Wenda Zhou, and Arian Maleki. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 4067–4077. PMLR, 2020.

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

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

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.

Ji Xu, Arian Maleki, Kamiar Rahnama Rad, and Daniel Hsu. Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9): 5997–6030, 2021.

William P Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120. Springer-Verlag New York, 1989. doi: 10.1007/978-1-4612-1015-3.

## Appendix A. Proof of the main results

**Notation.** For vectors in $\mathbb{R}^q$ or $\mathbb{R}^n$, the Euclidean norm is $\|\cdot\|$ and $\|\cdot\|_q$ is the $\ell_q$-norm for $1 \le q \le +\infty$. For matrices, $\|\cdot\|_{op}$ is the operator norm (largest singular value), $\|\cdot\|_F$ the Frobenius norm. We use index $i$ only to loop or sum over $[n] = \{1, ..., n\}$ and $j$ only to loop or sum over $[p] = \{1, ..., p\}$, so that $\boldsymbol{e}_i \in \mathbb{R}^n$ refers to the $i$-th canonical basis vector in $\mathbb{R}^n$ and $\boldsymbol{e}_j \in \mathbb{R}^p$ the $j$-th canonical basis vector in $\mathbb{R}^p$. Positive absolute constants are denoted $C_0, C_1, C_2, ...,$, constants that depend on $\gamma$ only are denoted $C_0(\gamma), C_1(\gamma), ...$ and constant that depend on $\gamma, \mu$ only are denoted by $C_0(\gamma, \mu), C_1(\gamma, \mu), \dots$ If $\boldsymbol{f} : \mathbb{R}^q \to \mathbb{R}^n$ is differentiable at $\boldsymbol{z} \in \mathbb{R}^q$, we denote the Jacobian matrix in $\mathbb{R}^{n\times q}$ by $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}}$ or $\partial \boldsymbol{f}/\partial \boldsymbol{z}$. For an event $\Omega$, its indicator function is denoted by $I_\Omega$ or $I\{\Omega\}$.

**Organization of the proofs.** Appendix B provides the proof of the main results from the main text (Theorems 4 to 9 and 11) and the overall proof strategy. Appendix C gives the proof of the probabilistic tools used in Appendix B. Appendix D proves the differentiability formulae in Theorems 1 and 3.

**Additional simulations.** Additional simulations and figures are given in Appendix F for Gaussian designs and in Appendix G for non-Gaussian Rademacher design. The simulations for Rademacher design suggests that our results generalize to non-Gaussian design, although it is unclear at this point how to extend the proofs to non-Gaussian $\boldsymbol{X}$.

## Appendix B. Proof of the main results

We perform the following change of variable to reduce the anisotropic design regression problem to an isotropic one, $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n\times p}$ a Gaussian matrix with iid $N(0,1)$ entries and

$$\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{G}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(\varepsilon_i - \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{u}) + g(\boldsymbol{\beta}^* + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}) \tag{B.1}$$

and denote by $(h_j)_{j=1,...,p}$ the components of (B.1). Then $\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}^*) = \boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ with $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ the $M$-estimator in (1.1). With $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, by the chain rule and (2.1),

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}(\partial/\partial g_{ij})\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{G}) \\
&= (\partial/\partial g_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{G}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \boldsymbol{G}\boldsymbol{\Sigma}^{1/2}) \\
&= \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*)\boldsymbol{e}_j + \widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_j\psi(r_i) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i\psi'(r_i)(\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}})\boldsymbol{e}_j
\end{aligned}$$

where $\boldsymbol{e}_i \in \mathbb{R}^n, \boldsymbol{e}_j \in \mathbb{R}^p$ denote canonical basis vectors. Define $\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{G}) = \psi(\boldsymbol{\varepsilon} - \boldsymbol{G}\boldsymbol{h})$ and let

$$\boldsymbol{A} := \boldsymbol{\Sigma}^{1/2} \widehat{\boldsymbol{A}\boldsymbol{\Sigma}}^{1/2}.$$

Then we have

$$(\partial/\partial g_{ij})\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{G}) = \boldsymbol{A}\boldsymbol{e}_j \psi(r_i) - \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_i \psi'(r_i) h_j \tag{B.2}$$

$$(\partial/\partial g_{ij})\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{G}) = -\operatorname{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j \psi(r_i) - \boldsymbol{V}\boldsymbol{e}_i h_j \tag{B.3}$$

where the second line follows by the chain rule for Lipschitz functions in in (Ziemer, 1989, Theorem 2.1.11). The crux of the argument is that the quantities of interest appearing in our results, $\|\boldsymbol{h}\|^2 = \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2, \|\psi(\boldsymbol{r})\|^2$, $\operatorname{tr}[\widehat{\boldsymbol{A}\boldsymbol{\Sigma}}] = \operatorname{tr}[\boldsymbol{A}], \operatorname{tr}[\boldsymbol{V}]$ and $\hat{\mathsf{df}}$ naturally appear from tensor contractions involving the derivatives in (B.2)-(B.3). For instance, denoting $\boldsymbol{D} = \operatorname{diag}\{\psi'(\boldsymbol{r})\} \in \mathbb{R}^{n\times n}$ if $h_j, \psi_i$ are the $j$-th and $i$-th component of (B.1) and $\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{G})$ and denoting $\sum_{i=1}^n \sum_{j=1}^p$ by $\sum_{ij}$ for brevity,

$$\sum_{j=1}^p \frac{\partial h_j}{g_{ij}} = \operatorname{tr}[\boldsymbol{A}]\psi_i - \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i \qquad \text{for a given } i = 1, ..., n, \tag{B.4}$$

$$\sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}} = -\boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j - \operatorname{tr}[\boldsymbol{V}]h_j \qquad \text{for a given } j = 1, ..., p, \tag{B.5}$$

$$\sum_{ij} \frac{\partial(h_j \psi_i)}{g_{ij}} = \|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] - \boldsymbol{h}^\top \boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{\psi} - \boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{h} - \|\boldsymbol{h}\|^2 \operatorname{tr}[\boldsymbol{V}], \tag{B.6}$$

$$\sum_{ij} \frac{\partial(h_j \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{h})}{g_{ij}} = \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{h} + n\|\boldsymbol{h}\|^2 + \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{A}\boldsymbol{h} - \|\boldsymbol{h}\|^2 \hat{\mathsf{df}}, \tag{B.7}$$

$$\sum_{ij} \frac{\partial(\psi_i \boldsymbol{e}_j^\top \boldsymbol{G}^\top \boldsymbol{\psi})}{g_{ij}} = -\boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{\psi} - \operatorname{tr}[\boldsymbol{V}]\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \boldsymbol{h}^\top \boldsymbol{G}^\top \boldsymbol{V}\boldsymbol{\psi} + (p - \hat{\mathsf{df}})\|\boldsymbol{\psi}\|^2 \tag{B.8}$$

where we used that $\hat{\mathsf{df}} = \sum_{i=1}^n \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i = \operatorname{tr}[\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}]$ in the fourth line and $\hat{\mathsf{df}} = \sum_{j=1}^p \boldsymbol{e}_j^\top \boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j = \operatorname{tr}[\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}]$ in the fifth thanks to the commutation property of the trace. The terms in colored purple indicate terms that will be proved to be negligible later on. The probabilistic tool that leads to asymptotic normality of the residuals is the following.

**Proposition 13** *[Variant of Bellec and Zhang (2019)] Let $\boldsymbol{z} \in N(\boldsymbol{0}, \boldsymbol{I}_q)$ and $\boldsymbol{f} := \boldsymbol{f}(\boldsymbol{z}) : \mathbb{R}^q \to \mathbb{R}^q \setminus \{\boldsymbol{0}\}$ be locally Lipschitz in $\boldsymbol{z}$ with $\mathbb{E}[\|\boldsymbol{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \boldsymbol{f}}{\partial z_k}\|^2] < +\infty$. Then*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{f}^\top \boldsymbol{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\boldsymbol{f}\|_2} - Z\Big)^2\Big] \le (7 + 2\sqrt{6})\mathbb{E}\Big[\|\boldsymbol{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \boldsymbol{f}}{\partial z_k}\|^2\Big] < +\infty. \tag{B.9}$$

Proposition 13 is proved in Appendix C. From here, asymptotic normality of the residuals in the square loss case is readily obtained using the explicit formulae for the derivatives and the contraction (B.4). We start with the square loss and the proof of Theorem 5.

**Proof (Proof of Theorem 5)** Apply Proposition 13 with $q = p + 1$ and $\boldsymbol{z} = (\boldsymbol{g}_i, \varepsilon_i/\sigma) \sim N(\boldsymbol{0}, \boldsymbol{I}_{p+1})$ conditionally on $(\boldsymbol{g}_l, \varepsilon_l)_{l\in[n]\setminus\{i\}}$, and with $\boldsymbol{f} = (\boldsymbol{h}, -\sigma) \in \mathbb{R}^{p+1}$. Note that the last

component of $\boldsymbol{f}$ is constant and $\|\boldsymbol{f}\|^2 = \|\boldsymbol{h}\|^2 + \sigma^2$. By (B.4) and $\boldsymbol{D} = \boldsymbol{I}_n$ for the square loss, $\mathrm{tr}[\partial \boldsymbol{f}/\partial \boldsymbol{z}] = \mathrm{tr}[\boldsymbol{A}]\psi_i - \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_i$ and by symmetry in $i = 1, ..., n$, $\mathbb{E}[|\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_i|^2/\|\boldsymbol{f}\|^2] = \frac{1}{n}\sum_{l=1}^n \mathbb{E}[|\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_l|^2/\|\boldsymbol{f}\|^2] = \frac{1}{n}\mathbb{E}[\|\boldsymbol{G}\boldsymbol{A}^\top \boldsymbol{h}\|^2/\|\boldsymbol{f}\|^2] \le \frac{1}{n}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2\|\boldsymbol{A}\|_{op}^2] \le n^{-2}C_6(\gamma,\mu)$ thanks to $\|\boldsymbol{A}\|_{op} \le 1/(n\mu)$ and $\mathbb{E}[\|\boldsymbol{G}\|_{op}^2] \le C_7(\gamma)n$. Similarly, for the square loss $r_i = \psi_i = \varepsilon_i - \boldsymbol{g}_i^\top \boldsymbol{h}$ and

$$\begin{aligned}\|\boldsymbol{f}\|^{-1}\|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F &= (\|\boldsymbol{h}\|^2+\sigma^2)^{-1/2}\|\boldsymbol{A}\psi_i - \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_i\boldsymbol{h}^\top\|_F \\ &\le \|\boldsymbol{A}\|_{op}[\sqrt{p}|\varepsilon_i|/\sigma + \sqrt{p}\|\boldsymbol{h}\|^{-1}|\boldsymbol{g}_i^\top \boldsymbol{h}| + \|\boldsymbol{G}\|_{op}].\end{aligned}$$

By the triangle inequality, $\|\boldsymbol{A}\|_{op} \le 1/(n\mu)$ and $p \le \gamma n$,

$$\mathbb{E}[\|\boldsymbol{f}\|^{-2}\|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F^2]^{1/2} \le \tfrac{\sqrt{p}}{n\mu}(\mathbb{E}[\varepsilon_i^2/\sigma^2]^{1/2} + \mathbb{E}[(\boldsymbol{g}_i^\top \boldsymbol{h})^2/\|\boldsymbol{h}\|^2]^{1/2}) + \tfrac{1}{n\mu}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2]^{1/2}.$$

By symmetry in $i = 1, ..., n$, $\mathbb{E}[(\boldsymbol{g}_i^\top \boldsymbol{h})^2/\|\boldsymbol{h}\|^2] = \frac{1}{n}\sum_{l=1}^n \mathbb{E}[(\boldsymbol{g}_l^\top \boldsymbol{h})^2/\|\boldsymbol{h}\|^2] \le \frac{1}{n}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2]$. Since $\frac{1}{n}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2] \le C_8(\gamma)$, the right-hand side in the previous display is bounded from above by $C_9(\gamma,\mu)n^{-1/2}$. Since $\boldsymbol{f}^\top \boldsymbol{z} = -r_i$ we obtain $-r_i - \mathrm{tr}[\boldsymbol{A}]r_i = (\|\boldsymbol{h}\|^2+\sigma^2)^{1/2}(Z + O_P(n^{-1/2}))$ which completes the proof of (3.4). ■

**Proof (Proof of Theorem 4)** Let $U \sim N(0,1)$ be independent of everything else. We apply the previous proposition with $\boldsymbol{z} = (\boldsymbol{g}_i, U) \sim N(\boldsymbol{0}, \boldsymbol{I}_{p+1})$ conditionally on $(\boldsymbol{\varepsilon}, \boldsymbol{g}_l, l \in [n] \setminus \{i\})$ to $\boldsymbol{f} = (\boldsymbol{h}, n^{-1/4}\psi(\varepsilon_i))$. Note that the last component of $\boldsymbol{f}$ is constant. By (B.4), $\mathrm{tr}[\partial \boldsymbol{f}/\partial \boldsymbol{z}] = \mathrm{tr}[\boldsymbol{A}]\psi_i - \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i$ and by (B.2),

$$\|\boldsymbol{f}\|^{-1}\|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F = (\|\boldsymbol{h}\|^2 + n^{-1/2}\psi(\varepsilon_i)^2)^{-1/2}\|\boldsymbol{A}\psi_i - \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i\boldsymbol{h}^\top\|_F \tag{B.10}$$

$$\le \|\boldsymbol{A}\|_{op}[n^{1/4}\sqrt{p} + \sqrt{p}\|\boldsymbol{h}\|^{-1}|\boldsymbol{g}_i^\top \boldsymbol{h}| + \|\boldsymbol{G}\|_{op}] \tag{B.11}$$

where we used $\|\boldsymbol{A}\|_F \le \sqrt{p}\|\boldsymbol{A}\|_{op}$ and $|\psi_i| \le \psi(\varepsilon_i) + |\boldsymbol{g}_i^\top \boldsymbol{h}|$ thanks to $\psi$ being 1-Lipschitz. We have $\|\boldsymbol{A}\|_{op} \le 1/(n\mu)$ and $\mathbb{E}[\|\boldsymbol{h}\|^{-2}|\boldsymbol{g}_i^\top \boldsymbol{h}|^2] = \frac{1}{n}\sum_{l=1}^n \mathbb{E}[\|\boldsymbol{h}\|^{-2}|\boldsymbol{g}_l^\top \boldsymbol{h}|^2] = \frac{1}{n}\mathbb{E}[\|\boldsymbol{h}\|^{-2}\|\boldsymbol{G}\boldsymbol{h}\|^2] \le \frac{1}{n}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2]$ by symmetry in $i = 1, ..., n$, so that $\mathbb{E}[\|\boldsymbol{f}\|^{-2}\|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F^2] \le n^{-1/2}C_{10}(\gamma,\mu)$. Thus by Proposition 13,

$$\begin{aligned}(-r_i - \mathrm{tr}[\boldsymbol{A}]\psi_i) + (\varepsilon_i - \|\boldsymbol{h}\|Z) &= \boldsymbol{g}_i^\top \boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\psi_i - \|\boldsymbol{h}\|Z \\ &= -Un^{-1/4}\psi(\varepsilon_i) + [\|\boldsymbol{f}\| - \|\boldsymbol{h}\|]Z + \|\boldsymbol{f}\|\,\mathrm{Rem} - \boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i\end{aligned}$$

where $\mathbb{E}[\mathrm{Rem}^2] \le C_{11}\mathbb{E}[\|\boldsymbol{f}\|^{-2}\|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F^2] \le n^{-1/2}C_{12}(\gamma,\mu)$. By properties of the operator norm and symmetry in $i = 1, ..., n$,

$$\mathbb{E}[\|\boldsymbol{h}\|^{-2}|\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i|^2] = \tfrac{1}{n}\mathbb{E}[\|\boldsymbol{h}\|^{-2}\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}^\top \boldsymbol{h}\|^2] \le \tfrac{1}{n}\mathbb{E}[\|\boldsymbol{G}\|_{op}^2\|\boldsymbol{A}\|_{op}^2] \le \tfrac{C_{13}(\gamma,\mu)}{n^2}. \tag{B.12}$$

By the triangle inequality, $|\|\boldsymbol{f}\| - \|\boldsymbol{h}\|| \le n^{-1/4}|\psi(\varepsilon_i)|$ so that the right-hand side is of the form $O_P(n^{-1/4})(|\psi(\varepsilon_i)| + \|\boldsymbol{h}\|)$ as desired. The previous display can be rewritten as $r_i + \mathrm{tr}[\boldsymbol{A}]\psi_i = \tilde{\varepsilon}_i^n + \|\boldsymbol{h}\|\tilde{Z}_i^n$ for

$$\tilde{\varepsilon}_i^n = \varepsilon_i + Un^{-1/4}\psi(\varepsilon_i) - [\|\boldsymbol{f}\| - \|\boldsymbol{h}\|](Z + \mathrm{Rem}), \qquad \tilde{Z}_i^n = -Z - \mathrm{Rem} + \|\boldsymbol{h}\|^{-1}\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{e}_i.$$

If $\varepsilon_i$ has a fixed distribution $F$, then $|\psi(\varepsilon_i)| \le |\psi(0)| + |\varepsilon_i| = |\varepsilon_i| = O_P(1)$ thanks to $\psi(0) = 0$ and $\psi$ being 1-Lipschitz so that $(\tilde{\varepsilon}_i^n, \tilde{Z}_i^n) = (\varepsilon_i, -Z) + O_P(n^{-1/4})$. Since $(\varepsilon_i, -Z)$ are independent, by Slutsky's theorem this proves that $(\tilde{\varepsilon}_i^n, \tilde{Z}_i^n)$ converges weakly to the product measure $F \otimes N(0,1)$. ■

**Proposition 14** *Let $\boldsymbol{h} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$, $\boldsymbol{\psi} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ be locally Lipschitz functions. If $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has iid $N(0,1)$ entries then*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \sum_{ij} \frac{\partial(\psi_i h_j)}{g_{ij}}}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big)^2 + \Big(\frac{\|\boldsymbol{G}\boldsymbol{h}\|^2 - \sum_{ij} \frac{\partial(h_j \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{h})}{g_{ij}}}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big)^2 + \Big(\frac{\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \sum_{ij} \frac{\partial(\psi_i \boldsymbol{e}_j^\top \boldsymbol{G}^\top\boldsymbol{\psi})}{g_{ij}}}{n\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2}\Big)^2\Big]$$
$$\leq C_{14}\mathbb{E}\Big[n + p + \|\boldsymbol{G}\|_{op}^2 + (n+p)\sum_{i=1}^n\sum_{j=1}^p \frac{1 + \|\boldsymbol{G}\|_{op}^2/n}{(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^2}\Big(\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{n}\Big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2\Big)\Big] \tag{B.13}$$

*for some positive absolute constant in the second line.*

Proposition 14 is proved in Appendix C; it is a consequence of (Bellec, 2020, Proposition 6.3). By Proposition 14 combined with the identities (B.6)-(B.7)-(B.8), and by showing that the purple-colored terms in (B.6)-(B.7)-(B.8) are negligible, we obtain the following.

**Proposition 15** *Let Assumption A be fulfilled. Then*

$$\mathbb{E}\Big[\Big\{n^{-\frac{1}{2}}(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1}\big(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\|\boldsymbol{\psi}\|^2 + \mathrm{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2\big)\Big\}^2\Big] \leq C_{15}(\gamma,\mu), \tag{B.14}$$
$$\mathbb{E}\Big[\Big\{n^{-\frac{1}{2}}(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1}\big(\tfrac{1}{n}\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \tfrac{p-\hat{\mathsf{df}}}{n}\|\boldsymbol{\psi}\|^2 + \tfrac{\mathrm{tr}[\boldsymbol{V}]}{n}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}\big)\Big\}^2\Big] \leq C_{16}(\gamma,\mu), \tag{B.15}$$
$$\mathbb{E}\Big[\Big\{n^{-\frac{1}{2}}(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1}\big(\|\boldsymbol{G}\boldsymbol{h}\|^2 - \mathrm{tr}[\boldsymbol{A}]\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - (n-\hat{\mathsf{df}})\|\boldsymbol{h}\|^2\big)\Big\}^2\Big] \leq C_{17}(\gamma,\mu). \tag{B.16}$$

**Proof** We bound from above the derivatives in (B.13). For the norm of $(\partial/\partial g_{ij})\boldsymbol{h}$ and $(\partial/\partial g_{ij})\boldsymbol{\psi}$, by (B.3)-(B.2) and $\frac{1}{2}(a+b)^2 \leq a^2 + b^2$,

$$\sum_{ij}\frac{1}{2}\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 \leq \|\boldsymbol{A}\|_F^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2\|\boldsymbol{h}\|^2,\ \sum_{ij}\frac{1}{2n}\Big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2 \leq \frac{\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_F^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{V}\|_F^2\|\boldsymbol{h}\|^2}{n}.$$

Using $\|\boldsymbol{A}\|_{op} \leq 1/(n\mu)$, $\|\boldsymbol{D}\|_{op} \leq 1$, $p/n \leq \gamma$ and $\boldsymbol{V}$ in (2.3), it follows that in (B.13) we have

$$\frac{1}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\sum_{i=1}^n\sum_{j=1}^p\Big(\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{n}\Big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2\Big) \leq C_{18}(\gamma,\mu)\Big(1 + \frac{\|\boldsymbol{G}\|_{op}^2}{n}\Big). \tag{B.17}$$

Since $\mathbb{E}[\|n^{-1/2}\boldsymbol{G}\|_{op}^4] \leq C_{19}(\gamma)$ (Davidson and Szarek, 2001, Theorem II.13), this shows that (B.13) is bounded from above by $C_{20}(\gamma,\mu)n$. The contractions appearing in the left-hand side of (B.13) are given in (B.6)-(B.7)-(B.8), so that it remains to bound from above the purple colored terms in these three equations. This is done by using the upper bounds on the operator norms $\|\boldsymbol{A}\|_{op} \leq 1/(n\mu)$, $\|\boldsymbol{D}\|_{op} \leq 1$ and again that $\mathbb{E}[\|n^{-1/2}\boldsymbol{G}\|_{op}^4] \leq C_{21}(\gamma)$, so that (B.13) yields the three inequalities in Proposition 15. ∎

The next result is another probabilistic result where the contractions in (B.4)-(B.5) appear.

**Proposition 16** *Let $\boldsymbol{h}:\mathbb{R}^{n\times p}\to\mathbb{R}^p$, $\boldsymbol{\psi}:\mathbb{R}^{n\times p}\to\mathbb{R}^n$ be locally Lipschitz functions. If $\boldsymbol{G}\in\mathbb{R}^{n\times p}$ has iid $N(0,1)$ entries then*

$$\mathbb{E}\Big[\frac{\big|\frac{p}{n}\|\boldsymbol{\psi}\|^2-\frac{1}{n}\sum_{j=1}^p\big(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j-\sum_{i=1}^n\frac{\partial\psi_i}{\partial g_{ij}}\big)^2\big|}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}\Big]+\mathbb{E}\Big[\frac{\big|n\|\boldsymbol{h}\|^2-\sum_{i=1}^n\big(\boldsymbol{g}_i^\top\boldsymbol{h}-\sum_{j=1}^p\frac{\partial h_j}{\partial g_{ij}}\big)^2\big|}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}\Big]$$
$$\le C_{22}\Big(\sqrt{n+p}(1+\Xi^{1/2})+\Xi\Big)\ \textit{where}\ \Xi=\mathbb{E}\Big[\frac{1}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}\sum_{i=1}^n\sum_{j=1}^p\Big(\Big\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\Big\|^2+\frac{1}{n}\Big\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2\Big)\Big].$$

Proposition 16 is proved in Appendix C; it is a consequence of (Bellec, 2020, Theorem 7.1). Using the contractions (B.4)-(B.5) in the left-hand side of Proposition 16, and by showing that the purple colored terms are negligible, we obtain the following two inequalities.

**Proposition 17** *Let Assumption A be fulfilled. Then*

$$\mathbb{E}\big|n^{-\frac12}(\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n)^{-1}\big(\tfrac{p}{n}\|\boldsymbol{\psi}\|^2-\tfrac{1}{n}\|\boldsymbol{G}^\top\boldsymbol{\psi}+\operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2\big)\big|\le C_{23}(\gamma,\mu),\tag{B.18}$$
$$\mathbb{E}|n^{-\frac12}(\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n)^{-1}\big(n\|\boldsymbol{h}\|^2-\|\boldsymbol{G}\boldsymbol{h}-\operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2\big)|\le C_{24}(\gamma,\mu).\tag{B.19}$$

**Proof** For $\Xi$ in Proposition 16, the fact that $\Xi\le C_{25}(\gamma,\mu)$ is already proved in (B.17). For the first inequality we use Proposition 16 and the contraction (B.5). To control the purple terms in (B.5) inside the left-hand side of Proposition 17,

$$\Big|\sum_{j=1}^p(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j-\sum_{i=1}^n\frac{\partial\psi_i}{\partial g_{ij}})^2-\|\boldsymbol{G}^\top\boldsymbol{\psi}+\operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2\Big|=\Big|\boldsymbol{\psi}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\Big(2\boldsymbol{G}^\top\boldsymbol{\psi}+2\operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}+\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi}\Big)\Big|$$
$$\le(\|\boldsymbol{\psi}\|^2/n+\|\boldsymbol{h}\|^2)(2n\|\boldsymbol{G}\|_{op}^2\|\boldsymbol{A}\|_{op}+2\sqrt{n}\|\boldsymbol{G}\|_{op}\|\boldsymbol{A}\|_{op}+n\|\boldsymbol{A}\|_{op}^2\|\boldsymbol{G}\|_{op}^2)$$

thanks to $|\operatorname{tr}\boldsymbol{V}|\le n$ in Theorem 1. With the bound obtained by multiplying the previous display by $n^{-3/2}(\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n)^{-1}$, and using the previous bounds on $\|\boldsymbol{A}\|_{op}$ and $\mathbb{E}[\|n^{-1/2}\boldsymbol{G}\|_{op}^2]$, we obtain (B.18) from Proposition 16 and (B.5). The second claim is obtained by Proposition 16, the contraction (B.4) and an argument similar to the previous display bound the purple term in (B.4). ■

We are now ready to prove Theorem 9.

**Proof (Proof of Theorem 9)** Define

$$\begin{aligned}
\xi_I&=\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}-\operatorname{tr}[\boldsymbol{A}]\|\boldsymbol{\psi}\|^2+\operatorname{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2 && \text{(bounded in (B.14))},\\
\xi_{II}&=\tfrac{1}{n}\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2-\tfrac{p-\hat{\mathsf{df}}}{n}\|\boldsymbol{\psi}\|^2+\tfrac{\operatorname{tr}[\boldsymbol{V}]}{n}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} && \text{(bounded in (B.15))},\\
\xi_{III}&=\|\boldsymbol{G}\boldsymbol{h}\|^2-\operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}-(n-\hat{\mathsf{df}})\|\boldsymbol{h}\|^2 && \text{(bounded in (B.16))},\\
\xi_{IV}&=\tfrac{p}{n}\|\boldsymbol{\psi}\|^2-\tfrac{1}{n}\|\boldsymbol{G}^\top\boldsymbol{\psi}+\operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 && \text{(bounded in (B.18))},\\
\xi_V&=n\|\boldsymbol{h}\|^2-\|\boldsymbol{G}\boldsymbol{h}-\operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 && \text{(bounded in (B.19))}.
\end{aligned}$$

Then by expanding the square in $\xi_{IV}$ and $\xi_V$ and simple algebra (for instance by computing first $\xi_{II}+\xi_{IV}$ and $\xi_{III}+\xi_V$ separately),

$$(\operatorname{tr}[\boldsymbol{V}]/n-\operatorname{tr}\boldsymbol{A})\xi_I+\xi_{II}+\xi_{III}+\xi_{VI}+\xi_V=(\|\boldsymbol{\psi}\|^2/n+\|\boldsymbol{h}\|^2)(\hat{\mathsf{df}}-\operatorname{tr}[\boldsymbol{A}]\operatorname{tr}[\boldsymbol{V}]).$$

Since $|\operatorname{tr}[\boldsymbol{V}]/n \leq 1$, $\operatorname{tr}[\boldsymbol{A}] \leq \gamma/\mu$ by Theorem 1, the previous display divided by $n^{1/2}(\|\boldsymbol{\psi}\|^2/n + \|\boldsymbol{h}\|^2)$ and the bounds (B.14), (B.15), (B.16), (B.18) and (B.19) complete the proof. ■

To prove Theorem 6, we need this extra proposition whose proof is closely related to Proposition 15.

**Proposition 18** *Let Assumption A be fulfilled. Then*

$$\mathbb{E}\big[\big\{(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-\frac{1}{2}}\|\boldsymbol{\varepsilon}\|^{-1}\xi_{VI}\big\}^2\big] \leq C_{26}(\gamma,\mu) \quad \textit{for} \quad \xi_{VI} = \boldsymbol{\varepsilon}^\top(\boldsymbol{G}\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}). \tag{B.20}$$

Proposition 18 is proved in Appendix C. We are now ready to prove Theorem 6.

**Proof (Proof of Theorem 6)** We have $n\|\boldsymbol{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2 - \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 = \xi_V + 2\xi_{VI}$ by simple algebra and the definitions of $\xi_V$ and $\xi_{VI}$. Hence

$$\mathbb{E}\Big[\frac{\big|\|\boldsymbol{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2/n\big|}{\max\{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n, (\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}(\|\boldsymbol{\varepsilon}\|^2/n)^{1/2}\}}\Big] \leq n^{-1/2}C_{27}(\gamma,\mu) \tag{B.21}$$

thanks to (B.20) and (B.19). ■

**Proof (Proof of Theorem 7)** We perform the change of variable (B.1) to $\widetilde{\boldsymbol{\beta}}$ as well, giving $\widetilde{\boldsymbol{h}}$ (the counterpart of $\boldsymbol{h}$), $\widetilde{\boldsymbol{\psi}}$ (counterpart of $\boldsymbol{\psi}$) and $\widetilde{\boldsymbol{A}}$ (counterpart of $\boldsymbol{A}$). Let $\Omega$ be the event defined in the theorem, i.e,

$$\Omega = \{\|\boldsymbol{G}\|_{op} \leq 2\sqrt{n} + \sqrt{p}\} \cap \{\|\boldsymbol{\varepsilon}\|^2 \leq n^{2/(1+q)}\}. \tag{B.22}$$

Then $\mathbb{P}(\Omega^c) \to 0$ by (Davidson and Szarek, 2001, Theorem 2.13) for the first event and Pinelis (2021) to show that $\|\boldsymbol{\varepsilon}\|^2/n^{2/(1+q)} \to^{\mathbb{P}} 0$ under the assumption that $\mathbb{E}[|\varepsilon_i|^{1+q}]$ is bounded.

Under Assumption B, $I_\Omega(\|\boldsymbol{\psi}\|^2/n + \|\boldsymbol{h}\|^2)$ is bounded by a constant. Indeed, since the penalty $g$ is minimized at $\mathbf{0}$, $(\widehat{\boldsymbol{\beta}} - \mathbf{0})^\top \boldsymbol{X}^\top\boldsymbol{\psi} \in n(\widehat{\boldsymbol{\beta}} - \mathbf{0})^\top(\partial g(\widehat{\boldsymbol{\beta}}) - \partial g(\mathbf{0}))$ since $\mathbf{0} \in \partial g(\mathbf{0})$. By strong convexity of $g$ in Assumption A, $(\widehat{\boldsymbol{\beta}} - \mathbf{0})^\top \boldsymbol{X}^\top\boldsymbol{\psi} \geq \mu\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}}\|^2$. In $\Omega$, this implies $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}}\| \leq \frac{1}{\mu n}\|\boldsymbol{G}\|_{op}\|\boldsymbol{\psi}\| \leq C_{28}(\gamma,\mu)\|\boldsymbol{\psi}\|/\sqrt{n}$ and $\|\boldsymbol{\psi}\|/\sqrt{n} \leq M$ in Assumption B. Since $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*\|^2 \leq M$ in Assumption B, this yields $I_\Omega(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n) \leq C_{29}(\gamma,\mu,M)$ and the same holds for $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{\psi}}$: $I_\Omega(\|\widetilde{\boldsymbol{h}}\|^2 + \|\widetilde{\boldsymbol{\psi}}\|^2/n) \leq C_{30}(\gamma,\mu,M)$.

Inequality (B.21) thus implies

$$\mathbb{E}[I_\Omega(\big|\|\boldsymbol{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2/n\big| + \big|\|\widetilde{\boldsymbol{h}}\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\widetilde{\boldsymbol{r}} + \operatorname{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\boldsymbol{\psi}}\|^2/n\big|)] \leq C_{31}(\gamma,\mu,M)(n^{-1/2} \vee n^{-q/(1+q)}).$$

Since $q \in (0,1)$ we have $n^{-1/2} \vee n^{-q/(1+q)} = n^{-q/(1+q)}$ in the right-hand side. Let $\hat{\Omega} = \{\|\boldsymbol{h}\|^2 - \|\widetilde{\boldsymbol{h}}\|^2 > \eta, \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 \leq \|\widetilde{\boldsymbol{r}} + \operatorname{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\boldsymbol{\psi}}\|^2\}$ be the event for which we are trying to control the probability. By the triangle inequality,

$$\mathbb{E}[I_\Omega\big|\|\boldsymbol{h}\|^2 - \|\widetilde{\boldsymbol{h}}\|^2 - \|\boldsymbol{r} + \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2/n + \|\widetilde{\boldsymbol{r}} + \operatorname{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\boldsymbol{\psi}}\|^2/n\big|] \leq C_{32}(\gamma,\mu,M)n^{-q/(1+q)}.$$

In $\hat{\Omega}$, the random variable in the expectation sign is larger than $\eta I_\Omega$. Thus $\eta\mathbb{E}[I_\Omega I_{\hat{\Omega}}] \leq C_{33}(\gamma,\mu,M)n^{-q/(1+q)}$ and $\mathbb{P}(\hat{\Omega}) \leq \eta^{-1}C_{34}(\gamma,\mu,M)n^{-q/(1+q)} + \mathbb{P}(\Omega^c)$. ■

**Proof (Proof of Theorem 8)** We follow the same strategy. Let $\Omega$ be the same event as in the previous proof, so that $\mathbb{P}(\Omega^c) \to 0$ as before. We perform the change of variable (B.1) for each

$k = 1, ..., K$ giving $\boldsymbol{h}_k$, $\boldsymbol{\psi}_k$ and $\boldsymbol{A}_k$. We have $I_\Omega \max_{k=1,...,K}(\|\boldsymbol{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \le C_{35}(\gamma, \mu, M)$ as explained in the previous proof.

Summing over $k$ the inequality (B.21) gives $\mathbb{E}[I_\Omega \sum_{k=1}^K |\|\boldsymbol{h}_k\|^2 + \|\boldsymbol{\varepsilon}\|^2 - \|\boldsymbol{r}_k + \operatorname{tr}[\boldsymbol{A}_k]\boldsymbol{\psi}_k\|^2|] \le KC_{36}(\gamma, \mu, M)n^{-q/(1+q)}$. Let $\hat{k}$ be the minimizer of $\|\boldsymbol{r}_k + \operatorname{tr}[\boldsymbol{A}_k]\boldsymbol{\psi}_k\|^2$ as defined in the statement of Theorem 8 and let $\tilde{k} \in \{1, ..., K\}$ be such that $\|\boldsymbol{h}_{\hat{k}}\|^2 \ge \|\boldsymbol{h}_{\tilde{k}}\|^2 + \eta$ in the event $\tilde{\Omega}$ where such $\tilde{k}$ exists. Then by the triangle inequality, $\eta\mathbb{E}[I_\Omega I_{\tilde{\Omega}}] \le C_{37}(\gamma, \mu, M)n^{-q/(1+q)}$. It follows that $\mathbb{P}(\tilde{\Omega}) \le \eta^{-1}C_{38}(\gamma, \mu, M)n^{-q/(1+q)} + \mathbb{P}(\Omega^c) \to 0$ as desired. ∎

**Proof (Proof of Theorem 11)** Using $\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2 = (\boldsymbol{a} - \boldsymbol{b})^\top(\boldsymbol{a} + \boldsymbol{b})$ we have

$$\|\boldsymbol{G}\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}\boldsymbol{h} - (\hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}])\|^2 = (\hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}] - \operatorname{tr}[\boldsymbol{A}])\boldsymbol{\psi}^\top(2\boldsymbol{G}\boldsymbol{h} - (\operatorname{tr}[\boldsymbol{A}] + \hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}])\boldsymbol{\psi}).$$

Hence using $|\operatorname{tr}[\boldsymbol{A}]| \le \gamma/\mu$, $|\hat{\mathsf{df}}| \le n$ and the Cauchy-Schwarz inequality

$$\begin{aligned}&|\|\boldsymbol{G}\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}\boldsymbol{h} - (\hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}])\|^2| \\ &\le C_{39}(\gamma, \mu)(\tfrac{n}{\operatorname{tr}\boldsymbol{V}} \vee 1)|\hat{\mathsf{df}}/n - \operatorname{tr}[\boldsymbol{V}]\operatorname{tr}[\boldsymbol{A}]/n|(\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{G}\|_{op}\|\boldsymbol{h}\|^2).\end{aligned}$$

Let $\Omega$ be the event in Theorem 7. Using the bound on the operator norm of $\boldsymbol{G}$ in $\Omega$, for any deterministic $\eta > 0$ we have proved

$$\mathbb{E}\Big[I\{\Omega\}I\{\operatorname{tr}[\boldsymbol{V}]n \ge \eta\}\frac{|\|\boldsymbol{G}\boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}\boldsymbol{h} - (\hat{\mathsf{df}}/\operatorname{tr}[\boldsymbol{V}])\|^2|}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big] \le \frac{C_{40}(\gamma, \mu)}{\eta \wedge 1}n^{1/2}$$

thanks to Theorem 9. By (D.8), in the event $\Omega$ where the operator norm of $\|n^{-1/2}\boldsymbol{G}\|_{op}$ is bounded by a constant, $\operatorname{tr}[\boldsymbol{V}] \ge \operatorname{tr}[\operatorname{diag}\{\psi'(\boldsymbol{r})\}]/C_{41}(\gamma, \mu)$. Hence combining the previous display with (B.21), we have proved

$$\mathbb{E}\Big[\frac{I\{\Omega\}I\{\sum_{i=1}^n \psi'(r_i) \ge n\eta\}\big|\|\boldsymbol{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\boldsymbol{r} + \frac{\hat{\mathsf{df}}}{\operatorname{tr}\boldsymbol{V}}\boldsymbol{\psi}\|^2/n\big|}{\max\{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n, (\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}(\|\boldsymbol{\varepsilon}\|^2/n)^{1/2}\}}\Big] \le \frac{C_{42}(\gamma, \mu, \eta)}{\sqrt{n}}.$$

At this point the proof is similar to that of Theorem 8: We perform the change of variable (B.1) for each $k = 1, ..., K$ giving $\boldsymbol{h}_k$, $\boldsymbol{\psi}_k$, $\hat{\mathsf{df}}_k$ and $\boldsymbol{V}_k$. We have $I_\Omega \max_{k=1,...,K}(\|\boldsymbol{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \le C_{43}(\gamma, \mu, M)$ as explained in the previous proofs. Summing over $k = 1, ..., K$ the previous display, using $I_\Omega \max_{k=1,...,K}(\|\boldsymbol{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \le C_{44}(\gamma, \mu, M)$ and $I_\Omega\|\boldsymbol{\varepsilon}\|^2 \le n^{2/(1+q)}$ we find

$$\mathbb{E}\Big[\sum_{k=1}^K I\{\Omega\}I\{\sum_{i=1}^n \psi_k'(r_{ki}) \ge n\eta\}\Big|\|\boldsymbol{h}_k\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\boldsymbol{r}_k + \frac{\hat{\mathsf{df}}_k}{\operatorname{tr}\boldsymbol{V}_k}\boldsymbol{\psi}_k\|^2/n\Big|\Big] \le \frac{KC_{45}(\gamma, \mu, \eta)}{n^{q/(1+q)}}.$$

Let $\tilde{\Omega}$ be the event that there exists $\tilde{k}$ with $\frac{1}{n}\sum_{i=1}^n \psi_{\tilde{k}}'(r_{\tilde{k}i}) \ge \eta$ satisfying $\|\boldsymbol{h}_{\tilde{k}}\|^2 + \tilde{\eta} \le \|\boldsymbol{h}_{\hat{k}}\|^2$, then by the previous display and the triangle inequality, using $\|\boldsymbol{r}_{\hat{k}} + \frac{\hat{\mathsf{df}}_{\hat{k}}}{\operatorname{tr}\boldsymbol{V}_{\hat{k}}}\boldsymbol{\psi}_{\hat{k}}\|^2 \le \|\boldsymbol{r}_{\tilde{k}} + \frac{\hat{\mathsf{df}}_{\tilde{k}}}{\operatorname{tr}\boldsymbol{V}_{\tilde{k}}}\boldsymbol{\psi}_{\tilde{k}}\|^2$ by definition of $\hat{k}$, we obtain $\tilde{\eta}\mathbb{P}(I_\Omega I_{\tilde{\Omega}}) = O(K/n^{q/(1+q)})$. Since $\tilde{\eta}$ is a constant independent of $n, p$ and $\mathbb{P}(\Omega) \to 1$, the probability $\mathbb{P}(\tilde{\Omega})$ converge to 0 if $K = o(n^{q/(1+q)})$. ∎

## Appendix C. Probabilistic results and their proofs

**Proposition 13** *[Variant of Bellec and Zhang (2019)] Let $\boldsymbol{z} \in N(\boldsymbol{0}, \boldsymbol{I}_q)$ and $\boldsymbol{f} := \boldsymbol{f}(\boldsymbol{z}) : \mathbb{R}^q \to \mathbb{R}^q \setminus \{\boldsymbol{0}\}$ be locally Lipschitz in $\boldsymbol{z}$ with $\mathbb{E}[\|\boldsymbol{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \boldsymbol{f}}{\partial z_k}\|^2] < +\infty$. Then*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{f}^\top \boldsymbol{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\boldsymbol{f}\|_2} - Z\Big)^2\Big] \le (7 + 2\sqrt{6}) \mathbb{E}\Big[\|\boldsymbol{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \boldsymbol{f}}{\partial z_k}\|^2\Big] < +\infty. \tag{B.9}$$

**Proof** Let $\boldsymbol{g} := \boldsymbol{g}(\boldsymbol{z}) = \frac{\boldsymbol{f}(\boldsymbol{z})}{\|\boldsymbol{f}(\boldsymbol{z})\|} - \mathbb{E}[\frac{\boldsymbol{f}(\boldsymbol{z})}{\|\boldsymbol{f}(\boldsymbol{z})\|}]$ and set

$$Z = \boldsymbol{z}^\top \mathbb{E}[\frac{\boldsymbol{f}(\boldsymbol{z})}{\|\boldsymbol{f}(\boldsymbol{z})\|}] \Big/ \sqrt{V}, \qquad V = \Big\|\mathbb{E}\Big[\frac{\boldsymbol{f}(\boldsymbol{z})}{\|\boldsymbol{f}(\boldsymbol{z})\|}\Big]\Big\|^2$$

so that $Z \sim N(0,1)$ and $V$ is deterministic with $V \le 1$ by Jensen's inequality. As a first step, we proceed to prove inequality

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{f}^\top \boldsymbol{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\boldsymbol{f}\|_2} - \sqrt{V} Z\Big)^2\Big] \le 6\, \mathbb{E}\Big[\|\boldsymbol{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \boldsymbol{f}}{\partial z_k}\|^2\Big]. \tag{C.1}$$

Then at any point $\boldsymbol{z}$ where $\boldsymbol{f}$ is differentiable we have

$$\frac{\partial \boldsymbol{g}}{\partial z_k} = \|\boldsymbol{f}(\boldsymbol{z})\|^{-1} \widehat{\boldsymbol{P}} \frac{\partial \boldsymbol{f}}{\partial z_k}, \qquad \text{where} \qquad \widehat{\boldsymbol{P}} = \boldsymbol{I}_q - \frac{\boldsymbol{f}\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^2}.$$

This implies that almost surely,

$$\frac{\boldsymbol{f}^T \boldsymbol{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\boldsymbol{f}\|_2} - \sqrt{V} Z = \boldsymbol{g}^T \boldsymbol{z} - \sum_{k=1}^q (\partial/\partial z_k) g_k - \frac{\boldsymbol{f}^T (\partial \boldsymbol{f}/\partial \boldsymbol{z}) \boldsymbol{f}}{\|\boldsymbol{f}\|^3}$$

where $\partial \boldsymbol{f}/\partial \boldsymbol{z}$ is the matrix with entries $(l,k)$ entry $(\partial/\partial z_k) f_l$ for all, $k, l = 1, ..., q$.

By the triangle inequality and $(a+b)^2 \le 2a^2 + 2b^2$, this implies that the left-hand side of (C.1) is bounded from above by $2\mathbb{E}[(\boldsymbol{z}^T \boldsymbol{g} - \operatorname{tr}[\partial \boldsymbol{g}/\partial \boldsymbol{z}])^2] + 2\mathbb{E}[\|\boldsymbol{f}\|^{-2} \|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F^2]$. The first term can be bounded using the main result of Bellec and Zhang (2018) and the Gaussian Poincaré inequality (Boucheron et al., 2013, Theorem 3.20)

$$\mathbb{E}[(\boldsymbol{z}^T \boldsymbol{g} - \operatorname{tr}[\partial \boldsymbol{g}/\partial \boldsymbol{z}])^2] = \mathbb{E}[\|\boldsymbol{g}\|^2] + \mathbb{E} \operatorname{tr}[(\partial \boldsymbol{g}/\partial \boldsymbol{z})^2] \le 2\mathbb{E}[\|\partial \boldsymbol{g}/\partial \boldsymbol{z}\|_F^2].$$

This proves (C.1). To bound $|\sqrt{V} - 1|$, we have by the triangle inequality

$$|\sqrt{V} - 1| = |\sqrt{V} - \big\|\tfrac{\boldsymbol{f}}{\|\boldsymbol{f}\|}\big\|| \le \big\|\mathbb{E}[\tfrac{\boldsymbol{f}}{\|\boldsymbol{f}\|}] - \tfrac{\boldsymbol{f}}{\|\boldsymbol{f}\|}\big\| = \|\boldsymbol{g}\|.$$

By another application of the Gaussian Poincaré inequality,

$$|\sqrt{V} - 1|^2 \le \mathbb{E}[\|\boldsymbol{g}\|_2^2] \le \mathbb{E}[\|\partial \boldsymbol{g}/\partial \boldsymbol{z}\|_F^2] \le \mathbb{E}[\|\boldsymbol{f}\|^{-2} \|\partial \boldsymbol{f}/\partial \boldsymbol{z}\|_F^2]. \tag{C.2}$$

Combining Equations (C.1) and (C.2) using $(a+b)^2 = a^2 + 2ab + b^2 \le a^2 + 1/\sqrt{6} a^2 + \sqrt{6} b^2 + b^2$, we obtain the constant $7 + 2\sqrt{6}$.

■

**Proposition 14** *Let $\boldsymbol{h} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$, $\boldsymbol{\psi} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ be locally Lipschitz functions. If $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has iid $N(0,1)$ entries then*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \sum_{ij}\frac{\partial(\psi_i h_j)}{g_{ij}}}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big)^2 + \Big(\frac{\|\boldsymbol{G}\boldsymbol{h}\|^2 - \sum_{ij}\frac{\partial(h_j \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{h})}{g_{ij}}}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big)^2 + \Big(\frac{\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \sum_{ij}\frac{\partial(\psi_i \boldsymbol{e}_j^\top \boldsymbol{G}^\top\boldsymbol{\psi})}{g_{ij}}}{n\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2}\Big)^2\Big]$$
$$\le C_{46}\mathbb{E}\Big[n + p + \|\boldsymbol{G}\|_{op}^2 + (n+p)\sum_{i=1}^n\sum_{j=1}^p \frac{1 + \|\boldsymbol{G}\|_{op}^2/n}{(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^2}\Big(\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{n}\Big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2\Big)\Big] \tag{B.13}$$

*for some positive absolute constant in the second line.*

**Proof (Proof of Proposition 14)** We prove the claim separately for the three terms in the left-hand side of Proposition 14; we start with the first of the three terms. We will apply the probabilistic result given in Proposition 6.3 in Bellec (2020): if $\boldsymbol{\eta} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ and $\boldsymbol{\rho} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ are locally Lipschitz and $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has iid $N(0,1)$ entries,

$$\mathbb{E}\Big[\Big(\boldsymbol{\rho}^\top \boldsymbol{G}\boldsymbol{\eta} - \sum_{ij}\frac{\partial(\rho_i\eta_j)}{g_{ij}}\Big)^2\Big] \le \mathbb{E}\Big[\|\boldsymbol{\rho}\|^2\|\boldsymbol{\eta}\|^2\Big] + 2\mathbb{E}\Big[\sum_{ij}\|\boldsymbol{\eta}\|^2\|\frac{\partial\boldsymbol{\rho}}{\partial g_{ij}}\|^2 + \|\boldsymbol{\rho}\|^2\|\frac{\partial\boldsymbol{\eta}}{\partial g_{ij}}\|^2\Big]. \tag{C.3}$$

The proof only relies on Gaussian integration by parts to transform the left-hand side. Let $\boldsymbol{f} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n+p}$ be locally Lipschitz. For any $i, j$ and at a point where both $\boldsymbol{h}$ and $\boldsymbol{\psi}$ are differentiable and $\boldsymbol{f} \neq \boldsymbol{0}$,

$$\frac{\partial}{\partial g_{ij}}\Big(\frac{\boldsymbol{f}}{\|\boldsymbol{f}\|}\Big) = \frac{1}{\|\boldsymbol{f}\|}\Big(\boldsymbol{I}_{n+p} - \frac{\boldsymbol{f}\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^2}\Big)\frac{\partial \boldsymbol{f}}{\partial g_{ij}} \quad \text{so that} \quad \Big\|\frac{\partial}{\partial g_{ij}}\Big(\frac{\boldsymbol{f}}{\|\boldsymbol{f}\|}\Big)\Big\|^2 \le \frac{1}{\|\boldsymbol{f}\|^2}\Big\|\frac{\partial \boldsymbol{f}}{\partial g_{ij}}\Big\|^2.$$

We use this inequality applied with

$$\boldsymbol{f} = (\boldsymbol{h}, \tfrac{1}{\sqrt{n}}\boldsymbol{\psi}), \qquad \boldsymbol{\rho} = \tfrac{1}{\sqrt{n}}\tfrac{\boldsymbol{\psi}}{\|\boldsymbol{f}\|}, \qquad \boldsymbol{\eta} = \tfrac{\boldsymbol{h}}{\|\boldsymbol{f}\|}. \tag{C.4}$$

To bound from above the right-hand side of (C.3), the inequality in the previous display can be rewritten

$$\|\frac{\partial\boldsymbol{\eta}}{\partial g_{ij}}\|^2 + \|\frac{\partial\boldsymbol{\rho}}{\partial g_{ij}}\|^2 \le \frac{1}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big(\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|^2\Big). \tag{C.5}$$

Since $\|\boldsymbol{\rho}\| \le 1$ and $\|\boldsymbol{\eta}\| \le 1$ by definition, the right-hand side of (C.3) is bounded from above by $1 + 2\mathbb{E}[\frac{1}{\|\boldsymbol{h}\|^2+\|\boldsymbol{\psi}\|^2/n}(\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|^2)]$. Thus the proof of Proposition 14 for the first term in the left-hand side is almost complete; it remains to control inside the parenthesis of the left-hand side,

$$\sum_{ij}\frac{1}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\frac{\partial(\psi_i n^{-1/2}h_j)}{\partial g_{ij}} - \frac{\partial}{\partial g_{ij}}\Big(\frac{\psi_i n^{-1/2}h_j}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big) = 2\sum_{ij}\psi_i n^{-1/2}h_j\frac{\boldsymbol{h}^\top\frac{\partial\boldsymbol{h}}{\partial g_{ij}} + \frac{1}{n}\boldsymbol{\psi}^\top\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}}{(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^2}.$$

By multiple applications of the Cauchy-Schwartz inequality, the absolute value of the previous display is bounded from above by $2(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1/2}(\sum_{ij}\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\| + \frac{1}{n}\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\|)^{1/2}$. This completes the proof of Proposition 14 for the first term in the left-hand side.

For the second and third term in the left-hand side of Proposition 14, apply instead (C.3) to $\boldsymbol{\rho} = \boldsymbol{G}\boldsymbol{\eta}$ and $\boldsymbol{\eta} = \boldsymbol{G}^\top \boldsymbol{\rho}$ to obtain

$$\mathbb{E}\Big[\Big(\|\boldsymbol{G}\boldsymbol{\eta}\|^2 - \sum_{ij} \frac{\partial(\eta_j \boldsymbol{e}_i^\top \boldsymbol{G}\boldsymbol{\eta})}{g_{ij}}\Big)^2\Big] \le \mathbb{E}\Big[\|\boldsymbol{G}\boldsymbol{\eta}\|^2\|\boldsymbol{\eta}\|^2\Big] + 2\mathbb{E}\Big[\sum_{ij} \|\boldsymbol{\eta}\|^2 \|\boldsymbol{e}_i\eta_j + \boldsymbol{G}\frac{\partial \boldsymbol{\eta}}{\partial g_{ij}}\|^2 + \|\boldsymbol{G}\boldsymbol{\eta}\|^2 \|\frac{\partial \boldsymbol{\eta}}{\partial g_{ij}}\|^2\Big],$$

$$\mathbb{E}\Big[\Big(\|\boldsymbol{G}^\top\boldsymbol{\rho}\|^2 - \sum_{ij} \frac{\partial(\rho_i \boldsymbol{\rho}^\top \boldsymbol{G}\boldsymbol{e}_j)}{g_{ij}}\Big)^2\Big] \le \mathbb{E}\Big[\|\boldsymbol{G}^\top\boldsymbol{\rho}\|^2\|\boldsymbol{\rho}\|^2\Big] + 2\mathbb{E}\Big[\sum_{ij} \|\boldsymbol{G}^\top\boldsymbol{\rho}\|^2 \|\frac{\partial \boldsymbol{\rho}}{\partial g_{ij}}\|^2 + \|\boldsymbol{\rho}\|^2 \|\boldsymbol{e}_j\rho_i + \boldsymbol{G}^\top\frac{\partial \boldsymbol{\rho}}{\partial g_{ij}}\|^2\Big].$$

Setting $\boldsymbol{\rho} = \frac{1}{\sqrt{n}}\boldsymbol{\psi}/\|\boldsymbol{f}\|$, $\boldsymbol{\eta} = \boldsymbol{h}/\|\boldsymbol{f}\|$ we obtain the claim in Equation (C.3) by bounding the right-hand side of the previous displays using the operator norm of $\boldsymbol{G}$ and arguments similar to (C.5). The term involving $\frac{\partial}{\partial g_{ij}}\big(\frac{1}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\big)$ in the left-hand side is controlled similarly to the previous paragraph.

■

**Proposition 16** *Let $\boldsymbol{h} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$, $\boldsymbol{\psi} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ be locally Lipschitz functions. If $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has iid $N(0,1)$ entries then*

$$\mathbb{E}\Big[\frac{\big|\frac{p}{n}\|\boldsymbol{\psi}\|^2 - \frac{1}{n}\sum_{j=1}^p\big(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n \frac{\partial\psi_i}{\partial g_{ij}}\big)^2\big|}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big] + \mathbb{E}\Big[\frac{\big|n\|\boldsymbol{h}\|^2 - \sum_{i=1}^n\big(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j=1}^p \frac{\partial h_j}{\partial g_{ij}}\big)^2\big|}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\Big]$$
$$\le C_{47}\Big(\sqrt{n+p}(1+\Xi^{1/2}) + \Xi\Big) \text{ where } \Xi = \mathbb{E}\Big[\frac{1}{\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\sum_{i=1}^n\sum_{j=1}^p\Big(\Big\|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{n}\Big\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\Big\|^2\Big)\Big].$$

**Proof (Proof of Proposition 16)** We first focus on the first term in the left-hand side. Theorem 7.1 in Bellec (2020) provides that if $\boldsymbol{\rho} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ is locally Lipschitz with $\|\boldsymbol{\rho}\| \le 1$ then

$$\mathbb{E}\Big|p\|\boldsymbol{\rho}\|^2 - \sum_{j=1}^p\Big(\boldsymbol{\rho}^\top\boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n\frac{\partial\rho_i}{\partial g_{ij}}\Big)^2\Big| \le C_{48}\sqrt{p}\Big(1 + \mathbb{E}\sum_{ij}\Big\|\frac{\partial\boldsymbol{\rho}}{\partial g_{ij}}\Big\|^2\Big)^{1/2} + C_{49}\mathbb{E}\sum_{ij}\Big\|\frac{\partial\boldsymbol{\rho}}{\partial g_{ij}}\Big\|^2. \tag{C.6}$$

Let $\boldsymbol{\rho} = n^{-1/2}\boldsymbol{\psi}/\|\boldsymbol{f}\|$ as in (C.4). Inequality (C.5) lets us bound from above the right-hand side of the previous display by the right-hand side of Proposition 16. In the left-hand side, $p\|\boldsymbol{\rho}\|^2 = \frac{p}{n}\|\boldsymbol{\psi}\|^2/(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)$ as desired. For the left-hand side, using some algebra in (Bellec, 2020, Section 7), for any random vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ by the triangle and Cauchy-Schwarz inequalities we have

$$\begin{aligned}
|p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{a}\|^2| - |p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{b}\|^2| &\le \|\boldsymbol{a}-\boldsymbol{b}\|\|\boldsymbol{a}+\boldsymbol{b}\| \\
&\le \|\boldsymbol{a}-\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}-\boldsymbol{b}\|\|\boldsymbol{b}\| \\
&\le \|\boldsymbol{a}-\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}-\boldsymbol{b}\|(\sqrt{|\|\boldsymbol{b}\|^2 - p\|\boldsymbol{\rho}\|^2|} + \sqrt{p\|\boldsymbol{\rho}\|^2}) \\
&\le 3\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \tfrac{1}{2}|\|\boldsymbol{b}\|^2 - p\|\boldsymbol{\rho}\|^2| + 2\|\boldsymbol{a}-\boldsymbol{b}\|\sqrt{p\|\boldsymbol{\rho}\|^2}
\end{aligned}$$

so that $|p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{a}\|^2| \le \frac{3}{2}|p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{b}\|^2| + 3\|\boldsymbol{a}-\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}-\boldsymbol{b}\|\sqrt{p\|\boldsymbol{\rho}\|^2}$. Applying this to $b_j = \boldsymbol{\rho}^\top\boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n \frac{\partial\rho_i}{\partial g_{ij}}$ we use (C.6) to bound $|p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{b}\|^2|$ and $\|\boldsymbol{\rho}\| \le 1$ to bound

$\sqrt{p\|\boldsymbol{\rho}\|^2} \le \sqrt{p}$. It remains to specify $\boldsymbol{a}$ so that $|p\|\boldsymbol{\rho}\|^2 - \|\boldsymbol{a}\|^2|$ coincides with the first term in the left-hand side of Proposition 16 and bound $\|\boldsymbol{a}-\boldsymbol{b}\|$. Consequently, we set

$$a_j = \frac{\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j - \sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}}}{\sqrt{n}(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}} = \boldsymbol{\rho}^\top \boldsymbol{G}\boldsymbol{e}_j - \frac{\sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}}}{\sqrt{n}(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}} = b_j - \sum_{i=1}^n \frac{\psi_i}{\sqrt{n}} \frac{\partial (D^{-1})}{\partial g_{ij}}$$

where $D = (\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}$ so that by the Cauchy-Schwarz inequality $\|\boldsymbol{a}-\boldsymbol{b}\|^2 \le \frac{1}{n}\|\boldsymbol{\psi}\|^2 \sum_{ij} (\frac{\partial (D^{-1})}{\partial g_{ij}})^2$ and

$$\sum_{ij} \Big( \frac{\partial (D^{-1})}{\partial g_{ij}} \Big)^2 = \frac{1}{D^6} \sum_{ij} \Big( \boldsymbol{h}^\top \frac{\partial \boldsymbol{h}}{\partial g_{ij}} + \frac{\boldsymbol{\psi}}{\sqrt{n}}^\top \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \Big)^2 \le \frac{2}{D^4} \sum_{ij} \|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2. \quad \text{(C.7)}$$

using again the Cauchy-Schwarz inequality and $\max\{\|\boldsymbol{h}\|^2, \|\boldsymbol{\psi}\|^2/n\} \le D^2$. We obtain $\|\boldsymbol{a}-\boldsymbol{b}\|^2 \le D^{-2} \sum_{ij} \|\frac{\partial \boldsymbol{h}}{\partial g_{ij}}\|^2 + \frac{1}{n}\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2$ which completes the proof for the first term in the left-hand side of Proposition 16. For the second term in the left-hand side, the proof is similar with by exchanging the role of $n$ and $p$ in (C.6) and applying (C.6) to $\boldsymbol{h}/D$ instead of $\boldsymbol{\psi}/(\sqrt{n}D)$. ■

**Proposition 18** *Let Assumption A be fulfilled. Then*

$$\mathbb{E}\big[\big\{(\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-\frac{1}{2}} \|\boldsymbol{\varepsilon}\|^{-1} \xi_{VI}\big\}^2\big] \le C_{50}(\gamma,\mu) \quad \textit{for} \quad \xi_{VI} = \boldsymbol{\varepsilon}^\top(\boldsymbol{G}\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\boldsymbol{\psi}). \quad \text{(B.20)}$$

**Proof (Proof of Proposition 18)** Apply (C.3) with $\boldsymbol{\rho} = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$ and $\boldsymbol{\eta} = \boldsymbol{h}/D$ where $D = (\|\boldsymbol{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}$ as in the previous proof (this scalar $D$ is not related to the diagonal matrix $\boldsymbol{D} = \mathrm{diag}\{\psi'(\boldsymbol{r})\}$). Since $\boldsymbol{\varepsilon}$ has 0 derivative with respect to $\boldsymbol{G}$ we find

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{\varepsilon}^\top \boldsymbol{G}\boldsymbol{h}}{\|\boldsymbol{\varepsilon}\|D} - \sum_{ij} \frac{\varepsilon_i}{\|\boldsymbol{\varepsilon}\|} \frac{\partial (h_j D^{-1})}{\partial g_{ij}}\Big)^2\Big] \le 1 + 2\sum_{ij} \mathbb{E}[\|\frac{\partial \boldsymbol{\eta}}{\partial g_{ij}}\|^2].$$

The right-hand side is bounded from above by $C_{51}(\gamma,\mu)$ thanks to (C.5) and (B.17). For the second term above we use product rule and (B.4),

$$\sum_{ij} \frac{\varepsilon_i}{\|\boldsymbol{\varepsilon}\|} \frac{\partial (h_j D^{-1})}{\partial g_{ij}} = \frac{\mathrm{tr}[\boldsymbol{A}]\boldsymbol{\psi}^\top \boldsymbol{\varepsilon}}{D\|\boldsymbol{\varepsilon}\|} - \frac{\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{G}^\top \mathrm{diag}(\psi'(\boldsymbol{r}))\boldsymbol{\varepsilon}}{D\|\boldsymbol{\varepsilon}\|} + \sum_{ij} \frac{\varepsilon_i h_j}{\|\boldsymbol{\varepsilon}\|} \frac{\partial (D^{-1})}{\partial g_{ij}}.$$

To complete the proof we need to prove that the expectation of the square of the second and third terms colored in purple are bounded by $C_{52}(\gamma,\mu)$. Since $\|\boldsymbol{h}\| \le D$, the second term is bounded from above by $\|\boldsymbol{A}\|_{op}\|\boldsymbol{G}\|_{op}$ since $|\psi'| \le 1$ and $\mathbb{E}[\|\boldsymbol{A}\|_{op}^2 \|\boldsymbol{G}\|_{op}^2] \le C_{53}(\gamma,\mu)$ thanks to $\|\boldsymbol{A}\|_{op} \le 1/(n\mu)$ and (Davidson and Szarek, 2001, Theorem II.13). For the third term, we use the Cauchy-Schwarz inequality $(\sum_{ij} \frac{\varepsilon_i h_j}{\|\boldsymbol{\varepsilon}\|} \frac{\partial (D^{-1})}{\partial g_{ij}})^2 \le \|\boldsymbol{h}\|^2 \sum_{ij} (\frac{\partial (D^{-1})}{\partial g_{ij}})^2$, (C.7) and (B.17). ■

## Appendix D. Proof of differentiability results

**Theorem 1** *Let Assumption A be fulfilled. For almost every* $(\boldsymbol{y}, \boldsymbol{X})$ *the map* $(\boldsymbol{y}, \boldsymbol{X}) \mapsto \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ *is differentiable at* $(\boldsymbol{y}, \boldsymbol{X})$ *and there exists a matrix* $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ *depending on* $(\boldsymbol{X}, \boldsymbol{y})$ *with* $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \leq (n\mu)^{-1}$ *s.t.*

$$\begin{aligned}(\partial/\partial y_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i),\\ (\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) &= \widehat{\boldsymbol{A}}\boldsymbol{e}_j \psi(r_i) - \widehat{\boldsymbol{A}}\boldsymbol{X}^\top \boldsymbol{e}_i \psi'(r_i)\widehat{\beta}_j,\end{aligned} \qquad \text{where } r_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}, \tag{2.1}$$

$\boldsymbol{e}_i \in \mathbb{R}^n, \boldsymbol{e}_j \in \mathbb{R}^p$ *are canonical basis vectors ,* $\psi := \rho'$ *and* $\psi'$ *denote the derivatives. Furthermore,*

$$\hat{\mathsf{df}} = \mathrm{tr}[\boldsymbol{X}(\partial/\partial \boldsymbol{y})\hat{\boldsymbol{\beta}}] = \mathrm{tr}[\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top \mathrm{diag}\{\psi'(\boldsymbol{r})\}], \tag{2.2}$$

$$\boldsymbol{V} = \mathrm{diag}\{\psi'(\boldsymbol{r})\}(\boldsymbol{I}_n - \boldsymbol{X}(\partial/\partial \boldsymbol{y})\widehat{\boldsymbol{\beta}}) = \mathrm{diag}\{\psi'(\boldsymbol{r})\} - \mathrm{diag}\{\psi'(\boldsymbol{r})\}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top \mathrm{diag}\{\psi'(\boldsymbol{r})\} \tag{2.3}$$

*satisfy* $0 \leq \hat{\mathsf{df}} \leq n$ *and* $0 \leq \mathrm{tr}[\boldsymbol{V}] \leq n$.

The first part of the following proof is similar to the argument using the KKT conditions in Bellec (2020). After (D.3), the argument is novel and lets us derive the convenient formula (2.1) and the existence of matrix $\widehat{\boldsymbol{A}}$ which plays a central role in the contractions (B.4)-(B.8).

**Proof (Proof of Theorem 1)** $\boldsymbol{X}_t = \boldsymbol{X} + t\boldsymbol{U}$ and $\boldsymbol{y}_t = \boldsymbol{y} + t\boldsymbol{v}$ with $t \in \mathbb{R}$ where $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{v} \in \mathbb{R}^n$ are fixed. Let $\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}_t, \boldsymbol{X}_t)$ and $\widehat{\boldsymbol{r}}_t = \boldsymbol{y}_t - \boldsymbol{X}_t\widehat{\boldsymbol{\beta}}(\boldsymbol{y}_t, \boldsymbol{X}_t)$ and $\widehat{\boldsymbol{\psi}}(\boldsymbol{y}_t, \boldsymbol{X}_t) = \psi(\widehat{\boldsymbol{r}}_t)$. By convention, without arguments $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}$ refer to $(\boldsymbol{y}, \boldsymbol{X})$ which is $(\boldsymbol{y}_t, \boldsymbol{X}_t)$ at $t = 0$. By the KKT conditions, $\boldsymbol{X}^\top\widehat{\boldsymbol{\psi}} \in n\partial g(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{X}_t^\top\widehat{\boldsymbol{\psi}}_t \in n\partial g(\widehat{\boldsymbol{\beta}}_t)$, by strong convexity of $g$, we have

$$n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})\|^2 \leq (\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})^\top(\boldsymbol{X}_t^\top\widehat{\boldsymbol{\psi}}_t - \boldsymbol{X}^\top\widehat{\boldsymbol{\psi}}). \tag{D.1}$$

By the fact that $\psi$ is non-decreasing and 1-Lipschitz, for any two real numbers $a < b$, $0 \leq \psi(b) - \psi(a) \leq b - a$. Multiplying $\psi(b) - \psi(a)$, we have $(\psi(b) - \psi(a))^2 \leq (\psi(b) - \psi(a))(b - a)$. Thus

$$\|\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}\|^2 \leq (\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}})^\top(\widehat{\boldsymbol{r}}_t - \widehat{\boldsymbol{r}}).$$

Adding up the above two displays we have

$$n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})\|^2 + \|\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}\|^2 \leq (\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})^\top(\boldsymbol{X}_t^\top\widehat{\boldsymbol{\psi}}_t - \boldsymbol{X}^\top\widehat{\boldsymbol{\psi}}) + (\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}})^\top(\widehat{\boldsymbol{r}}_t - \widehat{\boldsymbol{r}}). \tag{D.2}$$

By $\boldsymbol{X}_t^\top\widehat{\boldsymbol{\psi}}_t - \boldsymbol{X}^\top\widehat{\boldsymbol{\psi}} = (\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\psi}} + \boldsymbol{X}_t^\top(\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}})$ and $\boldsymbol{X}_t(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{r}}_t - \widehat{\boldsymbol{r}} = \boldsymbol{y}_t - \boldsymbol{y} - (\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\beta}}$, we have

$$n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})\|^2 + \|\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}\|^2 \leq (\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})^\top(\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\psi}} + (\boldsymbol{y}_t - \boldsymbol{y} - (\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}).$$

By the Cauchy-Schwartz inequality, the above implies

$$\big(n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})\|^2 + \|\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}\|^2\big)^{1/2} \leq (n\mu)^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\psi}}\|_2 + \|\boldsymbol{y}_t - \boldsymbol{y} - (\boldsymbol{X}_t - \boldsymbol{X})^\top\widehat{\boldsymbol{\beta}}\|_2,$$

Since $t, \boldsymbol{U}, \boldsymbol{v}$ are arbitrary, for $(\boldsymbol{y}_t, \boldsymbol{X}_t)$ and $(\boldsymbol{y}, \boldsymbol{X})$ both in a compact subset $K$ of $\mathbb{R}^p \times \mathbb{R}^{n\times p}$, the above display also implies

$$\big(n\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}})\|^2 + \|\widehat{\boldsymbol{\psi}}_t - \widehat{\boldsymbol{\psi}}\|^2\big)^{1/2} \leq \mathrm{const}(K)\big(\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_t - \boldsymbol{X})\|_{op} + \|\boldsymbol{y}_t - \boldsymbol{y}\|_2\big),$$

where $\mathrm{const}(K) := \sup_{(\boldsymbol{y},\boldsymbol{X})\in K}\{(n\mu)^{-1/2}\|\widehat{\boldsymbol{\psi}}\|_2+1+\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\beta}}\|_2\}$. This says that $\widehat{\boldsymbol{\beta}}(\boldsymbol{y},\boldsymbol{X}), \widehat{\boldsymbol{\psi}}(\boldsymbol{y},\boldsymbol{X})$ are locally Lipschitz in $(\boldsymbol{y},\boldsymbol{X})$. By Rademacher's Theorem, $\partial\widehat{\boldsymbol{\beta}}/\partial y_i$ and $\partial\widehat{\boldsymbol{\beta}}/\partial x_{ij}$ exist almost everywhere.

Taking the limit $t \to 0^+$ in (D.1) and using the chain rule, where the derivatives exist we have

$$\begin{aligned}
&n\mu\big\|\boldsymbol{\Sigma}^{1/2}\big(\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\big)\big\|_2^2\\
&\le \Big(\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\Big)^\top\Big(\boldsymbol{U}^\top\widehat{\boldsymbol{\psi}}+\boldsymbol{X}^\top\mathrm{diag}(\widehat{\boldsymbol{\psi}}')\big(-\boldsymbol{U}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})+\Big(I_n-\boldsymbol{X}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\Big)\boldsymbol{v}\big)\Big)\\
&= \Big(\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\Big)^\top B(\boldsymbol{U},\boldsymbol{v})-\Big\|\mathrm{diag}(\widehat{\boldsymbol{\psi}}')^{\frac12}\boldsymbol{X}\Big(\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\Big)\Big\|_2^2
\end{aligned} \tag{D.3}$$

where $(\partial\widehat{\boldsymbol{\beta}}/\partial\boldsymbol{y})\boldsymbol{v} := \sum_{i\in[n]}(\partial\widehat{\boldsymbol{\beta}}/\partial y_i)v_i$, the Jacobian with respect to $\boldsymbol{X}$ and the linear map $B : \mathbb{R}^{n\times p}\times\mathbb{R}^n \to \mathbb{R}^p$ are defined as

$$\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U}) := \sum_{i,j\in[n]\times[p]}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial x_{ij}}u_{ij}\in\mathbb{R}^p,\quad B(\boldsymbol{U},\boldsymbol{v}) := \boldsymbol{U}^\top\widehat{\boldsymbol{\psi}}+\boldsymbol{X}^\top\mathrm{diag}(\widehat{\boldsymbol{\psi}}')(-\boldsymbol{U}\widehat{\boldsymbol{\beta}}+\boldsymbol{v})\in\mathbb{R}^p$$

where $(u_{ij})_{i=1,\dots,n,j=1,\dots,p}$ are the entries of $\boldsymbol{U}$. By the Cauchy-Schwartz inequality, (D.3) provides us the following two main ingredients:

$$\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U}) = 0 \text{ for all } (\boldsymbol{U},\boldsymbol{v}) \text{ such that } B(\boldsymbol{U},\boldsymbol{v}) = 0, \tag{D.4}$$

$$\Big\|\boldsymbol{\Sigma}^{1/2}\Big(\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\Big)\Big\|_2 \le \mu^{-1}n^{-1}\|\boldsymbol{\Sigma}^{-1/2}B(\boldsymbol{U},\boldsymbol{v})\|_2. \tag{D.5}$$

Since both $\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U})$ and $B(\boldsymbol{U},\boldsymbol{v})$ are linear in $(\boldsymbol{U},\boldsymbol{v})\in\mathbb{R}^{n\times p}\times\mathbb{R}^n$ into $\mathbb{R}^p$, Proposition 19 implies that there exists a matrix $\widehat{\boldsymbol{A}}\in\mathbb{R}^{p\times p}$ such that $\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U}) = \widehat{\boldsymbol{A}}B(\boldsymbol{U},\boldsymbol{v})$ for all $(\boldsymbol{U},\boldsymbol{v})$, and by (D.5), $\widehat{\boldsymbol{A}}$ can be chosen such that $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le (n\mu)^{-1}$ thanks to the operator norm identity in Proposition 19. With $(\boldsymbol{U},\boldsymbol{v}) = (\boldsymbol{e}_i\boldsymbol{e}_j^\top,\boldsymbol{0})$ for $(i,j)\in[n]\times[p]$ and $(\boldsymbol{U},\boldsymbol{v}) = (\boldsymbol{0},\boldsymbol{e}_k)$ for $k\in[n]$, we obtain the stated formulae for $(\partial/\partial x_{ij})\widehat{\boldsymbol{\beta}}$ and $(\partial/\partial y_k)\widehat{\boldsymbol{\beta}}$ in (2.1).

Now we show that both $\mathrm{tr}[\boldsymbol{V}] := \mathrm{tr}[\boldsymbol{D}-\boldsymbol{D}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}]$ and $\hat{\mathsf{df}} := \mathrm{tr}[\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}]$ are in $[0,n]$ where $\boldsymbol{D} := \mathrm{diag}\{\psi'(\boldsymbol{r})\}$. Using the symmetric part of $\boldsymbol{A}$ defined as $\widetilde{\boldsymbol{A}} := (\widehat{\boldsymbol{A}}+\widehat{\boldsymbol{A}}^\top)/2$ we have $\mathrm{tr}[\boldsymbol{V}] = \mathrm{tr}[\boldsymbol{D}-\boldsymbol{D}\boldsymbol{X}\widetilde{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}]$ and $\hat{\mathsf{df}} = \mathrm{tr}[\boldsymbol{D}^{1/2}\boldsymbol{X}\widetilde{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2}]$ by property of the trace. In (D.3), take $\boldsymbol{U} = \boldsymbol{0}$ so that $\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}}\boldsymbol{v}+\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U}) = \widehat{\boldsymbol{A}}B(\boldsymbol{U},\boldsymbol{v}) = \widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v}$ and we have with $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$

$$\begin{aligned}
(1+\tfrac{n\mu}{\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2})\|\boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v}\|^2 &\le n\mu\|\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v}\|^2+\|\boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v}\|^2 &\text{(D.6)}\\
&\le \boldsymbol{v}^\top\boldsymbol{D}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{D}\boldsymbol{X}\widetilde{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{v} &\text{(D.7)}
\end{aligned}$$

for all $\boldsymbol{v}$. This implies the positive semi-definite property of the symmetric matrix $\boldsymbol{D}\boldsymbol{X}\widetilde{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}$, and thus $\hat{\mathsf{df}}\ge 0$ and $\mathrm{tr}[\boldsymbol{V}]\le\mathrm{tr}[\boldsymbol{D}]\le n$. With $\widetilde{\boldsymbol{v}} = \boldsymbol{D}^{1/2}\boldsymbol{v}$, it also implies

$$(1+n\mu/\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2)\|\boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2}\widetilde{\boldsymbol{v}}\|^2 \le \widetilde{\boldsymbol{v}}^\top\boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2}\widetilde{\boldsymbol{v}},$$

which, by the Cauchy-Schwartz inequality, yields $(1+n\mu/\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2)\|\boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2}\|_{op} \leq 1$. The same operator norm inequality with $\widehat{\boldsymbol{A}}$ replaced by its symmetric part $\widetilde{\boldsymbol{A}} = (\widehat{\boldsymbol{A}} + \widehat{\boldsymbol{A}}^\top)/2$ thanks to the triangle inequality. Thus $\widehat{\mathrm{df}} \leq \mathrm{tr}[\boldsymbol{D}](1+n\mu/\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2)^{-1} \leq n$ as well as

$$\begin{aligned}\mathrm{tr}[\boldsymbol{V}] = \mathrm{tr}[\boldsymbol{D}^{1/2}(\boldsymbol{I}_n - \boldsymbol{D}^{1/2}\boldsymbol{X}\widetilde{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2})\boldsymbol{D}^{1/2}] &\geq \mathrm{tr}[\boldsymbol{D}](1-(1+n\mu/\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2)^{-1}) \\ &= \mathrm{tr}[\boldsymbol{D}]/(\|\boldsymbol{D}^{1/2}\boldsymbol{G}\|_{op}^2/(n\mu)+1) \\ &\geq \mathrm{tr}[\boldsymbol{D}]/(\|\boldsymbol{G}\|_{op}^2/(n\mu)+1) \qquad \text{(D.8)} \\ &\geq 0\end{aligned}$$

thanks to $\psi' \in [0,1]$. Inequality (D.7) with $\widetilde{\boldsymbol{v}} = \boldsymbol{D}^{1/2}\boldsymbol{v}$ and $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{D}^{1/2}\boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^\top\boldsymbol{D}^{1/2}$ implies $\|(\boldsymbol{M}-\boldsymbol{I}_n)\widetilde{\boldsymbol{v}}\|^2 \leq \widetilde{\boldsymbol{v}}^\top(\boldsymbol{I}_n-\boldsymbol{M})\widetilde{\boldsymbol{v}}$. As the left-hand side is $\|\boldsymbol{M}\widetilde{\boldsymbol{v}}\|^2 - 2\widetilde{\boldsymbol{v}}^\top\boldsymbol{M}\widetilde{\boldsymbol{v}} + \|\widetilde{\boldsymbol{v}}\|^2$, this yields $\|\boldsymbol{M}\widetilde{\boldsymbol{v}}\|^2 \leq \widetilde{\boldsymbol{v}}^\top\boldsymbol{M}\widetilde{\boldsymbol{v}} \leq \|\widetilde{\boldsymbol{v}}\|\|\boldsymbol{M}\widetilde{\boldsymbol{v}}\|$. If $\widetilde{\boldsymbol{v}}$ has unit norm and is such that $\|\boldsymbol{M}\widetilde{\boldsymbol{v}}\| = \|\boldsymbol{M}\|_{op}$ this gives $\|\boldsymbol{M}\|_{op} \leq 1$ so that $\|\boldsymbol{V}\|_{op} = \|\boldsymbol{D}^{1/2}\boldsymbol{M}\boldsymbol{D}^{1/2}\|_{op} \leq \|\boldsymbol{D}\|_{op} \leq 1$. This gives another proof of $\mathrm{tr}[\boldsymbol{V}] \leq n$. ■

**Proof (Proof of Theorem 3)**

The proof for the intercept term included is the same to that of Theorem 1. The only difference is that when computing the derivatives,

$$\begin{aligned}&\frac{d\widehat{\boldsymbol{\psi}}_t}{dt}|_{t=0} = \boldsymbol{U}^\top\widehat{\boldsymbol{\psi}} + \boldsymbol{X}^\top\Big(\frac{\partial\widehat{\boldsymbol{\psi}}}{\partial\boldsymbol{y}}\boldsymbol{v} + \frac{\partial\widehat{\boldsymbol{\psi}}}{\partial\boldsymbol{X}}(\boldsymbol{U})\Big), \quad \frac{\partial\widehat{\boldsymbol{\psi}}}{\partial\boldsymbol{y}}\boldsymbol{v} = \mathrm{diag}(\widehat{\boldsymbol{\psi}}')(\boldsymbol{I}_n - \mathbf{1}\frac{\partial\widehat{\beta}_0}{\partial\boldsymbol{y}} - \boldsymbol{X}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{y}})\boldsymbol{v}, \\ &\frac{\partial\widehat{\boldsymbol{\psi}}}{\partial\boldsymbol{X}}(\boldsymbol{U}) = \mathrm{diag}(\widehat{\boldsymbol{\psi}}')(-\mathbf{1}\frac{\partial\widehat{\beta}_0}{\partial\boldsymbol{X}}(\boldsymbol{U}) - \boldsymbol{U}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{X}}(\boldsymbol{U}))\end{aligned}$$

$$\implies \frac{d\widehat{\boldsymbol{\psi}}_t}{dt}|_{t=0} = -\widehat{\boldsymbol{\psi}}'\frac{d\widehat{\beta}_{0,t}}{dt}|_{t=0} - \mathrm{diag}(\widehat{\boldsymbol{\psi}}')\boldsymbol{X}\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0} + \mathrm{diag}(\widehat{\boldsymbol{\psi}}')\boldsymbol{v} - \mathrm{diag}(\widehat{\boldsymbol{\psi}}')\boldsymbol{U}\widehat{\boldsymbol{\beta}}.$$

We have an additional KKT conditions providing us $0 = \mathbf{1}^\top(d\widehat{\boldsymbol{\psi}}_t/dt)|_{t=0}$. Multiplying $\mathbf{1}^\top$ on both sides of the above display, we have

$$\begin{aligned}\frac{d\widehat{\beta}_{0,t}}{dt}|_{t=0} &= -\frac{\widehat{\boldsymbol{\psi}}'^\top\boldsymbol{X}}{\mathbf{1}^\top\widehat{\boldsymbol{\psi}}'}\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0} + \frac{\widehat{\boldsymbol{\psi}}'^\top\boldsymbol{v}}{\mathbf{1}^\top\widehat{\boldsymbol{\psi}}'} - \frac{\widehat{\boldsymbol{\psi}}'^\top\boldsymbol{U}\widehat{\boldsymbol{\beta}}}{\mathbf{1}^\top\widehat{\boldsymbol{\psi}}'}, \\ \implies \frac{d\widehat{\boldsymbol{\psi}}_t}{dt}|_{t=0} &= -\boldsymbol{\Psi}'\boldsymbol{X}\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0} + \boldsymbol{\Psi}'\boldsymbol{v} - \boldsymbol{\Psi}'\boldsymbol{U}\widehat{\boldsymbol{\beta}},\end{aligned}$$

where $\boldsymbol{\Psi}' := \mathrm{diag}(\widehat{\boldsymbol{\psi}}') - \widehat{\boldsymbol{\psi}}'\widehat{\boldsymbol{\psi}}'^\top/\mathbf{1}^\top\widehat{\boldsymbol{\psi}}'$. By taking limit of $t \to 0$ in Equation (D.2),

$$\begin{aligned}n\mu\Big\|\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}\Big\|_2^2 \leq \frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}^\top\frac{d(\boldsymbol{X}^\top\widehat{\boldsymbol{\psi}})}{dt}|_{t=0} &= \frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}^\top\Big(\boldsymbol{U}^\top\widehat{\boldsymbol{\psi}} + \boldsymbol{X}^\top\frac{d\widehat{\boldsymbol{\psi}}_t}{dt}|_{t=0}\Big) \\ &= \frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}^\top\Big(\boldsymbol{U}^\top\widehat{\boldsymbol{\psi}} + \boldsymbol{X}^\top\big(-\boldsymbol{\Psi}'\boldsymbol{X}\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0} + \boldsymbol{\Psi}'\boldsymbol{v} - \boldsymbol{\Psi}'\boldsymbol{U}\widehat{\boldsymbol{\beta}}\big)\Big) \\ &= \frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}^\top\Big(\boldsymbol{U}^\top\widehat{\boldsymbol{\psi}} + \boldsymbol{X}^\top\boldsymbol{\Psi}'\boldsymbol{v} - \boldsymbol{X}^\top\boldsymbol{\Psi}'\boldsymbol{U}\widehat{\boldsymbol{\beta}}\Big) - \Big\|\boldsymbol{\Psi}'^{1/2}\boldsymbol{X}\frac{d\widehat{\boldsymbol{\beta}}_t}{dt}|_{t=0}\Big\|^2.\end{aligned}$$

■

**Proposition 19** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two real matrices with shape $n$ by $p$. Assume that $\boldsymbol{Bv} = \boldsymbol{0}$ for all $\boldsymbol{v} \in \mathbb{R}^p$ such that $\boldsymbol{Av} = \boldsymbol{0}$. Then the matrix $\boldsymbol{C} := \boldsymbol{BA}^+$ where $\boldsymbol{A}^+$ is the Moore-Penrose pseudoinverse of $\boldsymbol{A}$ satisfies $\boldsymbol{B} = \boldsymbol{CA}$ and $\|\boldsymbol{C}\|_{op} = \max_{\boldsymbol{u}\in\mathbb{R}^n:\boldsymbol{Au}\neq\boldsymbol{0}}\{\|\boldsymbol{Bu}\|_2/\|\boldsymbol{Au}\|_2\}$.*

**Proof** Let $r$ be the rank of $\boldsymbol{A}$. We let $\boldsymbol{A} = \boldsymbol{UDV}^\top$ be the SVD of $\boldsymbol{A}$ with $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ diagonal with positive entries, $\boldsymbol{V} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{U}, \boldsymbol{V}$ both with orthonormal columns. Then $\boldsymbol{A}^+ = \boldsymbol{VD}^{-1}\boldsymbol{U}^\top$ and $\boldsymbol{CA} = \boldsymbol{BVV}^\top = \boldsymbol{B} - \boldsymbol{B}(\boldsymbol{I}_p - \boldsymbol{VV}^\top)$. Since $\ker \boldsymbol{A} \subset \ker \boldsymbol{B}$ and $(\boldsymbol{I}_p - \boldsymbol{VV}^\top)$ is the orthogonal projection onto $\ker \boldsymbol{A}$ we have $\boldsymbol{B}(\boldsymbol{I}_p - \boldsymbol{VV}^\top) = \boldsymbol{0}$. This proves $\boldsymbol{B} = \boldsymbol{CA}$.

For $\|\boldsymbol{BA}^+\|_{op}$, for any vector $\boldsymbol{u}$, $\|\boldsymbol{Bu}\| = \|\boldsymbol{CAu}\| \le \|\boldsymbol{C}\|_{op}\|\boldsymbol{Au}\|$ by definition of $\|\boldsymbol{C}\|_{op}$. This proves $\|\boldsymbol{C}\|_{op} \ge M$ for $M = \max_{\boldsymbol{u}\in\mathbb{R}^n:\boldsymbol{Au}\neq\boldsymbol{0}}\{\|\boldsymbol{Bu}\|_2/\|\boldsymbol{Au}\|_2\}$. For the inequality $M \ge \|\boldsymbol{C}\|_{op}$, if $\|\boldsymbol{C}\|_{op} = \|\boldsymbol{Cv}\|$ for some unit vector $\boldsymbol{v}$ then $\boldsymbol{AA}^+\boldsymbol{v} = \boldsymbol{UU}^\top\boldsymbol{v} \neq \boldsymbol{0}$ since $\boldsymbol{v}$ is a right singular vector of $\boldsymbol{C} = \boldsymbol{BVDU}^\top$ and cannot belong to $\ker(\boldsymbol{U}^\top)$. Next, $\|\boldsymbol{Cv}\| = \|\boldsymbol{BA}^+\boldsymbol{v}\| \le M\|\boldsymbol{AA}^+\boldsymbol{v}\|$ and the conclusion follows since $\|\boldsymbol{AA}^+\|_{op} \le 1$ and $\|\boldsymbol{v}\| = 1$. ■

## Appendix E. Relaxing strong convexity: Proof of Proposition 12

Consider the notation for $\boldsymbol{G}, \boldsymbol{\psi}(\varepsilon, \boldsymbol{G}), \boldsymbol{h}(\varepsilon, \boldsymbol{G}), D$ defined around (B.1). Let $\Omega = \{\boldsymbol{X} \in U_\varepsilon\}$. Let us first rewrite the Lipschitz condition (7.1) using the change of variable $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ explained around equation (B.1): the mapping

$$\tilde{\Phi}_\varepsilon \begin{cases} \{\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}, \boldsymbol{M} \in U_\varepsilon\} \to \mathbb{R}^{n+p}, \\ \qquad\qquad \boldsymbol{G} \mapsto D^{-1}\big(n^{-1/2}\boldsymbol{\psi}(\varepsilon, \boldsymbol{G}),\ \boldsymbol{h}(\varepsilon, \boldsymbol{G})\big) \end{cases} \text{is } \frac{L}{\sqrt{n}}\text{-Lipschitz}$$

where we recall the notation $D = (\frac{1}{n}\|\boldsymbol{\psi}(\varepsilon, \boldsymbol{G})\|^2 + \|\boldsymbol{h}(\varepsilon, \boldsymbol{G})\|^2)^{1/2}$. After the change of variable, the identities (B.2)-(B.8) all hold in the event $\Omega$.

All previous calculations made in the strongly convex case hold here is well, but only in the event $\Omega$. Outside of event $\Omega$, the derivatives may not exist at all. As in Theorem 6, the device that lets us work around this is Kirszbraun's theorem: there exists an $L/\sqrt{n}$ Lipschitz function $\bar{\Phi}_\varepsilon$ (the "extension") such that $\bar{\Phi}_\varepsilon(\boldsymbol{G}) = \tilde{\Phi}_\varepsilon(\boldsymbol{G})$ for all $\boldsymbol{G}$ in the domain $\{\boldsymbol{M}\boldsymbol{\Sigma}^{-1/2}, \boldsymbol{M} \in U_\varepsilon\}$ of $\tilde{\Phi}_\varepsilon$.

Now define $\boldsymbol{\rho} : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ and $\boldsymbol{\eta} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ by $\bar{\Phi}(\boldsymbol{G}) = (\boldsymbol{\rho}(\boldsymbol{G}), \boldsymbol{\eta}(\boldsymbol{G}))$. Using the Lipschitz condition and the fact that the Frobenius norm of a Jacobian is bounded from above by it rank times the square of its operator norm,

$$\sum_{ij} \|\frac{\partial \boldsymbol{\rho}}{\partial g_{ij}}\|^2 \le L^2, \qquad \sum_{ij} \|\frac{\partial \boldsymbol{\eta}}{\partial g_{ij}}\|^2 \le \frac{p}{n}L^2$$

so that the right-hand side of (C.6) is bounded above by $C_{54}(L, \gamma)\sqrt{p}$ and the right-hand side of (C.3) is bounded from above by $C_{55}(L, \gamma)$. After we have bounded the right-hand side, we are allowed to add the indicator function $I\{\Omega\}$ in the left-hand sides of (C.6) and of (C.3) since adding an indicator function only makes it smaller. This devices lets us obtain analogs of Proposition 14 and Proposition 16 where the right-hand sides are of the same order as in the strongly convex case, provided that we add the indicator function $I\{\Omega\}$ in the left-hand sides.

From here, in the event $\Omega$ we use the bound $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le L/n$ assumed in Proposition 12 instead of the bound $\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{1/2}\|_{op} \le 1/(n\mu)$ from Theorem 1. This device provides the bounds

(B.14), (B.15), (B.16), (B.18), (B.19) and (B.20) on $\xi_I, ...\xi_{VI}$, with the modification that the left-hand sides present the indicator function $I\{\Omega\}$ and the right-hand sides are $C_{56}(\gamma, L)$ instead of $C_{57}(\gamma, \mu)$ in the strongly convex case. The algebra is then the same as in the strongly convex case and Proposition 12 follows.

## Appendix F. Additional Figures (anisotropic Gaussian design)

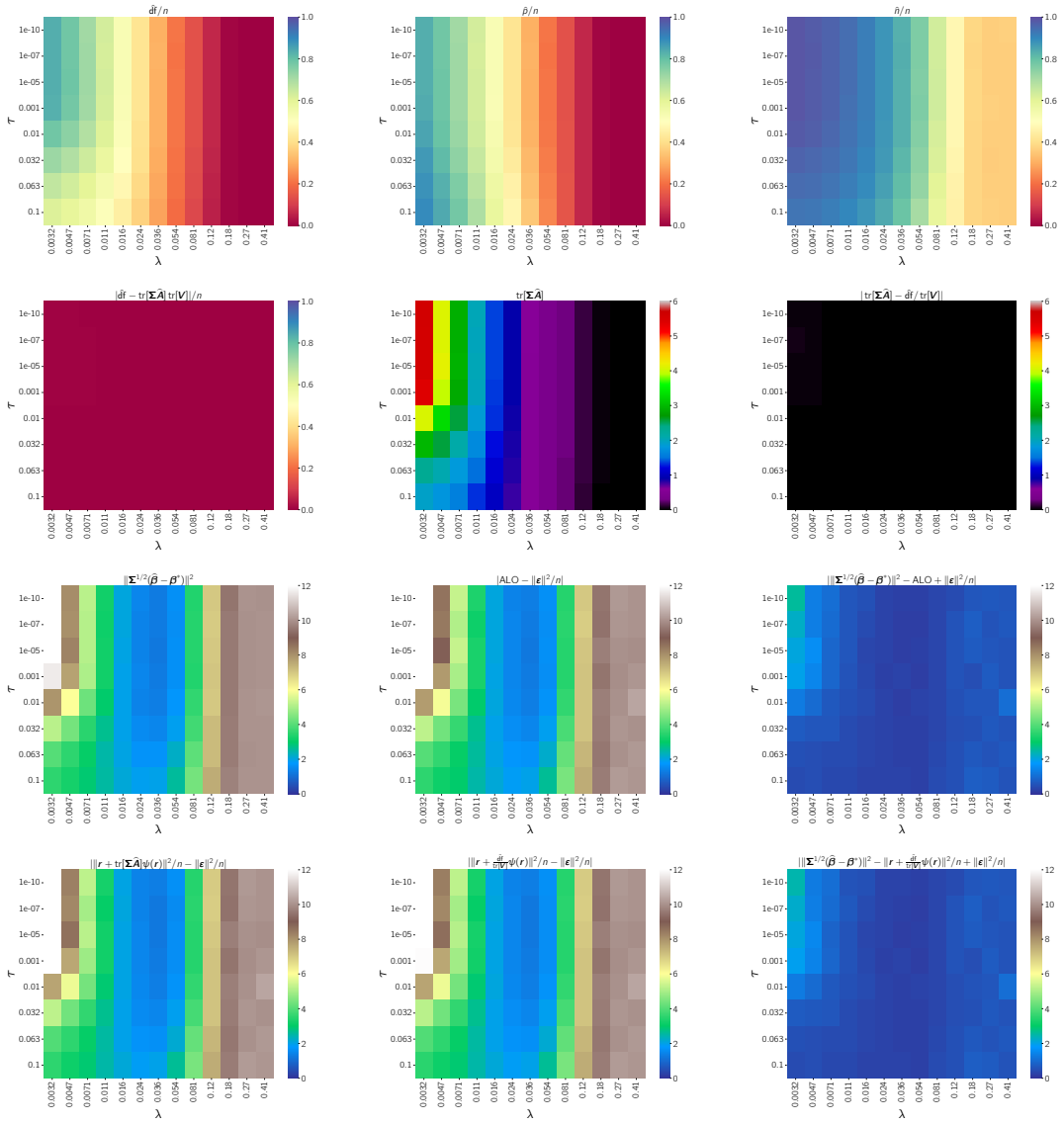

Figure 4: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.054n^{1/2}$ and $(\lambda, \tau)$ where $\lambda \in [0.0032, 0.41]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is the average over 100 repetitions. See the simulation setup in Section 6 in the paper for more details.

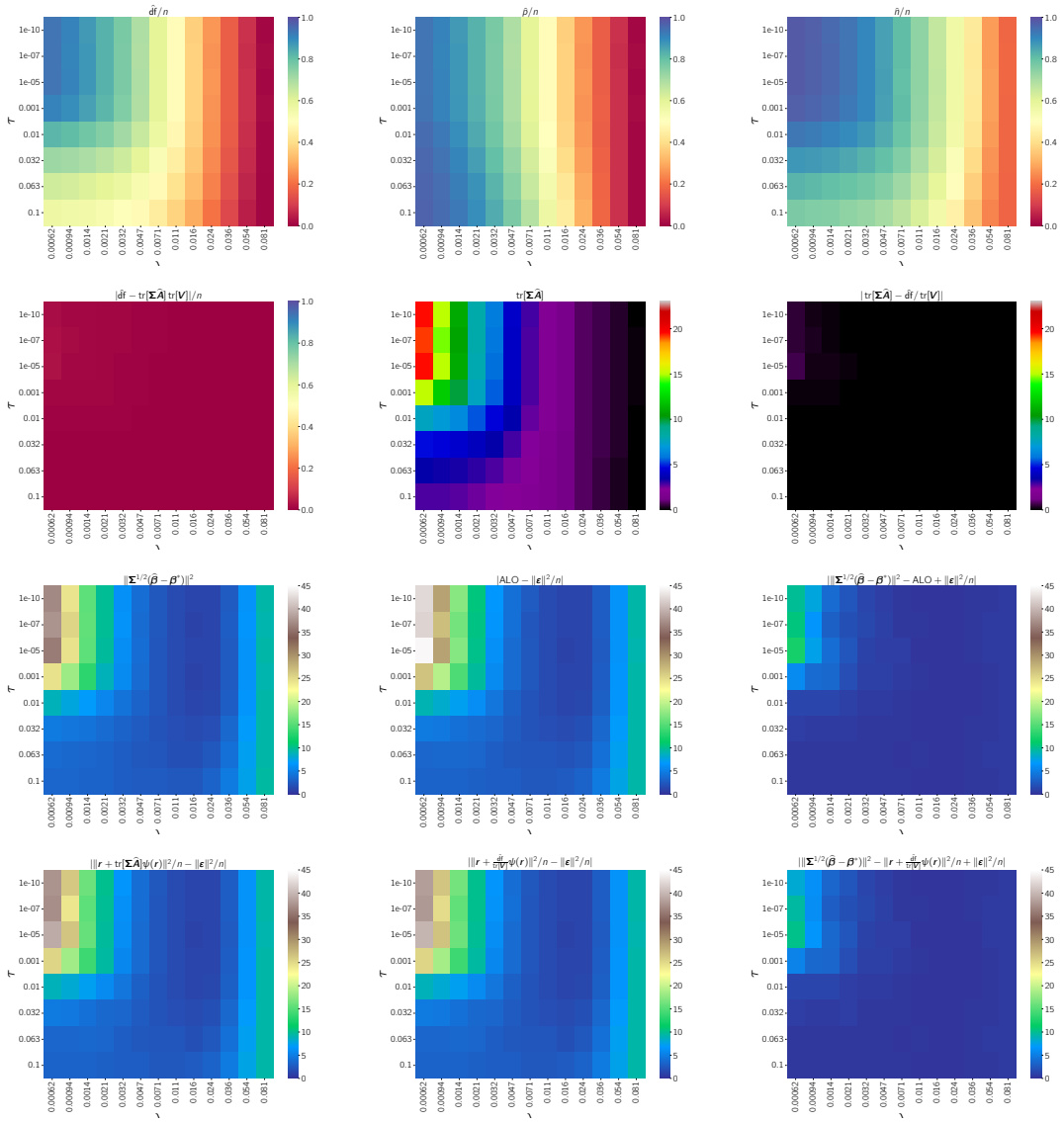

Figure 5: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.024n^{1/2}$ and $(\lambda, \tau)$ where $\lambda \in [0.00062, 0.081]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is the average over 50 repetitions. See the simulation setup in Section 6 in the paper for more details.

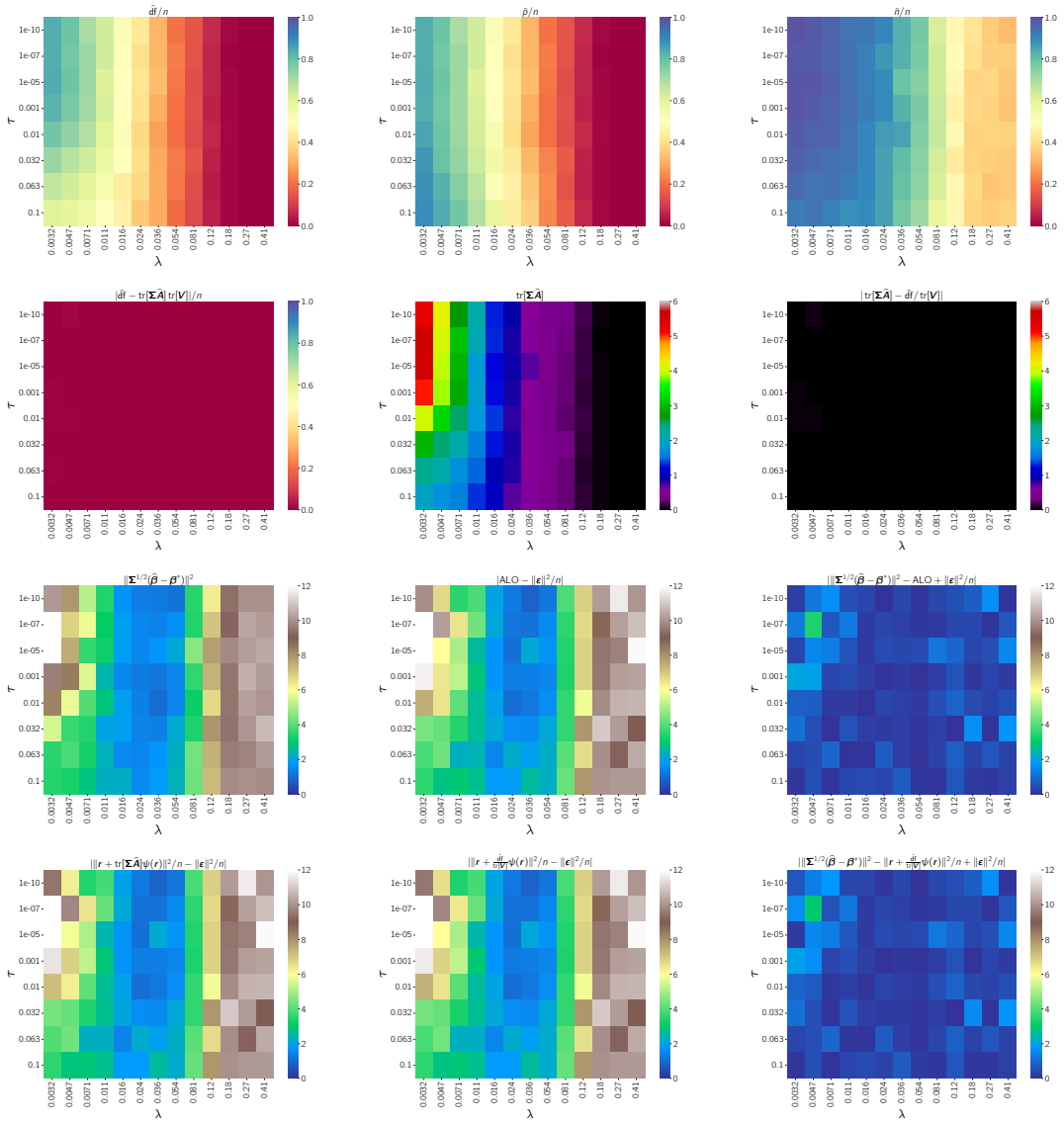

Figure 6: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.054n^{1/2}$ and $(\lambda, \tau)$ where $\lambda \in [0.0032, 0.41]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is over 1 repetition. See the simulation setup in Section 6 in the paper for more details.

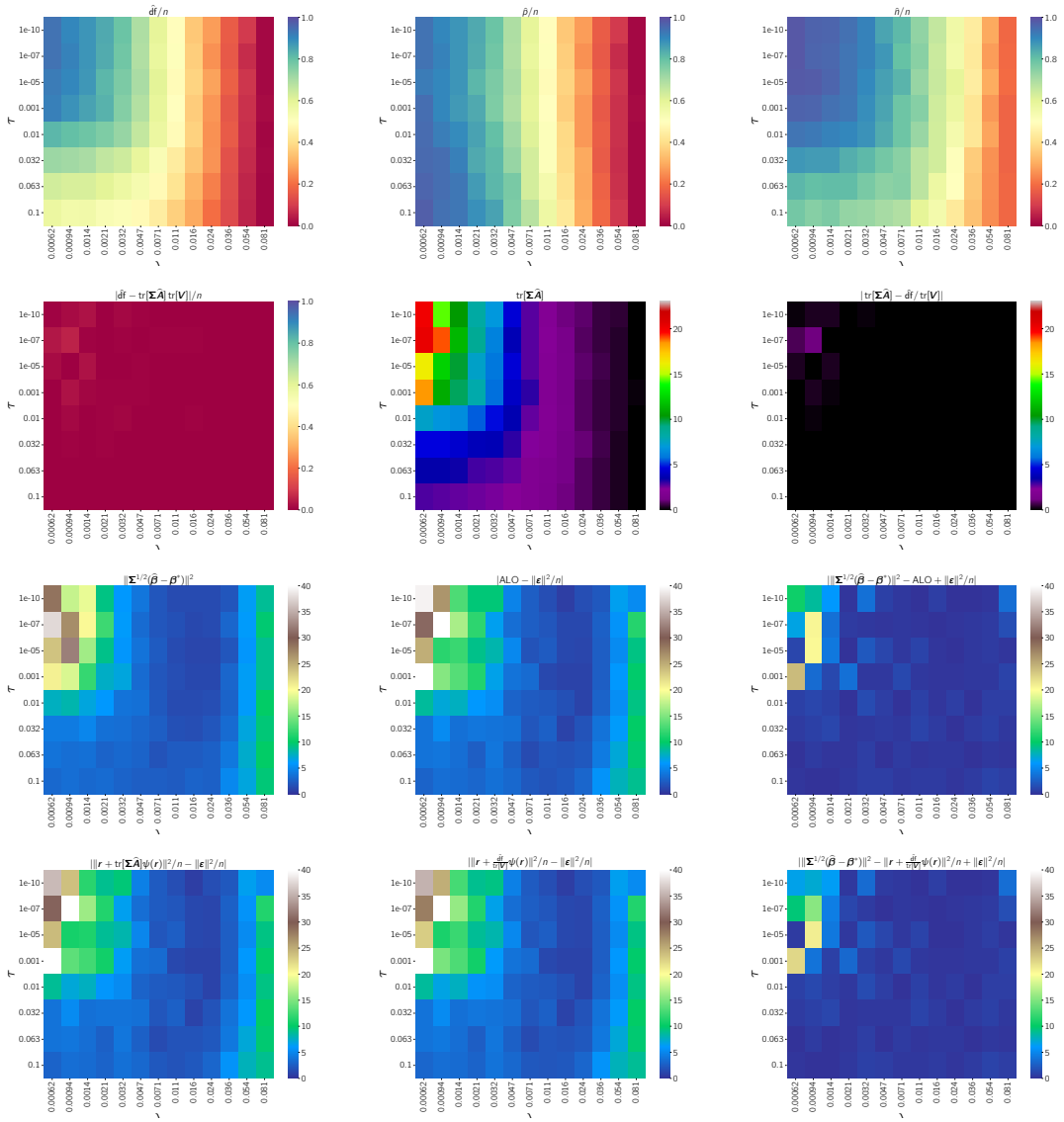

Figure 7: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.024n^{1/2}$ and $(\lambda, \tau)$ where $\lambda \in [0.00062, 0.081]$ and $\tau \in [10^{-10}, 0.1]$. Each cell over 1 repetition. See the simulation setup in Section 6 in the paper for more details.

## Appendix G. Additional Figures (non-Gaussian, Rademacher design)

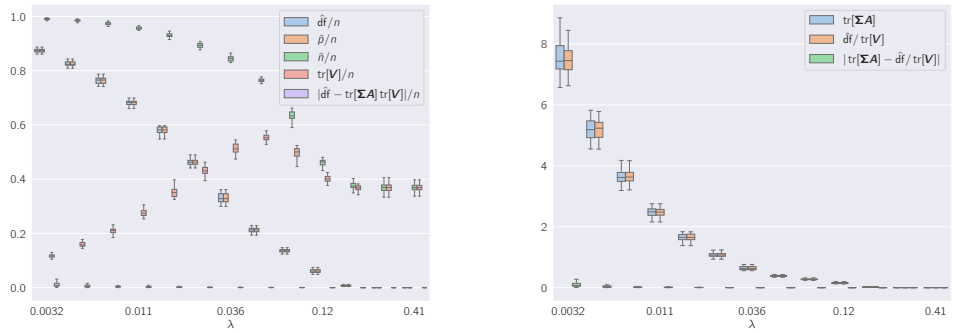


Figure 8: Boxplots for $\hat{\mathrm{df}}, \hat{p}, \hat{n}, \mathrm{tr}[\boldsymbol{V}], \mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}]$ and $|\,\mathrm{tr}[\boldsymbol{\Sigma}\widehat{\boldsymbol{A}}] - \hat{\mathrm{df}}/\,\mathrm{tr}[\boldsymbol{V}]|$ in Huber Elastic-Net regression with $\tau = 10^{-10}$ and $\lambda \in [0.0032, 0.41]$. The data are generated with $\boldsymbol{X}$ having iid entries taking value $\pm 1$ each with probability 0.5 (so that $\boldsymbol{\Sigma} = \boldsymbol{I}_p$). Each box contains 30 data points.

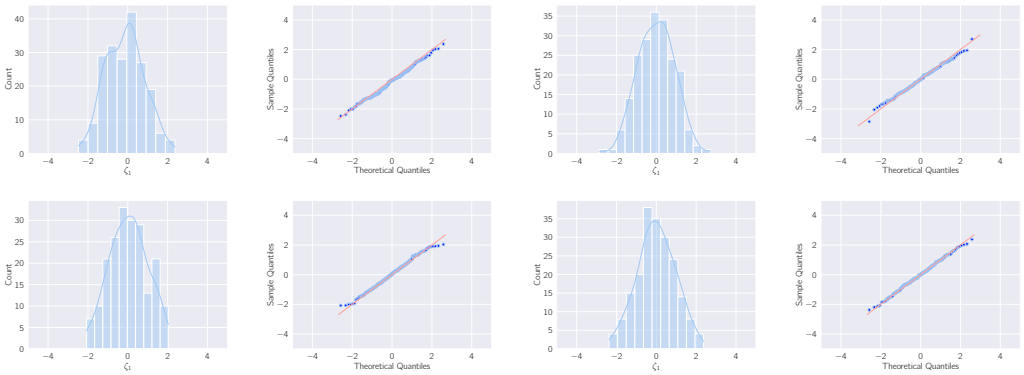


Figure 9: Histogram and QQ-plot for $\zeta_1$ in (3.3) under Huber Elastic-Net regression for different choices of tuning parameters $(\lambda, \tau)$. Left Top: $(0.036, 10^{-10})$, Right Top: $(0.054, 0.01)$, Left Bottom: $(0.036, 0.01)$, Right Bottom: $(0.024, 0.1)$. Each figure contains 100 data points generated with Rademacher design matrix (each entry has value $\pm 1$ with probability 0.5) and iid $\varepsilon_i$ from the $t$-distribution with 2 degrees of freedom.