

---

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

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

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

## 21 1 Introduction

22 This paper studies properties of robust estimators in linear models  $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$  with response  
23  $\mathbf{y} \in \mathbb{R}^n$ , unknown regression vector  $\beta^*$  where  $\mathbf{X}$  is a design matrix with  $n$  rows  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , each row  
24  $\mathbf{x}_i$  being a high-dimensional feature vector in  $\mathbb{R}^p$  with covariance  $\Sigma$ . Throughout, let  $\hat{\beta} = \hat{\beta}(\mathbf{y}, \mathbf{X})$   
25 be a regularized M-estimator given as a solution of the convex minimization problem

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

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

32 The paper introduces the following criterion to select a loss-penalty pair  $(\rho, g)$  with small out-of-  
33 sample error  $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2$ : for a given set of candidate loss-penalty pairs  $\{(\rho, g)\}$  and the

34 corresponding  $M$ -estimator  $\hat{\beta}_{\rho,g}$  in (1), select the pair  $(\rho, g)$  that minimizes the criterion

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

35 where  $\text{tr}[\cdot]$  is the trace,  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  is the derivative of  $\rho$ ,  $\psi'$  the derivative of  $\psi$  and we extend  $\psi$   
36 and  $\psi'$  to functions  $\mathbb{R}^n \rightarrow \mathbb{R}^n$  by componentwise application of the univariate function of the same  
37 symbol. Above,  $(\partial/\partial \mathbf{y}) \hat{\beta}_{\rho,g} \in \mathbb{R}^{p \times n}$  denotes the Jacobian of (1) with respect to  $\mathbf{y}$  for  $\mathbf{X}$  fixed,  
38 at the observed data  $(\mathbf{y}, \mathbf{X})$ . As we will see while studying particular examples, for pairs  $(\rho, g)$   
39 commonly used in robust high-dimensional statistics such as the square loss, Huber loss with the  
40  $\ell_1$ -penalty or Elastic-Net penalty, the ratio  $\hat{\text{df}}/\text{tr}[\mathbf{V}]$  in (2) admits simple, closed-form expressions  
41 and can be computed at a negligible computational cost once  $\hat{\beta}_{\rho,g}(\mathbf{y}, \mathbf{X})$  itself has been computed.  
42 The criterion (2) has an appealing adaptivity property: it does not require any knowledge of the noise  
43  $\varepsilon$  or its distribution, nor any knowledge of the covariance  $\Sigma$  of the design.

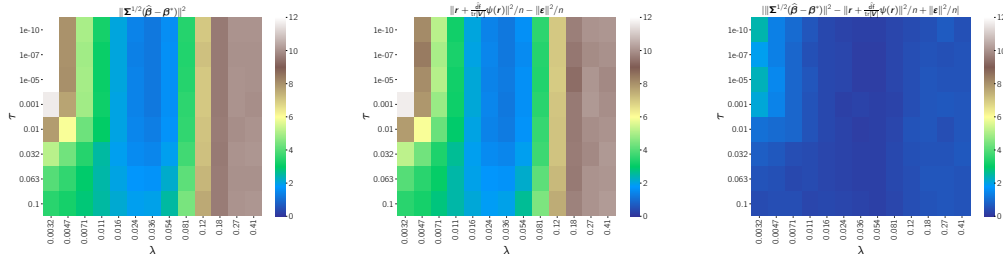


Figure 1: Heatmaps for  $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2$ , its approximation  $\|\mathbf{r} + (\hat{\text{df}}/\text{tr}[\mathbf{V}])\psi(\mathbf{r})\|^2/n - \|\varepsilon\|^2/n$  and the approximation error  $|\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 - \|\mathbf{r} + (\hat{\text{df}}/\text{tr}[\mathbf{V}])\psi(\mathbf{r})\|^2/n - \|\varepsilon\|^2/n|$  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.

## 44 1.1 Contributions

- 45 1. The end goal of paper is to provide theoretical justification and theoretical guarantees for the  
46 criterion (2) in the high-dimensional regime where the ratio  $p/n$  has a finite limit and  $\mathbf{X}$  has  
47 anisotropic Gaussian distribution. The theoretical results will justify the approximation

$$\left\| \mathbf{r} + \left( \hat{\text{df}} / \text{tr}[\mathbf{V}] \right) \psi(\mathbf{r}) \right\|^2 / n \approx \|\varepsilon\|^2 / n + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2. \quad (3)$$

48 Figure 1 illustrates the accuracy of (3) on simulated data. To study the criterion (2) and derive the  
49 approximation (3), we develop novel results of independent interest regarding  $M$ -estimators in (1):

- 50 2. The paper derives general formula for the derivatives  $(\partial/\partial y_i) \hat{\beta}$  and  $(\partial/\partial x_{ij}) \hat{\beta}$ . This sheds light  
51 on the differentiability structure of  $M$ -estimators for general loss-penalty pairs: for any  $\rho, g$  with  $g$   
52 strongly convex, there exists  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  depending on  $(\mathbf{y}, \mathbf{X})$  such that for almost every  $(\mathbf{y}, \mathbf{X})$ ,

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

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

- 54 3. The paper obtains a stochastic representation for the residual  $y_i - \mathbf{x}_i^\top \hat{\beta}$  for some fixed  $i = 1, \dots, n$ ,  
55 extending some results of [12] on unregularized  $M$ -estimators to penalized ones as in (1). In  
56 short, for each  $i = 1, \dots, n$  the  $i$ -th residual satisfies  $r_i = y_i - \mathbf{x}_i^\top \hat{\beta}$

$$r_i + (\hat{\text{df}} / \text{tr} \mathbf{V}) \psi(r_i) \approx \varepsilon_i + Z_i \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\| \quad (4)$$

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

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

63 **Assumption 1.1.** For constants  $\gamma, \mu > 0$  independent of  $n, p$  we have  $p/n \leq \gamma$ , the loss  $\rho : \mathbb{R} \rightarrow \mathbb{R}$   
64 is convex with a unique minimizer at 0, continuously differentiable and its derivative  $\psi = \rho'$  is  
65 1-Lipschitz. The design matrix  $\mathbf{X}$  has iid  $N(\mathbf{0}, \boldsymbol{\Sigma})$  rows for some invertible covariance  $\boldsymbol{\Sigma}$  and the  
66 noise  $\boldsymbol{\varepsilon}$  is independent of  $\mathbf{X}$  with continuous distribution. The penalty  $g : \mathbb{R}^p \rightarrow \mathbb{R}$  is  $\mu$ -strongly  
67 convex w.r.t.  $\boldsymbol{\Sigma}$  in the sense that  $\mathbf{b} \mapsto g(\mathbf{b}) - (\mu/2)\mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}$  is convex in  $\mathbf{b} \in \mathbb{R}^p$ .

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

73 Since Assumption 1.1 requires  $p/n \leq \gamma$ , the Bolzano-Weierstrass theorem lets us extract a subse-  
74 quence of regression problems such that  $p/n \rightarrow \gamma'$  along this subsequence, for some constant  $\gamma$ . This  
75 is the asymptotic regime we have in mind throughout the paper, although our results do not require a  
76 specific limit for the ratio  $p/n$ . For some results, we will require the following additional assumption  
77 which is satisfied by robust loss functions and penalty that shrink towards 0.

78 **Assumption 1.2.** The penalty is minimized at  $\mathbf{0}$ , that is,  $g(\mathbf{0}) = \min_{\mathbf{b} \in \mathbb{R}^p} g(\mathbf{b})$ ; the loss is Lipschitz as  
79 in  $|\psi| \leq M$  for some constant  $M$  independent of  $n, p$ ; the signal is bounded as in  $\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*\|^2 \leq M$ .

## 80 1.2 Related works

81 The context of the present work is the study of  $M$ -estimators in the regime  $\frac{p}{n}$  has a finite limit. This  
82 literature pioneered in [2, 12, 11, 18] typically describes the subtle behavior of  $\hat{\boldsymbol{\beta}}$  in this regime by  
83 solving a system of nonlinear equations. This system typically depends on a prior distribution for the  
84 components of  $\boldsymbol{\beta}^*$ , and either depends on the covariance  $\boldsymbol{\Sigma}$  [8] or assume  $\boldsymbol{\Sigma} = \mathbf{I}_p$  [2, 19, 7, among  
85 many others]. Solutions to the nonlinear system are a powerful tool to understand  $\hat{\boldsymbol{\beta}}$  in theory, e.g.,  
86 to characterize the deterministic limit of  $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|$ , see e.g., the general results in [7] for the  
87 square loss and [19] for general loss-penalty pairs. However, since the system and its solution depend  
88 on unobservable quantity ( $\boldsymbol{\Sigma}$  and prior on  $\boldsymbol{\beta}^*$ ), the system solution is not directly usable for practical  
89 purposes such as parameter tuning.

90 The present work distinguishes itself from most of this literature as the goal is to describe the behavior  
91 of  $\hat{\boldsymbol{\beta}}$  using observable quantities that only depend on the data  $(\mathbf{y}, \mathbf{X})$  (and not unobservable ones such  
92 as  $\boldsymbol{\Sigma}$  or a prior distribution on  $\boldsymbol{\beta}^*$  that appear in the aforementioned nonlinear system of equations).  
93 As we will see this view lets us perform adaptive tuning of parameters in a fully adaptive manner  
94 using the criterion (2). The criterion (2) appeared in previous works for the square loss only: [1, 15]  
95 studied (2) for the Lasso with  $\boldsymbol{\Sigma} = \mathbf{I}_p$  and [3, Section 3] for the square loss and general penalty  
96 (note that for the square loss  $\rho(u) = u^2/2$ , (2) reduces to  $n^2 \|\mathbf{r}\|^2 / (n - \hat{\text{df}})^2$  due to  $\psi(u) = u$  and  
97  $\text{tr}[\mathbf{V}] = n - \hat{\text{df}}$ . The property  $\psi(u) = u$  of the square loss hides the subtle interplay between  
98  $\mathbf{r}, \psi(\mathbf{r}), \hat{\text{df}}$  and  $\text{tr}[\mathbf{V}]$  in (2) for  $\rho$  different than the square loss). A criterion different from (2) is  
99 studied in [15, 3] to estimate the out-of-sample error. That criterion has the drawback to require the  
100 knowledge of  $\boldsymbol{\Sigma}$ , unlike (2) which is fully adaptive.

101 This work leverages probabilistic results on functions of standard normal random variables [5][3,  
102 §6, §7] which are consequences of Stein's formula [17]. Consequently, the main limitation of our  
103 work is that it currently requires Gaussian design for the probabilistic results (on the other hand, the  
104 differentiability result (5) is deterministic and does not rely on any probabilistic assumption).

## 105 2 Differentiability of regularized M-estimators

106 The first step towards the study of the criterion (2) is to justify the almost sure existence of the  
107 derivatives of  $\hat{\boldsymbol{\beta}}$  that appear in (2) through the scalar  $\hat{\text{df}}$  and the matrix  $\mathbf{V}$  in (2). Although the  
108 criterion (2) only involves the derivatives of  $\hat{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{X})$  with respect to  $\mathbf{y}$  for a fixed  $\mathbf{X}$ , the proof of

our results rely on the interplay between the derivatives with respect to  $\mathbf{y}$  and with respect to  $\mathbf{X}$ : this differentiability structure of  $M$ -estimators is the content of the following result.

**Theorem 2.1.** *Let Assumption 1.1 be fulfilled. For almost every  $(\mathbf{y}, \mathbf{X})$  the map  $(\mathbf{y}, \mathbf{X}) \mapsto \hat{\beta}(\mathbf{y}, \mathbf{X})$  is differentiable at  $(\mathbf{y}, \mathbf{X})$  and there exists a matrix  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  with  $\|\Sigma^{1/2} \hat{\mathbf{A}} \Sigma^{1/2}\|_{op} \leq (n\mu)^{-1}$  s.t.*

$$\begin{aligned} (\partial/\partial y_i) \hat{\beta}(\mathbf{y}, \mathbf{X}) &= \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i \psi'(r_i), \\ (\partial/\partial x_{ij}) \hat{\beta}(\mathbf{y}, \mathbf{X}) &= \hat{\mathbf{A}} \mathbf{e}_j \psi(r_i) - \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i \psi'(r_i) \hat{\beta}_j, \end{aligned} \quad \text{where } r_i = y_i - \mathbf{x}_i^\top \hat{\beta}, \quad (5)$$

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

$$\text{df} = \text{tr}[\mathbf{X}(\partial/\partial \mathbf{y}) \hat{\beta}] = \text{tr}[\mathbf{X} \hat{\mathbf{A}} \mathbf{X} \text{diag}\{\psi'(\mathbf{r})\}], \quad (6)$$

$$\mathbf{V} = \text{diag}\{\psi'(\mathbf{r})\}(\mathbf{I}_n - \mathbf{X}(\partial/\partial \mathbf{y}) \hat{\beta}) = \text{diag}\{\psi'(\mathbf{r})\} - \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X} \hat{\mathbf{A}} \mathbf{X} \text{diag}\{\psi'(\mathbf{r})\}. \quad (7)$$

satisfy  $0 \leq \text{df} \leq n$  and  $0 \leq \text{tr}[\mathbf{V}] \leq n$ .

Since the same matrix  $\hat{\mathbf{A}}$  appears in both the derivatives with respect to  $y_i$  and to  $x_{ij}$ , (5) provides relationship between  $(\partial/\partial y_i) \hat{\beta}$  and  $(\partial/\partial x_{ij}) \hat{\beta}$ , for instance  $(\partial/\partial x_{ij}) \hat{\beta} = \hat{\mathbf{A}} \mathbf{e}_j \psi(r_i) - \hat{\beta}_j (\partial/\partial y_i) \hat{\beta}$ . Although the matrix  $\hat{\mathbf{A}}$  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 2.1.** *For the square loss  $\rho(u) = u^2/2$ , the differentiability formulae (5) reduce to*

$$(\partial/\partial y_i) \hat{\beta}(\mathbf{y}, \mathbf{X}) = \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i, \quad (\partial/\partial x_{ij}) \hat{\beta}(\mathbf{y}, \mathbf{X}) = \hat{\mathbf{A}} \mathbf{e}_j (y_i - \mathbf{x}_i^\top \hat{\beta}) - \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i \hat{\beta}_j$$

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

In the simple case where  $g$  is twice continuously differentiable, (5) follows [5] with

$$\hat{\mathbf{A}} = (\mathbf{X}^\top \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X} + n \nabla^2 g(\hat{\beta}))^{-1} \quad (8)$$

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

$$\dot{\mathbf{X}}^\top \psi(\mathbf{r}) - \mathbf{X}^\top \text{diag}\{\psi'(\mathbf{r})\}(\dot{\mathbf{y}} - \dot{\mathbf{X}} \hat{\beta}(\mathbf{y}, \mathbf{X})) = \hat{\mathbf{A}}^{-1} \frac{d}{dt} \hat{\beta}(\mathbf{y}(t), \mathbf{X}(t))|_{t=0}$$

with  $\hat{\mathbf{A}}$  in (8). This gives (5) if the penalty  $g$  is twice-differentiable. Theorem 2.1 reveals that for arbitrary convex penalty functions including non-differentiable ones, the differentiability structure (5) always holds, as in the case of twice differentiable penalty  $g$ , even for penalty functions such as  $g(\mathbf{b}) = \mu \|\mathbf{b}\|^2/2 + \lambda \|\text{mat}(\mathbf{b})\|_{\text{nuc}}$  where  $\text{mat} : \mathbb{R}^p \rightarrow \mathbb{R}^{d_1 \times d_2}$  is a linear isomorphism to the space of  $d_1 \times d_2$  matrices and  $\|\cdot\|_{\text{nuc}}$  is the nuclear norm: in this case by Theorem 2.1 there exists a matrix  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  such that (5) holds although no closed-form expression for  $\hat{\mathbf{A}}$  is known.

The representation (5) is a powerful tool as it provides explicit derivatives of quantities of interest such as  $\mathbf{r} = \mathbf{y} - \mathbf{X} \hat{\beta}$ ,  $\|\psi(\mathbf{r})\|^2$  or  $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2$ . These explicit derivatives can then be used in probabilistic identities and inequalities that involve derivatives, for instance Stein's formulae [17], the Gaussian Poincaré inequality [6, Theorem 3.20], or normal approximations [9, 5].

**Remark 2.2.** *Similar derivative formulae hold if an intercept is included in the minimization, as in*

$$(\hat{\beta}_0(\mathbf{y}, \mathbf{X}), \hat{\beta}(\mathbf{y}, \mathbf{X})) = \underset{b_0 \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n \rho(y_i - b_0 - \mathbf{x}_i^\top \mathbf{b}) + g(\mathbf{b}) \quad (9)$$

Let Assumption 1.1 be fulfilled, and assume further  $\|\psi'(\mathbf{r})\|_2 > 0$  with  $\mathbf{r} := \mathbf{y} - \mathbf{1}_n \hat{\beta}_0 - \mathbf{X}^\top \hat{\beta}$  where  $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$ . For almost every  $(\mathbf{y}, \mathbf{X})$  the map  $(\mathbf{y}, \mathbf{X}) \mapsto \hat{\beta}(\mathbf{y}, \mathbf{X})$  is differentiable at  $(\mathbf{y}, \mathbf{X})$ , and there exists  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  depending on  $(\mathbf{y}, \mathbf{X})$  with  $\|\Sigma^{1/2} \hat{\mathbf{A}} \Sigma^{1/2}\|_{op} \leq (n\mu)^{-1}$  such that

$$(\partial/\partial y_i) \hat{\beta}(\mathbf{y}, \mathbf{X}) = \hat{\mathbf{A}} \mathbf{X}^\top \Psi' \mathbf{e}_i, \quad (\partial/\partial x_{ij}) \hat{\beta}(\mathbf{y}, \mathbf{X}) = \hat{\mathbf{A}} \mathbf{e}_j \psi(r_i) - \hat{\mathbf{A}} \mathbf{X}^\top \Psi' \mathbf{e}_i \hat{\beta}_j, \quad (10)$$

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

### 3 Distribution of individual residuals

We now turn to the distribution of a single residual  $r_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$  for some fixed observation  $i \in \{1, \dots, n\}$  (for instance, fix  $i = 1$ ). By leveraging the differentiability structure (5) and the normal approximation from [5], the following result provides a clear picture of the distribution of  $r_i$ .

**Theorem 3.1.** *Let Assumption 1.1 be fulfilled and let  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  be given by Theorem 2.1. Then for every  $i = 1, \dots, n$  there exists  $Z_i \sim N(0, 1)$  such that*

$$\left| \left( r_i + \text{tr}[\Sigma \hat{\mathbf{A}}] \psi(r_i) \right) - \left( \varepsilon_i + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| Z_i \right) \right| \leq O_P(n^{-1/4})(|\psi(\varepsilon_i)| + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|) \quad (11)$$

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

$$r_i + \text{tr}[\Sigma \hat{\mathbf{A}}] \psi(r_i) = \tilde{\varepsilon}_i^n + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| \tilde{Z}_i^n. \quad (12)$$

If  $\varepsilon_i$  has a fixed distribution  $F$  and Assumption 1.2 holds then  $|\psi(\varepsilon_i)| + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\| = O_P(1)$ .

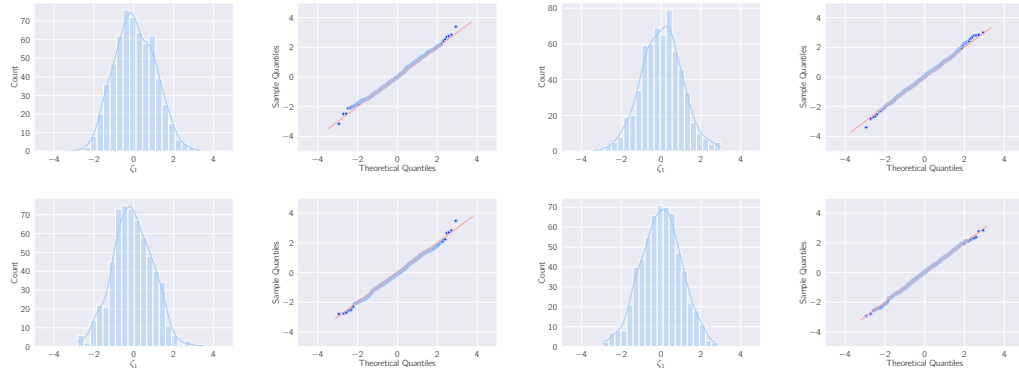


Figure 2: Histogram and QQ-plot for  $\zeta_i$  in (13) under Huber Elastic-Net regression for different choices of tuning parameters  $(\lambda, \tau)$ . Left Top:  $(0.036, 10^{-10})$ , Right Top:  $(0.054, 0.01)$ , Left Bottom:  $(0.036, 0.01)$ , Right Bottom:  $(0.024, 0.1)$ . Each figure contains 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 3.1 is a formal statement regarding the informal normal approximation

$$\zeta_i := \frac{r_i + \text{tr}[\Sigma \hat{\mathbf{A}}] \psi(r_i) - \varepsilon_i}{\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|} \approx N(0, 1). \quad (13)$$

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 3.2.** *Let Assumption 1.1 hold with  $\rho(u) = u^2/2$  and  $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Then for  $i = 1$ ,*

$$(\sigma^2 + \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2)^{-1/2} (1 + \text{tr}[\Sigma \hat{\mathbf{A}}]) (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) \rightarrow^d N(0, 1) \quad \text{as } n \rightarrow +\infty. \quad (14)$$

It is informative to provide a sketch of the proof of Theorem 3.1 explain the appearance of  $\psi(r_i)$  and  $\text{tr}[\Sigma \hat{\mathbf{A}}]$  in the normal approximation results (11) and (13). A variant of the normal approximation of [5] proved in the supplement states that for a differentiable function  $\mathbf{f} : \mathbb{R}^q \rightarrow \mathbb{R}^q \setminus \{\mathbf{0}\}$  and  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_q)$ , there exists  $Z \sim N(0, 1)$  such that

$$\mathbb{E} \left[ \left| \frac{\mathbf{f}(\mathbf{z})^\top \mathbf{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k(\mathbf{z})}{\|\mathbf{f}(\mathbf{z})\|} - Z \right|^2 \right] \leq C_1 \mathbb{E} \left[ \frac{\sum_{k=1}^q \|(\partial/\partial z_k) \mathbf{f}(\mathbf{z})\|^2}{\|\mathbf{f}(\mathbf{z})\|^2} \right]. \quad (15)$$

Some technical hurdles aside, the proof sketch is the following: Apply the previous display to  $q = p$ ,  $\mathbf{z} = \Sigma^{-1/2} \mathbf{x}_i$  conditionally on  $(\varepsilon, (\mathbf{x}_l)_{l \in [n] \setminus \{i\}})$  and to  $\mathbf{f}(\mathbf{z}) = \Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$  in the simple case where  $\boldsymbol{\beta}^* = \mathbf{0}$  (this amounts to performing a change of variable by translation of  $\hat{\boldsymbol{\beta}}$  to  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ ). Then the right-hand side of the previous display is negligible in probability compared to  $Z$ , and in the left-hand side  $\mathbf{f}(\mathbf{z})^\top \mathbf{z} = \mathbf{x}_i^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$  and  $\sum_{k=1}^q (\partial/\partial z_k) f_k(\mathbf{z}) \approx \text{tr}[\Sigma \hat{\mathbf{A}}] \psi(r_i)$  as the second term in (5) is negligible. This completes the sketch of the proof of (13).

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

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

where  $(\hat{\varepsilon}_i^n, \tilde{Z}_i^n)$  converges in distribution to 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 $\Sigma$ is known

The approximations of the previous sections for  $r_i + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(r_i)$  and the fact that  $\varepsilon_i$  is independent of  $Z_i \sim N(0, 1)$  in (11) suggest that  $(r_i + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(r_i))^2 \approx \varepsilon_i^2 + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 Z_i^2$ ; and averaging over  $\{1, \dots, n\}$  one can hope for the approximation  $\|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2/n \approx \|\varepsilon\|^2/n + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2$ . The following result makes this heuristic precise.

**Theorem 4.1.** *Let Assumption 1.1 be fulfilled and  $\hat{\mathbf{A}}$  be given by Theorem 2.1. Then*

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

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

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 + \|\varepsilon\|^2/n = (1 + O_P(n^{-1/2}))\|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2/n.$$

Theorem 4.1 provides a first candidate,  $\|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2/n$  to estimate

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 + \|\varepsilon\|^2/n. \quad (16)$$

Estimation of (16) is useful as  $\|\varepsilon\|^2/n$  is independent of the choice of the estimator  $\hat{\beta}$  and in particular independent of the chosen loss-penalty pair in (1). Given two or more estimators (1), choosing the one with smallest  $\|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2$  is thus a good proxy for minimizing the out-of-sample error.

**Corollary 4.2.** *Let  $\hat{\beta}, \tilde{\beta}$  be two  $M$ -estimators (1) Assumption 1.1 with loss-penalty pair  $(\rho, g)$  and  $(\tilde{\rho}, \tilde{g})$  respectively. Assume that both satisfy Assumption 1.1 and let  $\psi = \rho'$  and  $\tilde{\psi} = \tilde{\rho}'$ . Let  $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\beta}, \tilde{\mathbf{r}} = \mathbf{y} - \mathbf{X}\tilde{\beta}$  be the residuals,  $\hat{\mathbf{A}}, \tilde{\mathbf{A}}$  be the corresponding matrices of size  $p \times p$  given by Theorem 2.1. Further assume that both estimators satisfy Assumption 1.2 and that  $\varepsilon$  has iid coordinates independent with  $\mathbb{E}[|\varepsilon_i|^{1+q}] \leq M$  for constants  $q \in (0, 1), M > 0$  independent of  $n, p$ . Let  $\Omega = \{\|\mathbf{X}\Sigma^{-1/2}\|_{op} \leq 2\sqrt{n} + \sqrt{p}\} \cap \{\|\varepsilon\|^2 \leq n^{2/(1+q)}\}$ . 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*

$$\mathbb{P}\left(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 - \|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|^2 > \eta, \|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2 \leq \|\tilde{\mathbf{r}} + \text{tr}[\Sigma \tilde{\mathbf{A}}]\tilde{\psi}(\tilde{\mathbf{r}})\|^2\right) \leq C(\gamma, \mu, \eta, q, M)n^{-q/(1+q)} + \mathbb{P}(\Omega^c) \rightarrow 0.$$

Provided that the noise random variables  $\varepsilon_i$  have at least  $1 + q$  moments, Corollary 4.2 implies that with probability approaching one given two  $M$ -estimators  $\hat{\beta}$  and  $\tilde{\beta}$ , choosing the estimator corresponding to the smallest criteria among  $\|\mathbf{r} + \text{tr}[\Sigma \hat{\mathbf{A}}]\psi(\mathbf{r})\|^2$  and  $\|\tilde{\mathbf{r}} + \text{tr}[\Sigma \tilde{\mathbf{A}}]\tilde{\psi}(\tilde{\mathbf{r}})\|^2$  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).

**Corollary 4.3.** *As in Corollary 4.2, assume  $\mathbb{E}[|\varepsilon_i|^{1+q}] \leq M$  and let  $\hat{\beta}_1, \dots, \hat{\beta}_K$  be  $M$ -estimators of the form (1) with loss-penalty pair  $(\rho_k, g_k)$  satisfying Assumptions 1.1 and 1.2. For each  $k = 1, \dots, K$ , let  $\mathbf{r}_k = \mathbf{y} - \mathbf{X}\hat{\beta}_k$  be the residuals and  $\hat{\mathbf{A}}_k$  be the corresponding matrix of size  $p \times p$  from Theorem 2.1. Let  $\hat{k} \in \text{argmin}_{k=1, \dots, K} \|\mathbf{r}_k + \text{tr}[\Sigma \hat{\mathbf{A}}_k]\psi_k(\mathbf{r}_k)\|$  where  $\psi_k = \rho'_k$ . Then if  $(\gamma, \mu, \eta, q, M)$  are constants independent of  $n, p$*

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

Given  $K$  different loss-penalty pairs and the corresponding  $M$ -estimators in (1), minimizing the criterion  $\|\mathbf{r} + \text{tr}[\Sigma\hat{\mathbf{A}}]\mathbf{r}\|$  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` [16]) with tuning parameters evenly spaced in a log-scale that consequently have cardinality logarithmic in the parameter range. The major drawback of the criterion  $\|\mathbf{r} + \text{tr}[\Sigma\hat{\mathbf{A}}]\mathbf{r}\|$  is the dependence through  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  on the covariance  $\Sigma$  of the design, which is typically unknown. The next section introduces an estimator of  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  that does not require the knowledge of  $\Sigma$ .

## 5 Degrees of freedom and estimating $\text{tr}[\Sigma\hat{\mathbf{A}}]$ without the knowledge of $\Sigma$

This section focuses on estimating  $\text{tr}[\Sigma\hat{\mathbf{A}}]$ . The matrix  $\hat{\mathbf{A}}$  from Theorem 2.1 can be estimated from the data  $(\mathbf{y}, \mathbf{X})$  in the sense that  $\hat{\mathbf{A}}$  is a measurable function of  $(\mathbf{y}, \mathbf{X})$  (thanks to the observation that derivatives are limits, and limits of measurable functions are again measurable). The difficulty is thus to estimate  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  without the knowledge of  $\Sigma$ . To illustrate this difficulty, consider Ridge regression with square loss  $\rho(u) = u^2/2$  and penalty  $g(\mathbf{b}) = \tau\|\mathbf{b}\|^2/2$ . Then  $\hat{\beta}(\mathbf{y}, \mathbf{X}) = (\mathbf{X}^\top \mathbf{X} + \tau n \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$  and  $\hat{\mathbf{A}}$  in Theorem 2.1 is given explicitly by  $\hat{\mathbf{A}} = (\mathbf{X}^\top \mathbf{X} + \tau n \mathbf{I}_p)^{-1}$  and

$$\text{tr}[\Sigma\hat{\mathbf{A}}] = \text{tr}[(\mathbf{G}^\top \mathbf{G} + n\tau\Sigma^{-1})^{-1}], \quad \text{where } \mathbf{G} = \mathbf{X}\Sigma^{-1/2}.$$

Above,  $\mathbf{G}$  is a random matrix with iid  $N(0, 1)$  entries the value of  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  is highly dependent on the spectrum of  $\Sigma^{-1}$ . In this particular case, the limit of  $\text{tr}[(\mathbf{G}^\top \mathbf{G} + n\tau\Sigma^{-1})^{-1}]$  can be obtained using random matrix theory [14] as the limiting behavior of the Stieltjes transform of  $\mathbf{G}^\top \mathbf{G}/n + \tau\Sigma^{-1}$  and its spectral distribution is known; however the limit of the spectral distribution depends on the spectrum of  $\tau\Sigma^{-1}$ . This is not desirable here as we wish to construct estimators that require no knowledge on  $\Sigma$ . For more involved loss-penalty pairs such as the Elastic-Net in Example 6.1, such random matrix theory results do not apply as  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  depends on the random support of  $\hat{\beta}$ .

Instead, we do not rely on known random matrix theory results. With the matrix  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  given by Theorem 2.1, our proposal to estimate  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  is the ratio  $\hat{\text{df}}/\text{tr}[\mathbf{V}]$  with  $\hat{\text{df}}$  and  $\mathbf{V}$  in (6)-(7). Both the scalar  $\hat{\text{df}}$  and the matrix  $\mathbf{V} \in \mathbb{R}^{n \times n}$  are observable; in particular they do not depend on  $\Sigma$ .

**Theorem 5.1.** *Let Assumption 1.1 be fulfilled and  $\hat{\mathbf{A}}$  be given by Theorem 2.1. Then*

$$\mathbb{E}[|\text{tr}[\Sigma\hat{\mathbf{A}}]\text{tr}[\mathbf{V}]/n - \hat{\text{df}}/n|] \leq C_2(\gamma, \mu)n^{-1/2}. \quad (17)$$

Simulations in Figure 3 and Table 1 confirm that the approximation  $\text{tr}[\Sigma\hat{\mathbf{A}}] \approx \hat{\text{df}}/\text{tr}[\mathbf{V}]$  is accurate for the Huber loss with Elastic-Net penalty. For the square loss,  $\psi' = 1$  and  $\text{tr}[\mathbf{V}] = n - \hat{\text{df}}$  so that (17) becomes  $\mathbb{E}[|(1 - \hat{\text{df}}/n)(1 + \text{tr}[\Sigma\hat{\mathbf{A}}]) - 1|] \leq C_3(\gamma, \mu)n^{-1/2}$  and the following result holds.

**Corollary 5.2.** *Let Assumption 1.1 be fulfilled with  $\rho(u) = u^2/2$  and  $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Then  $(1 - \hat{\text{df}}/n)(1 + \text{tr}[\Sigma\hat{\mathbf{A}}]) \xrightarrow{\mathbb{P}} 1$  and the normality (14) holds with  $1 + \text{tr}[\Sigma\hat{\mathbf{A}}]$  replaced by  $(1 - \hat{\text{df}}/n)^{-1}$ .*

For general loss  $\rho$ , the criterion (2) replaces  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  by  $\hat{\text{df}}/\text{tr}[\mathbf{V}]$  in the proxy of the out-of-sample error  $\|\mathbf{r} + \text{tr}[\Sigma\hat{\mathbf{A}}]\psi(\mathbf{r})\|^2$  studied in the previous section. Thanks to (17), this replacement preserves the good properties of  $\|\mathbf{r} + \text{tr}[\Sigma\hat{\mathbf{A}}]\psi(\mathbf{r})\|^2$  proved in Corollaries 4.2 and 4.3.

**Theorem 5.3.** *For  $k = 1, \dots, K$ , let  $(\rho_k, g_k)$  be a loss-penalty pair satisfying Assumptions 1.1 and 1.2 with  $\psi_k = \rho'_k$ , let  $\hat{\beta}_k, \mathbf{r}_k, \hat{\mathbf{A}}_k$  be the corresponding  $M$ -estimator residual vector and matrix of size  $p \times p$  given by Theorem 2.1 as in Corollary 4.3 and let  $\hat{\text{df}}_k = \text{tr}[\mathbf{X}\mathbf{A}_k\mathbf{X}^\top \text{diag}\{\psi'_k(\mathbf{r}_k)\}]$  and  $\mathbf{V}_k = \text{diag}\{\psi'_k(\mathbf{r}_k)\}(\mathbf{I}_n - \mathbf{X}\mathbf{A}_k\mathbf{X}^\top \text{diag}\{\psi'_k(\mathbf{r}_k)\})$ . For a small constant  $\eta > 0$  independent of  $n, p$ , say  $\eta = 0.05$ , define*

$$\hat{k} \in \underset{k=1, \dots, K}{\text{argmin}} \left\| \mathbf{r}_k + \frac{\hat{\text{df}}_k}{\text{tr}[\mathbf{V}_k]} \psi_k(\mathbf{r}_k) \right\|^2 \quad \text{subject to} \quad \frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta.$$

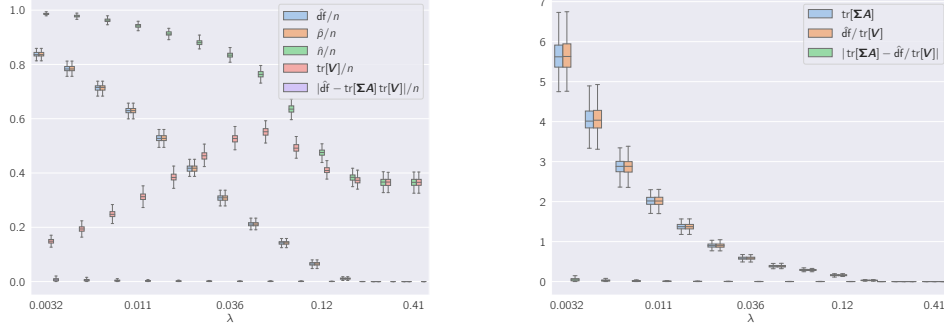
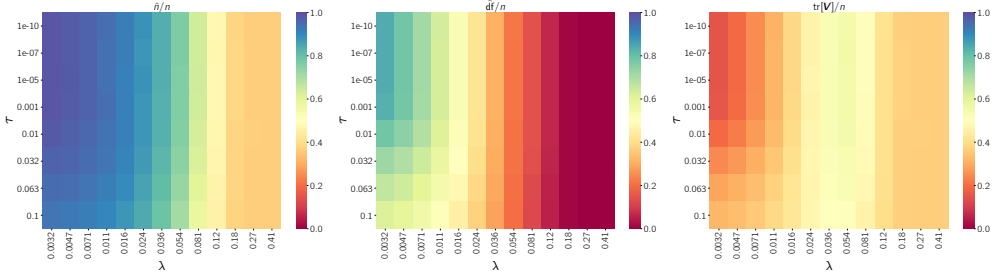


Figure 3: Above: Boxplots for  $\hat{df}$ ,  $\hat{p}$ ,  $\hat{n}$ ,  $\text{tr}[\mathbf{V}]$ ,  $\text{tr}[\Sigma\hat{\mathbf{A}}]$  and  $|\text{tr}[\Sigma\hat{\mathbf{A}}] - \hat{df}/\text{tr}[\mathbf{V}]|$  in Huber Elastic-Net regression with  $\tau = 10^{-10}$  and  $\lambda \in [0.0032, 0.41]$ . Each box contains 200 data points. Below: heatmaps for  $\hat{df}/n$ ,  $\text{tr}[\mathbf{V}]/n$  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.



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

$$\mathbb{P}\left(\|\Sigma^{1/2}(\hat{\beta}_k - \beta^*)\| > \min_{k=1, \dots, K: \frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta} \|\Sigma^{1/2}(\hat{\beta}_k - \beta^*)\| + \tilde{\eta}\right) \rightarrow 0 \quad \text{if } K = o(n^{q/(1+q)}).$$

223 Figure 1 illustrates on simulations the success of the criterion (2) over a grid of tuning parameters  
 224 for  $M$ -estimators with the Huber loss and Elastic-Net penalty. The criterion (2) is thus successful  
 225 at selecting a  $M$ -estimator with smallest out-of-sample error up to an additive constant  $\tilde{\eta}$ , among  
 226 those  $M$ -estimators indexed in  $\{1, \dots, K\}$  that are such that  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta$ . On the one hand  
 227 it is unclear to us whether the restriction  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta$ ; on the other hand there is a practical  
 228 meaning in excluding  $M$ -estimators with small  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki})$ : For the Huber loss  $H(u) := u^2/2$   
 229 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  
 230 in  $\{1, \dots, n\}$  such that the residual  $y_i - \mathbf{x}_i^\top \hat{\beta}_k$  fall within the quadratic regime of the loss function.  
 231 Observations  $i \in \{1, \dots, n\}$  that fall in the linear regime of the loss are excluded from the fit, in the  
 232 sense that for some  $i$  with  $r_{ki} = y_i - \mathbf{x}_i^\top \hat{\beta}_k > 1$ , replacing  $y_i$  by  $\tilde{y}_i = y_i + 1000$  (or any positive value)  
 233 does not change the  $M$ -estimator solution  $\hat{\beta}_k$  (this can be seen from the KKT conditions directly,  
 234 or by integration the derivative with respect to  $y_i$  in (5)). Thus the constraint  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta$   
 235 requires that at most a constant fraction of the observations are excluded from the fit (or equivalently,  
 236 at least a constant fraction of the  $n$  observations participate in the fit). For scaled versions of the  
 237 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  
 238 the number of residuals falling in the quadratic regime of the loss, i.e., the number of observations  
 239 participating in the fit. The heatmaps of Figure 3 illustrate  $\hat{n}$  in a simulation for a wide range of  
 240 parameters. Similarly, for smooth robust loss functions such as  $\rho_k(u) = \sqrt{1 + u^2}$ , the constraint  
 241  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{ki}) \geq \eta$  requires that at most a constant fraction of the  $n$  observations are such that  
 242  $\psi'_k(r_{ki}) < \eta/2$ , i.e., such that the second derivative  $\psi'_k$  is too small (and the loss  $\rho_k$  too flat).

243 Theorems 2.1, 3.2, 4.1 and 5.1 provide our general results applicable to a single regularized  $M$ -  
 244 estimator (1) while corollaries such as Theorem 5.3 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

$$\rho(u; \Lambda) = \Lambda^2 H(\Lambda^{-1} u), \quad g(\mathbf{b}; \lambda, \tau) = \lambda \|\mathbf{b}\|_1 + (\tau/2) \|\mathbf{b}\|_2^2 \quad (18)$$

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 6.1.** With  $(\rho, g)$  in (18), matrix  $\hat{\mathbf{A}}$  in (5) matrix  $\mathbf{V}$  in (7) and  $\hat{\mathbf{d}}\mathbf{f}$  in (6) we have

$$\begin{aligned} \hat{\mathbf{A}}_{\hat{S}, \hat{S}} &= (\mathbf{X}_{\hat{S}}^\top \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X}_{\hat{S}} + n\tau \mathbf{I}_{\hat{p}})^{-1}, \quad A_{i,j} = 0 \text{ if } i \notin \hat{S} \text{ or } j \notin \hat{S}, \\ \mathbf{V} &= \text{diag}\{\psi'(\mathbf{r})\} - \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X}_{\hat{S}} (\mathbf{X}_{\hat{S}}^\top \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X}_{\hat{S}} + n\tau \mathbf{I}_{\hat{p}})^{-1} \mathbf{X}_{\hat{S}}^\top \text{diag}\{\psi'(\mathbf{r})\}, \\ \hat{\mathbf{d}}\mathbf{f} &= \text{tr}[\mathbf{X}_{\hat{S}} (\mathbf{X}_{\hat{S}}^\top \text{diag}\{\psi'(\mathbf{r})\} \mathbf{X}_{\hat{S}} + n\tau \mathbf{I}_{\hat{p}})^{-1} \mathbf{X}_{\hat{S}}^\top \text{diag}\{\psi'(\mathbf{r})\}], \end{aligned} \quad (19)$$

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

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

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 (19) are proved in [3, §2.6]. Simulations in Figures 1 to 3 and Table 1 illustrate typical values for  $\hat{\mathbf{d}}\mathbf{f}$ ,  $\text{tr}[\mathbf{V}]$ ,  $\text{tr}[\Sigma \hat{\mathbf{A}}]$ , the out-of-sample error and the criterion (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  $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$  with anisotropic Gaussian design  $\Sigma$  and heavy-tail noise vector  $\varepsilon$ . The design matrix  $\mathbf{X}$  has  $n = 1001$  rows and  $p = 1000$  columns. Each row of  $\mathbf{X}$  is i.i.d.  $N(\mathbf{0}, \Sigma)$ , with the same  $\Sigma$  across all repetitions, generated once by  $\Sigma = \mathbf{R}^\top \mathbf{R}/(2p)$  with  $\mathbf{R} \in \mathbb{R}^{2p \times p}$  being a Rademacher matrix with i.i.d. entries  $\mathbb{P}(\mathbf{R}_{ij} = \pm 1) = \frac{1}{2}$ . The true signal vector  $\beta^* \in \mathbb{R}^p$  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  $\varepsilon \in \mathbb{R}^n$  has i.i.d. entries from the  $t$ -distribution with 2 degrees of freedom (so that  $\text{Var}[\varepsilon_i] = \infty$ , i.e.,  $\varepsilon_i$  is heavy-tailed).

**Estimation Process.** Each dataset  $(\mathbf{y}, \mathbf{X})$  is fitted by a Huber Elastic-Net estimator with loss-penalty pair in (18). 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  $(\lambda, \tau)$  is then set so that  $\hat{\mathbf{d}}\mathbf{f}/n$  varies on the grid from 0 to 1 (cf. the middle heatmap in Figure 3). The Elastic-Net penalty  $g(\mathbf{b}; \lambda, \tau) = \lambda \|\mathbf{b}\|_1 + (\tau/2) \|\mathbf{b}\|_2^2$  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.

## References

- [1] 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.

- [2] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.
- [3] Pierre C Bellec. Out-of-sample error estimate for robust m-estimators with convex penalty. *arXiv:2008.11840*, 2020.
- [4] 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.
- [5] Pierre C Bellec and Cun-Hui Zhang. Second order poincare inequalities and de-biasing arbitrary convex regularizers when  $p/n \rightarrow \gamma$ . *arXiv:1912.11943*, 2019.
- [6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
- [7] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019.
- [8] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.
- [9] Sourav Chatterjee. Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields*, 143(1):1–40, 2009.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Iosif Pinelis (<https://mathoverflow.net/users/36721/iosif-pinelis>). Large deviations: Growth of empirical average of iid non-negative random varialbes with infinite expectations? MathOverflow. URL:<https://mathoverflow.net/q/390939> (version: 2021-05-24).
- [14] 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.
- [15] 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.
- [16] 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.
- [17] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.
- [18] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.
- [19] 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.
- [20] William P Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120. Springer-Verlag New York, 1989.

## Checklist

### 1. For all authors...

- (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
- (b) Did you describe the limitations of your work? [Yes] See Assumptions 1.1 and 1.2 and the limitations mentioned in Section 1.2
- (c) Did you discuss any potential negative societal impacts of your work? [N/A]
- (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

### 2. If you are including theoretical results...

- (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1.1 and 1.2.
- (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary material.

### 3. If you ran experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See code in supplementary material.
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The code is also provided.
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] It takes an Amazon EC2 server approximately 40 hours to generate all our simulation results. This is also mentioned in supplementary.

### 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

- (a) If your work uses existing assets, did you cite the creators? [N/A] Simulations are implemented using Python.
- (b) Did you mention the license of the assets? [N/A]
- (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Simulated data only.

### 5. If you used crowdsourcing or conducted research with human subjects...

- (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
- (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

362 **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  
 363  $1 \leq q \leq +\infty$ . For matrices,  $\|\cdot\|_{op}$  is the operator norm (largest singular value),  $\|\cdot\|_F$  the Frobenius  
 364 norm. We use index  $i$  only to loop or sum over  $[n] = \{1, \dots, n\}$  and  $j$  only to loop or sum over  
 365  $[p] = \{1, \dots, p\}$ , so that  $e_i \in \mathbb{R}^n$  refers to the  $i$ -th canonical basis vector in  $\mathbb{R}^n$  and  $e_j \in \mathbb{R}^p$  the  $j$ -th  
 366 canonical basis vector in  $\mathbb{R}^p$ . Positive absolute constants are denoted  $C_0, C_1, C_2, \dots$ , constants that  
 367 depend on  $\gamma$  only are denoted  $C_0(\gamma), C_1(\gamma), \dots$  and constant that depend on  $\gamma, \mu$  only are denoted by  
 368  $C_0(\gamma, \mu), C_1(\gamma, \mu), \dots$ . If  $\mathbf{f} : \mathbb{R}^q \rightarrow \mathbb{R}^n$  is differentiable at  $\mathbf{z} \in \mathbb{R}^q$ , we denote the Jacobian matrix  
 369 in  $\mathbb{R}^{n \times q}$  by  $\frac{\partial \mathbf{f}}{\partial \mathbf{z}}$  or  $\partial \mathbf{f} / \partial \mathbf{z}$ . For an event  $\Omega$ , its indicator function is denoted by  $I_\Omega$  or  $I\{\Omega\}$ .

370 **Organization of the proofs.** Section 7 provides the proof of the main results from the main text  
 371 (Theorems 3.1, 3.2, 4.1, 5.1 and 5.3 and Corollaries 4.2 and 4.3) and the overall proof strategy. Sec-  
 372 tion 8 gives the proof of the probabilistic tools used in Section 7. Section 9 proves the differentiability  
 373 formulae in Theorem 2.1 and Remark 2.2.

374 **Additional simulations.** Additional simulations and figures are given in Section 10 for Gaussian  
 375 designs and in Section 11 for non-Gaussian Rademacher design. The simulations for Rademacher  
 376 design suggests that our results generalize to non-Gaussian design, although it is unclear at this point  
 377 how to extend the proofs to non-Gaussian  $\mathbf{X}$ .

378 Simulations were run on an Amazon EC2 c5.4xlarge instance for about 40 hours.

## 379 7 Proof of the main results

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

$$\mathbf{h}(\varepsilon, \mathbf{G}) = \underset{\mathbf{u} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \rho(\varepsilon_i - \mathbf{e}_i^\top \mathbf{G} \mathbf{u}) + g(\beta^* + \mathbf{\Sigma}^{-1/2} \mathbf{u}) \quad (20)$$

382 and denote by  $(h_j)_{j=1, \dots, p}$  the components of (20). Then  $\mathbf{\Sigma}^{1/2}(\widehat{\beta}(\mathbf{y}, \mathbf{X}) - \beta^*) = \mathbf{h}(\varepsilon, \mathbf{X})$  with  
 383  $\widehat{\beta}(\mathbf{y}, \mathbf{X})$  the  $M$ -estimator in (1). With  $\mathbf{y} = \mathbf{G}\mathbf{\Sigma}^{1/2}\beta^* + \varepsilon$ , by the chain rule and (5),

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

384 Define  $\psi(\varepsilon, \mathbf{G}) = \psi(\varepsilon - \mathbf{G}\mathbf{h})$ . With  $\mathbf{e}_i \in \mathbb{R}^n, \mathbf{e}_j \in \mathbb{R}^p$  denoting canonical basis vectors,

$$(\partial/\partial g_{ij})\mathbf{h}(\varepsilon, \mathbf{G}) = \mathbf{A}\mathbf{e}_j \psi(r_i) - \mathbf{A}\mathbf{G}^\top \mathbf{e}_i \psi'(r_i)h_j \quad (21)$$

$$(\partial/\partial g_{ij})\psi(\varepsilon, \mathbf{G}) = -\operatorname{diag}\{\psi'(\mathbf{r})\}\mathbf{G}\mathbf{A}\mathbf{e}_j \psi(r_i) - \mathbf{V}\mathbf{e}_i h_j \quad (22)$$

385 where the second line follows by the chain rule for Lipschitz functions in in [20, Theorem 2.1.11].  
 386 The crux of the argument is that the quantities of interest appearing in our results,  $\|\mathbf{h}\|^2 = \|\mathbf{\Sigma}^{1/2}(\widehat{\beta} - \beta^*)\|^2$ ,  $\|\psi(\mathbf{r})\|^2$ ,  $\operatorname{tr}[\widehat{\mathbf{A}}\mathbf{\Sigma}] = \operatorname{tr}[\mathbf{A}], \operatorname{tr}[\mathbf{V}]$  and  $\widehat{\mathbf{d}}$  naturally appear from tensor contractions involving  
 387 the derivatives in (21)-(22). For instance, denoting  $\mathbf{D} = \operatorname{diag}\{\psi'(\mathbf{r})\} \in \mathbb{R}^{n \times n}$  if  $h_j, \psi_i$  are the  $j$ -th  
 388

and  $i$ -th component of (20) and  $\psi(\varepsilon, G)$  and denoting  $\sum_{i=1}^n \sum_{j=1}^p$  by  $\sum_{ij}$  for brevity,

$$\sum_{j=1}^p \frac{\partial h_j}{g_{ij}} = \text{tr}[\mathbf{A}] \psi_i - \mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{D} \mathbf{e}_i \quad \text{for a given } i = 1, \dots, n, \quad (23)$$

$$\sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}} = -\psi^\top \mathbf{D} \mathbf{G} \mathbf{A} \mathbf{e}_j - \text{tr}[\mathbf{V}] h_j \quad \text{for a given } j = 1, \dots, p, \quad (24)$$

$$\sum_{ij} \frac{\partial(h_j \psi_i)}{g_{ij}} = \|\psi\|^2 \text{tr}[\mathbf{A}] - \mathbf{h}^\top \mathbf{G}^\top \mathbf{D} \psi - \psi^\top \mathbf{D} \mathbf{G} \mathbf{A} \mathbf{h} - \|\mathbf{h}\|^2 \text{tr}[\mathbf{V}], \quad (25)$$

$$\sum_{ij} \frac{\partial(h_j \mathbf{e}_i^\top \mathbf{G} \mathbf{h})}{g_{ij}} = \text{tr}[\mathbf{A}] \psi^\top \mathbf{G} \mathbf{h} - \mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{G} \mathbf{h} + n \|\mathbf{h}\|^2 + \psi^\top \mathbf{G} \mathbf{A} \mathbf{h} - \|\mathbf{h}\|^2 \hat{\text{df}}, \quad (26)$$

$$\sum_{ij} \frac{\partial(\psi_i \mathbf{e}_j^\top \mathbf{G}^\top \psi)}{g_{ij}} = -\psi^\top \mathbf{D} \mathbf{G} \mathbf{A} \mathbf{G}^\top \psi - \text{tr}[\mathbf{V}] \psi^\top \mathbf{G} \mathbf{h} - \mathbf{h}^\top \mathbf{G}^\top \mathbf{V} \psi + (p - \hat{\text{df}}) \|\psi\|^2 \quad (27)$$

where we used that  $\hat{\text{df}} = \sum_{i=1}^n \mathbf{e}_i^\top \mathbf{G} \mathbf{A} \mathbf{G}^\top \mathbf{D} \mathbf{e}_i = \text{tr}[\mathbf{G} \mathbf{A} \mathbf{G}^\top \mathbf{D}]$  in the fourth line and  $\hat{\text{df}} = \sum_{j=1}^p \mathbf{e}_j^\top \mathbf{G}^\top \mathbf{D} \mathbf{G} \mathbf{A} \mathbf{e}_j = \text{tr}[\mathbf{G}^\top \mathbf{D} \mathbf{G} \mathbf{A}]$  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 7.1.** [Variant of [5]] Let  $\mathbf{z} \in N(\mathbf{0}, \mathbf{I}_q)$  and  $\mathbf{f} := \mathbf{f}(\mathbf{z}) : \mathbb{R}^q \rightarrow \mathbb{R}^q \setminus \{\mathbf{0}\}$  be locally Lipschitz in  $\mathbf{z}$  with  $\mathbb{E}[\|\mathbf{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \mathbf{f}}{\partial z_k}\|^2] < +\infty$ . Then

$$\mathbb{E}\left[\left(\frac{\mathbf{f}^\top \mathbf{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\mathbf{f}\|_2} - Z\right)^2\right] \leq (7 + 2\sqrt{6}) \mathbb{E}\left[\|\mathbf{f}\|^{-2} \sum_{k=1}^q \left\|\frac{\partial \mathbf{f}}{\partial z_k}\right\|^2\right] < +\infty. \quad (28)$$

Proposition 7.1 is proved in Section 8. 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 (23). We start with the square loss and the proof of Theorem 3.2.

*Proof of Theorem 3.2.* Apply Proposition 7.1 with  $q = p + 1$  and  $\mathbf{z} = (\mathbf{g}_i, \varepsilon_i/\sigma) \sim N(\mathbf{0}, \mathbf{I}_{p+1})$  conditionally on  $(\mathbf{g}_l, \varepsilon_l)_{l \in [n] \setminus \{i\}}$ , and with  $\mathbf{f} = (\mathbf{h}, -\sigma) \in \mathbb{R}^{p+1}$ . Note that the last component of  $\mathbf{f}$  is constant and  $\|\mathbf{f}\|^2 = \|\mathbf{h}\|^2 + \sigma^2$ . By (23) and  $\mathbf{D} = \mathbf{I}_n$  for the square loss,  $\text{tr}[\partial \mathbf{f} / \partial \mathbf{z}] = \text{tr}[\mathbf{A}] \psi_i - \mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{e}_i$  and by symmetry in  $i = 1, \dots, n$ ,  $\mathbb{E}[\|\mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{e}_i\|^2 / \|\mathbf{f}\|^2] = \frac{1}{n} \sum_{l=1}^n \mathbb{E}[\|\mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{e}_l\|^2 / \|\mathbf{f}\|^2] = \frac{1}{n} \mathbb{E}[\|\mathbf{G} \mathbf{A}^\top \mathbf{h}\|^2 / \|\mathbf{f}\|^2] \leq \frac{1}{n} \mathbb{E}[\|\mathbf{G}\|_{op}^2 \|\mathbf{A}\|_{op}^2] \leq n^{-2} C_4(\gamma, \mu)$  thanks to  $\|\mathbf{A}\|_{op} \leq 1/(n\mu)$  and  $\mathbb{E}[\|\mathbf{G}\|_{op}^2] \leq C_5(\gamma)n$ . Similarly, for the square loss  $r_i = \psi_i = \varepsilon_i - \mathbf{g}_i^\top \mathbf{h}$  and

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

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

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

By symmetry in  $i = 1, \dots, n$ ,  $\mathbb{E}[(\mathbf{g}_i^\top \mathbf{h})^2 / \|\mathbf{h}\|^2] = \frac{1}{n} \sum_{l=1}^n \mathbb{E}[(\mathbf{g}_l^\top \mathbf{h})^2 / \|\mathbf{h}\|^2] \leq \frac{1}{n} \mathbb{E}[\|\mathbf{G}\|_{op}^2]$ . Since  $\frac{1}{n} \mathbb{E}[\|\mathbf{G}\|_{op}^2] \leq C_6(\gamma)$ , the right-hand side in the previous display is bounded from above by  $C_7(\gamma, \mu) n^{-1/2}$ . Since  $\mathbf{f}^\top \mathbf{z} = -r_i$  we obtain  $-r_i - \text{tr}[\mathbf{A}] r_i = (\|\mathbf{h}\|^2 + \sigma^2)^{1/2} (Z + O_P(n^{-1/2}))$  which completes the proof of (14).  $\square$

*Proof of Theorem 3.1.* Let  $U \sim N(0, 1)$  be independent of everything else. We apply the previous proposition with  $\mathbf{z} = (\mathbf{g}_i, U) \sim N(\mathbf{0}, \mathbf{I}_{p+1})$  conditionally on  $(\varepsilon, \mathbf{g}_l, l \in [n] \setminus \{i\})$  to  $\mathbf{f} = (\mathbf{h}, n^{-1/4} \psi(\varepsilon_i))$ . Note that the last component of  $\mathbf{f}$  is constant. By (23),  $\text{tr}[\partial \mathbf{f} / \partial \mathbf{z}] = \text{tr}[\mathbf{A}] \psi_i - \mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{D} \mathbf{e}_i$  and by (21),

$$\|\mathbf{f}\|^{-1} \|\partial \mathbf{f} / \partial \mathbf{z}\|_F = (\|\mathbf{h}\|^2 + n^{-1/2} \psi(\varepsilon_i)^2)^{-1/2} \|\mathbf{A} \psi_i - \mathbf{A} \mathbf{G}^\top \mathbf{D} \mathbf{e}_i \mathbf{h}^\top\|_F \quad (29)$$

$$\leq \|\mathbf{A}\|_{op} [n^{1/4} \sqrt{p} + \sqrt{p} \|\mathbf{h}\|^{-1} |\mathbf{g}_i^\top \mathbf{h}| + \|\mathbf{G}\|_{op}] \quad (30)$$

414 where we used  $\|\mathbf{A}\|_F \leq \sqrt{p}\|\mathbf{A}\|_{op}$  and  $|\psi_i| \leq \psi(\varepsilon_i) + |\mathbf{g}_i^\top \mathbf{h}|$  thanks to  $\psi$  being 1-Lipschitz. We  
 415 have  $\|\mathbf{A}\|_{op} \leq 1/(n\mu)$  and  $\mathbb{E}[\|\mathbf{h}\|^{-2}|\mathbf{g}_i^\top \mathbf{h}|^2] = \frac{1}{n} \sum_{l=1}^n \mathbb{E}[\|\mathbf{h}\|^{-2}|\mathbf{g}_l^\top \mathbf{h}|^2] = \frac{1}{n} \mathbb{E}[\|\mathbf{h}\|^{-2}\|\mathbf{G}\mathbf{h}\|^2] \leq$   
 416  $\frac{1}{n} \mathbb{E}[\|\mathbf{G}\|_{op}^2]$  by symmetry in  $i = 1, \dots, n$ , so that  $\mathbb{E}[\|\mathbf{f}\|^{-2}\|\partial \mathbf{f}/\partial \mathbf{z}\|_F^2] \leq n^{-1/2}C_8(\gamma, \mu)$ . Thus by  
 417 Proposition 7.1,

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

418 where  $\mathbb{E}[\text{Rem}^2] \leq C_9 \mathbb{E}[\|\mathbf{f}\|^{-2}\|\partial \mathbf{f}/\partial \mathbf{z}\|_F^2] \leq n^{-1/2}C_{10}(\gamma, \mu)$ . By properties of the operator norm  
 419 and symmetry in  $i = 1, \dots, n$ ,

$$\mathbb{E}[\|\mathbf{h}\|^{-2}|\mathbf{h}^\top \mathbf{A} \mathbf{G}^\top \mathbf{D} \mathbf{e}_i|^2] = \frac{1}{n} \mathbb{E}[\|\mathbf{h}\|^{-2}\|\mathbf{D} \mathbf{G} \mathbf{A}^\top \mathbf{h}\|^2] \leq \frac{1}{n} \mathbb{E}[\|\mathbf{G}\|_{op}^2\|\mathbf{A}\|_{op}^2] \leq \frac{C_{11}(\gamma, \mu)}{n^{-2}}. \quad (31)$$

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

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

420 If  $\varepsilon_i$  has a fixed distribution  $F$ , then  $|\psi(\varepsilon_i)| \leq |\psi(0)| + |\varepsilon_i| = |\varepsilon_i| = O_P(1)$  thanks to  $\psi(0) = 0$  and  
 421  $\psi$  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  
 422 Slutsky's theorem this proves that  $(\tilde{\varepsilon}_i^n, \tilde{Z}_i^n)$  converges weakly to the product measure  $F \otimes N(0, 1)$ .  $\square$

423 **Proposition 7.2.** Let  $\mathbf{h} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$ ,  $\psi : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$  be locally Lipschitz functions. If  $\mathbf{G} \in \mathbb{R}^{n \times p}$   
 424 has iid  $N(0, 1)$  entries then

$$\begin{aligned} \mathbb{E} \left[ \left( \frac{\psi^\top \mathbf{G} \mathbf{h} - \sum_{ij} \frac{\partial(\psi_i h_j)}{g_{ij}}}{\|\mathbf{h}\|^2 + \|\psi\|^2/n} \right)^2 + \left( \frac{\|\mathbf{G} \mathbf{h}\|^2 - \sum_{ij} \frac{\partial(h_j \mathbf{e}_i^\top \mathbf{G} \mathbf{h})}{g_{ij}}}{\|\mathbf{h}\|^2 + \|\psi\|^2/n} \right)^2 + \left( \frac{\|\mathbf{G}^\top \psi\|^2 - \sum_{ij} \frac{\partial(\psi_i \mathbf{e}_j^\top \mathbf{G}^\top \psi)}{g_{ij}}}{n\|\mathbf{h}\|^2 + \|\psi\|^2} \right)^2 \right] \\ \leq C_{12} \mathbb{E} \left[ n + p + \|\mathbf{G}\|_{op}^2 + (n + p) \sum_{i=1}^n \sum_{j=1}^p \frac{1 + \|\mathbf{G}\|_{op}^2/n}{(\|\mathbf{h}\|^2 + \|\psi\|^2/n)^2} \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \psi}{\partial g_{ij}} \right\|^2 \right) \right] \quad (32) \end{aligned}$$

425 for some positive absolute constant in the second line.

426 Proposition 7.2 is proved in Section 8. By Proposition 7.2 combined with the identities (25)-(26)-(27),  
 427 and by showing that the colored terms in purple (25)-(26)-(27) are negligible, we obtain the following.

428 **Proposition 7.3.** Let Assumption 1.1 be fulfilled. Then

$$\mathbb{E} \left[ \left\{ n^{-\frac{1}{2}} (\|\mathbf{h}\|^2 + \|\psi\|^2/n)^{-1} (\psi^\top \mathbf{G} \mathbf{h} - \text{tr}[\mathbf{A}]\|\psi\|^2 + \text{tr}[\mathbf{V}]\|\mathbf{h}\|^2) \right\}^2 \right] \leq C_{13}(\gamma, \mu), \quad (33)$$

$$\mathbb{E} \left[ \left\{ n^{-\frac{1}{2}} (\|\mathbf{h}\|^2 + \|\psi\|^2/n)^{-1} \left( \frac{1}{n} \|\mathbf{G}^\top \psi\|^2 - \frac{p - \hat{\text{df}}}{n} \|\psi\|^2 + \frac{\text{tr}[\mathbf{V}]}{n} \psi^\top \mathbf{G} \mathbf{h} \right) \right\}^2 \right] \leq C_{14}(\gamma, \mu), \quad (34)$$

$$\mathbb{E} \left[ \left\{ n^{-\frac{1}{2}} (\|\mathbf{h}\|^2 + \|\psi\|^2/n)^{-1} (\|\mathbf{G} \mathbf{h}\|^2 - \text{tr}[\mathbf{A}]\psi^\top \mathbf{G} \mathbf{h} - (n - \hat{\text{df}})\|\mathbf{h}\|^2) \right\}^2 \right] \leq C_{15}(\gamma, \mu). \quad (35)$$

429 *Proof.* We bound from above the derivatives in (32). For the norm of  $(\partial/\partial g_{ij})\mathbf{h}$  and  $(\partial/\partial g_{ij})\psi$ , by  
 430 (22)-(21) and  $\frac{1}{2}(a + b)^2 \leq a^2 + b^2$ ,

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

431 Using  $\|\mathbf{A}\|_{op} \leq 1/(n\mu)$ ,  $\|\mathbf{D}\|_{op} \leq 1$ ,  $p/n \leq \gamma$  and  $\mathbf{V}$  in (7), it follows that in (32) we have

$$\frac{1}{\|\mathbf{h}\|^2 + \|\psi\|^2/n} \sum_{i=1}^n \sum_{j=1}^p \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \psi}{\partial g_{ij}} \right\|^2 \right) \leq C_{16}(\gamma, \mu) (1 + \|\mathbf{G}\|_{op}^2/n). \quad (36)$$

432 Since  $\mathbb{E}[\|n^{-1/2}\mathbf{G}\|_{op}^4] \leq C_{17}(\gamma)$  [10, Theorem II.13], this shows that (32) is bounded from above  
 433 by  $C_{18}(\gamma, \mu)n$ . The contractions appearing in the left-hand side of (32) are given in (25)-(26)-(27),  
 434 so that it remains to bound from above the purple colored terms in these three equations. This is  
 435 done by using the upper bounds on the operator norms  $\|\mathbf{A}\|_{op} \leq 1/(n\mu)$ ,  $\|\mathbf{D}\|_{op} \leq 1$  and again that  
 436  $\mathbb{E}[\|n^{-1/2}\mathbf{G}\|_{op}^4] \leq C_{19}(\gamma)$ , so that (32) yields the three inequalities in Proposition 7.3.  $\square$

437 The next result is another probabilistic result where the contractions in (23)-(24) appear.

438 **Proposition 7.4.** *Let  $\mathbf{h} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$ ,  $\boldsymbol{\psi} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$  be locally Lipschitz functions. If  $\mathbf{G} \in \mathbb{R}^{n \times p}$*   
 439 *has iid  $N(0, 1)$  entries then*

$$\begin{aligned} & \mathbb{E} \left[ \frac{\left| \frac{p}{n} \|\boldsymbol{\psi}\|^2 - \frac{1}{n} \sum_{j=1}^p (\boldsymbol{\psi}^\top \mathbf{G} \mathbf{e}_j - \sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}})^2 \right|}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right] + \mathbb{E} \left[ \frac{\left| n \|\mathbf{h}\|^2 - \sum_{i=1}^n (\mathbf{g}_i^\top \mathbf{h} - \sum_{j=1}^p \frac{\partial h_j}{\partial g_{ij}})^2 \right|}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right] \\ & \leq C_{20} \left( \sqrt{n+p} (1 + \Xi^{1/2}) + \Xi \right) \text{ where } \Xi = \mathbb{E} \left[ \frac{1}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \sum_{i=1}^n \sum_{j=1}^p \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2 \right) \right]. \end{aligned}$$

440 The proof of Proposition 7.4 is given in Section 8. Using the contractions (23)-(24) in the left-hand  
 441 side of Proposition 7.4, and by showing that the purple colored terms are negligible, we obtain the  
 442 following two inequalities.

443 **Proposition 7.5.** *Let Assumption 1.1 be fulfilled. Then*

$$\mathbb{E} \left| n^{-\frac{1}{2}} (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1} \left( \frac{p}{n} \|\boldsymbol{\psi}\|^2 - \frac{1}{n} \|\mathbf{G}^\top \boldsymbol{\psi} + \text{tr}[\mathbf{V}] \mathbf{h}\|^2 \right) \right| \leq C_{21}(\gamma, \mu), \quad (37)$$

$$\mathbb{E} \left| n^{-\frac{1}{2}} (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1} (n \|\mathbf{h}\|^2 - \|\mathbf{G} \mathbf{h} - \text{tr}[\mathbf{A}] \boldsymbol{\psi}\|^2) \right| \leq C_{22}(\gamma, \mu). \quad (38)$$

444 *Proof.* For  $\Xi$  in Proposition 7.4, the fact that  $\Xi \leq C_{23}(\gamma, \mu)$  is already proved in (36). For the first  
 445 inequality we use Proposition 7.4 and the contraction (24). To control the purple terms in (24) inside  
 446 the left-hand side of Proposition 7.5,

$$\begin{aligned} & \left| \sum_{j=1}^p (\boldsymbol{\psi}^\top \mathbf{G} \mathbf{e}_j - \sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}})^2 - \|\mathbf{G}^\top \boldsymbol{\psi} + \text{tr}[\mathbf{V}] \mathbf{h}\|^2 \right| = \left| \boldsymbol{\psi}^\top \mathbf{D} \mathbf{G} \mathbf{A} (2 \mathbf{G}^\top \boldsymbol{\psi} + 2 \text{tr}[\mathbf{V}] \mathbf{h} + \mathbf{A}^\top \mathbf{G}^\top \mathbf{D} \boldsymbol{\psi}) \right| \\ & \leq (\|\boldsymbol{\psi}\|^2/n + \|\mathbf{h}\|^2) (2n \|\mathbf{G}\|_{op}^2 \|\mathbf{A}\|_{op} + 2\sqrt{n} \|\mathbf{G}\|_{op} \|\mathbf{A}\|_{op} + n \|\mathbf{A}\|_{op}^2 \|\mathbf{G}\|_{op}^2) \end{aligned}$$

447 thanks to  $|\text{tr} \mathbf{V}| \leq n$  in Theorem 2.1. With the bound obtained by multiplying the previous display  
 448 by  $n^{-3/2} (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1}$ , and using the previous bounds on  $\|\mathbf{A}\|_{op}$  and  $\mathbb{E}[\|n^{-1/2} \mathbf{G}\|_{op}^2]$ , we  
 449 obtain (37) from Proposition 7.4 and (24). The second claim is obtained by Proposition 7.4, the  
 450 contraction (23) and an argument similar to the previous display bound the purple term in (23).  $\square$

451 We are now ready to prove Theorem 5.1.

452 *Proof of Theorem 5.1.* Define

$$\begin{aligned} \xi_I &= \boldsymbol{\psi}^\top \mathbf{G} \mathbf{h} - \text{tr}[\mathbf{A}] \|\boldsymbol{\psi}\|^2 + \text{tr}[\mathbf{V}] \|\mathbf{h}\|^2 && \text{(bounded in (33))}, \\ \xi_{II} &= \frac{1}{n} \|\mathbf{G}^\top \boldsymbol{\psi}\|^2 - \frac{p - \hat{\text{df}}}{n} \|\boldsymbol{\psi}\|^2 + \frac{\text{tr}[\mathbf{V}]}{n} \boldsymbol{\psi}^\top \mathbf{G} \mathbf{h} && \text{(bounded in (34))}, \\ \xi_{III} &= \|\mathbf{G} \mathbf{h}\|^2 - \text{tr}[\mathbf{A}] \boldsymbol{\psi}^\top \mathbf{G} \mathbf{h} - (n - \hat{\text{df}}) \|\mathbf{h}\|^2 && \text{(bounded in (35))}, \\ \xi_{IV} &= \frac{p}{n} \|\boldsymbol{\psi}\|^2 - \frac{1}{n} \|\mathbf{G}^\top \boldsymbol{\psi} + \text{tr}[\mathbf{V}] \mathbf{h}\|^2 && \text{(bounded in (37))}, \\ \xi_V &= n \|\mathbf{h}\|^2 - \|\mathbf{G} \mathbf{h} - \text{tr}[\mathbf{A}] \boldsymbol{\psi}\|^2 && \text{(bounded in (38))}. \end{aligned}$$

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),

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

453 Since  $|\text{tr}[\mathbf{V}]/n| \leq 1$ ,  $\text{tr}[\mathbf{A}] \leq \gamma/\mu$  by Theorem 2.1, the previous display divided by  $n^{1/2} (\|\boldsymbol{\psi}\|^2/n +$   
 454  $\|\mathbf{h}\|^2)$  and the bounds (33), (34), (35), (37) and (38) complete the proof.  $\square$

455 To prove Theorem 4.1, we need this extra proposition whose proof is closely related to Proposition 7.3.  
 456

457 **Proposition 7.6.** *Let Assumption 1.1 be fulfilled. Then*

$$\mathbb{E} \left[ \left\{ (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-\frac{1}{2}} \|\boldsymbol{\varepsilon}\|^{-1} \xi_{VI} \right\}^2 \right] \leq C_{24}(\gamma, \mu) \quad \text{for} \quad \xi_{VI} = \boldsymbol{\varepsilon}^\top (\mathbf{G} \mathbf{h} - \text{tr}[\mathbf{A}] \boldsymbol{\psi}). \quad (39)$$

458 Proposition 7.6 is proved in Section 8. We are now ready to prove Theorem 4.1.

459 *Proof of Theorem 4.1.* We have  $n\|\mathbf{h}\|^2 + \|\varepsilon\|^2 - \|\mathbf{r} + \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2 = \xi_V + 2\xi_{VI}$  by simple algebra  
460 and the definitions of  $\xi_V$  and  $\xi_{VI}$ . Hence

$$\mathbb{E}\left[\frac{|\|\mathbf{h}\|^2 + \|\varepsilon\|^2/n - \|\mathbf{r} + \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2/n|}{\max\{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n, (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}(\|\varepsilon\|^2/n)^{1/2}\}}\right] \leq n^{-1/2}C_{25}(\gamma, \mu) \quad (40)$$

461 thanks to (39) and (38).  $\square$

462 *Proof of Corollary 4.2.* We perform the change of variable (20) to  $\tilde{\boldsymbol{\beta}}$  as well, giving  $\tilde{\mathbf{h}}$  (the counterpart  
463 of  $\mathbf{h}$ ),  $\tilde{\boldsymbol{\psi}}$  (counterpart of  $\boldsymbol{\psi}$ ) and  $\tilde{\mathbf{A}}$  (counterpart of  $\mathbf{A}$ ). Let  $\Omega$  be the event defined in the theorem, i.e.,

$$\Omega = \{\|\mathbf{G}\|_{op} \leq 2\sqrt{n} + \sqrt{p}\} \cap \{\|\varepsilon\|^2 \leq n^{2/(1+q)}\}. \quad (41)$$

464 Then  $\mathbb{P}(\Omega^c) \rightarrow 0$  by [10, Theorem 2.13] for the first event and [13] to show that  $\|\varepsilon\|^2/n^{2/(1+q)} \rightarrow^{\mathbb{P}} 0$   
465 under the assumption that  $\mathbb{E}[\|\varepsilon_i\|^{1+q}]$  is bounded.

466 Under Assumption 1.2,  $I_\Omega(\|\boldsymbol{\psi}\|^2/n + \|\mathbf{h}\|^2)$  is bounded by a constant. Indeed, since the penalty  
467  $g$  is minimized at  $\mathbf{0}$ ,  $(\tilde{\boldsymbol{\beta}} - \mathbf{0})^\top \mathbf{X}^\top \boldsymbol{\psi} \in n(\tilde{\boldsymbol{\beta}} - \mathbf{0})^\top (\partial g(\tilde{\boldsymbol{\beta}}) - \partial g(\mathbf{0}))$  since  $\mathbf{0} \in \partial g(\mathbf{0})$ . By strong  
468 convexity of  $g$  in Assumption 1.1,  $(\tilde{\boldsymbol{\beta}} - \mathbf{0})^\top \mathbf{X}^\top \boldsymbol{\psi} \geq \mu\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\beta}}\|^2$ . In  $\Omega$ , this implies  $\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{\beta}}\| \leq$   
469  $\frac{1}{\mu n}\|\mathbf{G}\|_{op}\|\boldsymbol{\psi}\| \leq C_{26}(\gamma, \mu)\|\boldsymbol{\psi}\|/\sqrt{n}$  and  $\|\boldsymbol{\psi}\|/\sqrt{n} \leq M$  in Assumption 1.2. Since  $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*\|^2 \leq$   
470  $M$  in Assumption 1.2, this yields  $I_\Omega(\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n) \leq C_{27}(\gamma, \mu, M)$  and the same holds for  $\tilde{\mathbf{h}}, \tilde{\boldsymbol{\psi}}$ :  
471  $I_\Omega(\|\tilde{\mathbf{h}}\|^2 + \|\tilde{\boldsymbol{\psi}}\|^2/n) \leq C_{28}(\gamma, \mu, M)$ .

472 Inequality (40) thus implies

$$\begin{aligned} \mathbb{E}[I_\Omega(|\|\mathbf{h}\|^2 + \|\varepsilon\|^2/n - \|\mathbf{r} + \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2/n| + |\|\tilde{\mathbf{h}}\|^2 + \|\varepsilon\|^2/n - \|\tilde{\mathbf{r}} + \text{tr}[\tilde{\mathbf{A}}]\tilde{\boldsymbol{\psi}}\|^2/n|)] \\ \leq C_{29}(\gamma, \mu, M)(n^{-1/2} \vee n^{-q/(1+q)}). \end{aligned}$$

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  $\hat{\Omega} = \{\|\mathbf{h}\|^2 -$   
 $\|\tilde{\mathbf{h}}\|^2 > \eta, \|\mathbf{r} + \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2 \leq \|\tilde{\mathbf{r}} + \text{tr}[\tilde{\mathbf{A}}]\tilde{\boldsymbol{\psi}}\|^2\}$  be the event for which we are trying to control the  
probability. By the triangle inequality,

$$\mathbb{E}[I_\Omega(|\|\mathbf{h}\|^2 - \|\tilde{\mathbf{h}}\|^2 - \|\mathbf{r} + \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2/n + \|\tilde{\mathbf{r}} + \text{tr}[\tilde{\mathbf{A}}]\tilde{\boldsymbol{\psi}}\|^2/n|)] \leq C_{30}(\gamma, \mu, M)n^{-q/(1+q)}.$$

473 In  $\hat{\Omega}$ , the random variable in the expectation sign is larger than  $\eta I_\Omega$ . Thus  $\eta \mathbb{E}[I_\Omega I_{\hat{\Omega}}] \leq$   
474  $C_{31}(\gamma, \mu, M)n^{-q/(1+q)}$  and  $\mathbb{P}(\hat{\Omega}) \leq \eta^{-1}C_{32}(\gamma, \mu, M)n^{-q/(1+q)} + \mathbb{P}(\Omega^c)$ .  $\square$

475 *Proof of Corollary 4.3.* We follow the same strategy. Let  $\Omega$  be the same event as in the previous  
476 proof, so that  $\mathbb{P}(\Omega^c) \rightarrow 0$  as before. We perform the change of variable (20) for each  $k = 1, \dots, K$   
477 giving  $\mathbf{h}_k, \boldsymbol{\psi}_k$  and  $\mathbf{A}_k$ . We have  $I_\Omega \max_{k=1, \dots, K}(\|\mathbf{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \leq C_{33}(\gamma, \mu, M)$  as explained  
478 in the previous proof.

479 Summing over  $k$  the inequality (40) gives  $\mathbb{E}[I_\Omega \sum_{k=1}^K (\|\mathbf{h}_k\|^2 + \|\varepsilon\|^2 - \|\mathbf{r}_k + \text{tr}[\mathbf{A}_k]\boldsymbol{\psi}_k\|^2)] \leq$   
480  $K C_{34}(\gamma, \mu, M)n^{-q/(1+q)}$ . Let  $\hat{k}$  be the minimizer of  $\|\mathbf{r}_k + \text{tr}[\mathbf{A}_k]\boldsymbol{\psi}_k\|^2$  as defined in the statement  
481 of Corollary 4.3 and let  $\tilde{k} \in \{1, \dots, K\}$  be such that  $\|\mathbf{h}_{\tilde{k}}\|^2 \geq \|\mathbf{h}_{\hat{k}}\|^2 + \eta$  in the event  $\tilde{\Omega}$  where  
482 such  $\tilde{k}$  exists. Then by the triangle inequality,  $\eta \mathbb{E}[I_\Omega I_{\tilde{\Omega}}] \leq C_{35}(\gamma, \mu, M)n^{-q/(1+q)}$ . It follows that  
483  $\mathbb{P}(\tilde{\Omega}) \leq \eta^{-1}C_{36}(\gamma, \mu, M)n^{-q/(1+q)} + \mathbb{P}(\Omega^c) \rightarrow 0$  as desired.  $\square$

484 *Proof of Theorem 5.3.* Using  $\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2 = (\mathbf{a} - \mathbf{b})^\top (\mathbf{a} + \mathbf{b})$  we have

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

485 Hence using  $|\text{tr}[\mathbf{A}]| \leq \gamma/\mu$ ,  $|\hat{\text{df}}| \leq n$  and the Cauchy-Schwarz inequality

$$\begin{aligned} & | \|\mathbf{G}\mathbf{h} - \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2 - \|\mathbf{G}\mathbf{h} - (\hat{\text{df}}/\text{tr}[\mathbf{V}])\|^2 | \\ & \leq C_{37}(\gamma, \mu)(\frac{n}{\text{tr}[\mathbf{V}]} \vee 1)|\hat{\text{df}}/n - \text{tr}[\mathbf{V}]| \text{tr}[\mathbf{A}]/n(\|\boldsymbol{\psi}\|^2 + \|\mathbf{G}\|_{op}\|\mathbf{h}\|^2). \end{aligned}$$



Let  $\Omega$  be the event in Corollary 4.2. Using the bound on the operator norm of  $\mathbf{G}$  in  $\Omega$ , for any deterministic  $\eta > 0$  we have proved

$$\mathbb{E}\left[I\{\Omega\}I\{\text{tr}[\mathbf{V}]n \geq \eta\} \frac{|\|\mathbf{G}\mathbf{h} - \text{tr}[\mathbf{A}]\boldsymbol{\psi}\|^2 - \|\mathbf{G}\mathbf{h} - (\hat{\text{df}}/\text{tr}[\mathbf{V}])\|^2|}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n}\right] \leq \frac{C_{38}(\gamma, \mu)}{\eta \wedge 1} n^{1/2}$$

thanks to Theorem 5.1. By (56), in the event  $\Omega$  where the operator norm of  $\|n^{-1/2}\mathbf{G}\|_{op}$  is bounded by a constant,  $\text{tr}[\mathbf{V}] \geq \text{tr}[\text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\}]/C_{39}(\gamma, \mu)$ . Hence combining the previous display with (40), we have proved

$$\mathbb{E}\left[\frac{I\{\Omega\}I\{\sum_{i=1}^n \psi'(r_i) \geq n\eta\} \|\mathbf{h}\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\mathbf{r} + \frac{\hat{\text{df}}}{\text{tr}[\mathbf{V}]} \boldsymbol{\psi}\|^2/n}{\max\{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n, (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{1/2}(\|\boldsymbol{\varepsilon}\|^2/n)^{1/2}\}}\right] \leq \frac{C_{40}(\gamma, \mu, \eta)}{\sqrt{n}}.$$

At this point the proof is similar to that of Corollary 4.3: We perform the change of variable (20) for each  $k = 1, \dots, K$  giving  $\mathbf{h}_k, \boldsymbol{\psi}_k, \hat{\text{df}}_k$  and  $\mathbf{V}_k$ . We have  $I_\Omega \max_{k=1, \dots, K} (\|\mathbf{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \leq C_{41}(\gamma, \mu, M)$  as explained in the previous proofs. Summing over  $k = 1, \dots, K$  the previous display, using  $I_\Omega \max_{k=1, \dots, K} (\|\mathbf{h}_k\|^2 + \|\boldsymbol{\psi}_k\|^2/n) \leq C_{42}(\gamma, \mu, M)$  and  $I_\Omega \|\boldsymbol{\varepsilon}\|^2 \leq n^{2/(1+q)}$  we find

$$\mathbb{E}\left[\sum_{k=1}^K I\{\Omega\}I\{\sum_{i=1}^n \psi'_k(r_{ki}) \geq n\eta\} \|\mathbf{h}_k\|^2 + \|\boldsymbol{\varepsilon}\|^2/n - \|\mathbf{r}_k + \frac{\hat{\text{df}}_k}{\text{tr}[\mathbf{V}_k]} \boldsymbol{\psi}_k\|^2/n\right] \leq \frac{KC_{43}(\gamma, \mu, \eta)}{n^{q/(1+q)}}.$$

Let  $\tilde{\Omega}$  be the event that there exists  $\tilde{k}$  with  $\frac{1}{n} \sum_{i=1}^n \psi'_k(r_{\tilde{k}i}) \geq \eta$  satisfying  $\|\mathbf{h}_{\tilde{k}}\|^2 + \tilde{\eta} \leq \|\mathbf{h}_{\hat{k}}\|^2$ , then by the previous display and the triangle inequality, using  $\|\mathbf{r}_{\tilde{k}} + \frac{\hat{\text{df}}_{\tilde{k}}}{\text{tr}[\mathbf{V}_{\tilde{k}}]} \boldsymbol{\psi}_{\tilde{k}}\|^2 \leq \|\mathbf{r}_{\tilde{k}} + \frac{\hat{\text{df}}_{\tilde{k}}}{\text{tr}[\mathbf{V}_{\tilde{k}}]} \boldsymbol{\psi}_{\tilde{k}}\|^2$  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  $\mathbb{P}(\Omega) \rightarrow 1$ , the probability  $\mathbb{P}(\tilde{\Omega})$  converge to 0 if  $K = o(n^{q/(1+q)})$ .  $\square$

## 8 Probabilistic results and their proofs

**Proposition 7.1.** [Variant of [5]] Let  $\mathbf{z} \in N(\mathbf{0}, \mathbf{I}_q)$  and  $\mathbf{f} := \mathbf{f}(\mathbf{z}) : \mathbb{R}^q \rightarrow \mathbb{R}^q \setminus \{\mathbf{0}\}$  be locally Lipschitz in  $\mathbf{z}$  with  $\mathbb{E}[\|\mathbf{f}\|^{-2} \sum_{k=1}^q \|\frac{\partial \mathbf{f}}{\partial z_k}\|^2] < +\infty$ . Then

$$\mathbb{E}\left[\left(\frac{\mathbf{f}^\top \mathbf{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\mathbf{f}\|_2} - Z\right)^2\right] \leq (7 + 2\sqrt{6}) \mathbb{E}\left[\|\mathbf{f}\|^{-2} \sum_{k=1}^q \left\|\frac{\partial \mathbf{f}}{\partial z_k}\right\|^2\right] < +\infty. \quad (28)$$

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

$$Z = \mathbf{z}^\top \mathbb{E}\left[\frac{\mathbf{f}(\mathbf{z})}{\|\mathbf{f}(\mathbf{z})\|}\right] / \sqrt{V}, \quad V = \left\|\mathbb{E}\left[\frac{\mathbf{f}(\mathbf{z})}{\|\mathbf{f}(\mathbf{z})\|}\right]\right\|^2$$

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

$$\mathbb{E}\left[\left(\frac{\mathbf{f}^\top \mathbf{z} - \sum_{k=1}^q (\partial/\partial z_k) f_k}{\|\mathbf{f}\|_2} - \sqrt{V}Z\right)^2\right] \leq 6 \mathbb{E}\left[\|\mathbf{f}\|^{-2} \sum_{k=1}^q \left\|\frac{\partial \mathbf{f}}{\partial z_k}\right\|^2\right]. \quad (42)$$

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

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

This implies that almost surely,

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

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

By the triangle inequality and  $(a + b)^2 \leq 2a^2 + 2b^2$ , this implies that the left-hand side of (42) is bounded from above by  $2\mathbb{E}[(\mathbf{z}^T \mathbf{g} - \text{tr}[\partial \mathbf{g} / \partial \mathbf{z}])^2] + 2\mathbb{E}[\|\mathbf{f}\|^{-2} \|\partial \mathbf{f} / \partial \mathbf{z}\|_F^2]$ . The first term can be bounded using the main result of [4] and the Gaussian Poincaré inequality [6, Theorem 3.20]

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

This proves (42). To bound  $|\sqrt{V} - 1|$ , we have by the triangle inequality

$$|\sqrt{V} - 1| = |\sqrt{V} - \|\frac{\mathbf{f}}{\|\mathbf{f}\|}\| \leq \|\mathbb{E}[\frac{\mathbf{f}}{\|\mathbf{f}\|}] - \frac{\mathbf{f}}{\|\mathbf{f}\|}\| = \|\mathbf{g}\|.$$

By another application of the Gaussian Poincaré inequality,

$$|\sqrt{V} - 1|^2 \leq \mathbb{E}[\|\mathbf{g}\|_2^2] \leq \mathbb{E}[\|\partial \mathbf{g} / \partial \mathbf{z}\|_F^2] \leq \mathbb{E}[\|\mathbf{f}\|^{-2} \|\partial \mathbf{f} / \partial \mathbf{z}\|_F^2]. \quad (43)$$

Combining Equations (42) and (43) using  $(a + b)^2 = a^2 + 2ab + b^2 \leq a^2 + 1/\sqrt{6}a^2 + \sqrt{6}b^2 + b^2$ , we obtain the constant  $7 + 2\sqrt{6}$ .

□

**Proposition 7.2.** Let  $\mathbf{h} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$ ,  $\boldsymbol{\psi} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$  be locally Lipschitz functions. If  $\mathbf{G} \in \mathbb{R}^{n \times p}$  has iid  $N(0, 1)$  entries then

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{\boldsymbol{\psi}^\top \mathbf{G} \mathbf{h} - \sum_{ij} \frac{\partial(\psi_i h_j)}{g_{ij}}}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right)^2 + \left( \frac{\|\mathbf{G} \mathbf{h}\|^2 - \sum_{ij} \frac{\partial(h_j \mathbf{e}_i^\top \mathbf{G} \mathbf{h})}{g_{ij}}}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right)^2 + \left( \frac{\|\mathbf{G}^\top \boldsymbol{\psi}\|^2 - \sum_{ij} \frac{\partial(\psi_i \mathbf{e}_j^\top \mathbf{G}^\top \boldsymbol{\psi})}{g_{ij}}}{n\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2} \right)^2 \right] \\ & \leq C_{44} \mathbb{E} \left[ n + p + \|\mathbf{G}\|_{op}^2 + (n + p) \sum_{i=1}^n \sum_{j=1}^p \frac{1 + \|\mathbf{G}\|_{op}^2/n}{(\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^2} \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2 \right) \right] \quad (32) \end{aligned}$$

for some positive absolute constant in the second line.

*Proof of Proposition 7.2.* We prove the claim separately for the three terms in the left-hand side of Proposition 7.2; we start with the first of the three terms. We will apply the probabilistic result given in Proposition 6.3 in [3]: if  $\boldsymbol{\eta} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$  and  $\boldsymbol{\rho} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$  are locally Lipschitz and  $\mathbf{G} \in \mathbb{R}^{n \times p}$  has iid  $N(0, 1)$  entries,

$$\mathbb{E} \left[ \left( \boldsymbol{\rho}^\top \mathbf{G} \boldsymbol{\eta} - \sum_{ij} \frac{\partial(\rho_i \eta_j)}{g_{ij}} \right)^2 \right] \leq \mathbb{E} [\|\boldsymbol{\rho}\|^2 \|\boldsymbol{\eta}\|^2] + 2\mathbb{E} \left[ \sum_{ij} \|\boldsymbol{\eta}\|^2 \left\| \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}} \right\|^2 + \|\boldsymbol{\rho}\|^2 \left\| \frac{\partial \boldsymbol{\eta}}{\partial g_{ij}} \right\|^2 \right]. \quad (44)$$

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

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

We use this inequality applied with

$$\mathbf{f} = (\mathbf{h}, \frac{1}{\sqrt{n}} \boldsymbol{\psi}), \quad \boldsymbol{\rho} = \frac{1}{\sqrt{n}} \frac{\boldsymbol{\psi}}{\|\mathbf{f}\|}, \quad \boldsymbol{\eta} = \frac{\mathbf{h}}{\|\mathbf{f}\|}. \quad (45)$$

To bound from above the right-hand side of (44), the inequality in the previous display can be rewritten

$$\left\| \frac{\partial \boldsymbol{\eta}}{\partial g_{ij}} \right\|^2 + \left\| \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}} \right\|^2 \leq \frac{1}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2 \right). \quad (46)$$

Since  $\|\boldsymbol{\rho}\| \leq 1$  and  $\|\boldsymbol{\eta}\| \leq 1$  by definition, the right-hand side of (44) is bounded from above by  $1 + 2\mathbb{E} \left[ \frac{1}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2 \right) \right]$ . Thus the proof of Proposition 7.2 for the first term in the left-hand side is almost complete; it remains to control inside the parenthesis of the left-hand side,

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

By multiple applications of the Cauchy-Schwartz inequality, the absolute value of the previous display is bounded from above by  $2(\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-1/2}(\sum_{ij} \|\frac{\partial \mathbf{h}}{\partial g_{ij}}\| + \frac{1}{n} \|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|)^{1/2}$ . This completes the proof of Proposition 7.2 for the first term in the left-hand side.

For the second and third term in the left-hand side of Proposition 7.2, apply instead (44) to  $\boldsymbol{\rho} = \mathbf{G}\boldsymbol{\eta}$  and  $\boldsymbol{\eta} = \mathbf{G}^\top \boldsymbol{\rho}$  to obtain

$$\begin{aligned} \mathbb{E} \left[ \left( \|\mathbf{G}\boldsymbol{\eta}\|^2 - \sum_{ij} \frac{\partial(\eta_j \mathbf{e}_i^\top \mathbf{G}\boldsymbol{\eta})}{g_{ij}} \right)^2 \right] &\leq \mathbb{E} [\|\mathbf{G}\boldsymbol{\eta}\|^2 \|\boldsymbol{\eta}\|^2] + 2\mathbb{E} \left[ \sum_{ij} \|\boldsymbol{\eta}\|^2 \|\mathbf{e}_i \eta_j + \mathbf{G} \frac{\partial \boldsymbol{\eta}}{\partial g_{ij}}\|^2 + \|\mathbf{G}\boldsymbol{\eta}\|^2 \left\| \frac{\partial \boldsymbol{\eta}}{\partial g_{ij}} \right\|^2 \right], \\ \mathbb{E} \left[ \left( \|\mathbf{G}^\top \boldsymbol{\rho}\|^2 - \sum_{ij} \frac{\partial(\rho_i \boldsymbol{\rho}^\top \mathbf{G} \mathbf{e}_j)}{g_{ij}} \right)^2 \right] &\leq \mathbb{E} [\|\mathbf{G}^\top \boldsymbol{\rho}\|^2 \|\boldsymbol{\rho}\|^2] + 2\mathbb{E} \left[ \sum_{ij} \|\mathbf{G}^\top \boldsymbol{\rho}\|^2 \left\| \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}} \right\|^2 + \|\boldsymbol{\rho}\|^2 \|\mathbf{e}_j \rho_i + \mathbf{G}^\top \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}}\|^2 \right]. \end{aligned}$$

Setting  $\boldsymbol{\rho} = \frac{1}{\sqrt{n}} \boldsymbol{\psi} / \|\mathbf{f}\|$ ,  $\boldsymbol{\eta} = \mathbf{h} / \|\mathbf{f}\|$  we obtain the claim in Equation (44) by bounding the right-hand side of the previous displays using the operator norm of  $\mathbf{G}$  and arguments similar to (46). The term involving  $\frac{\partial}{\partial g_{ij}} \left( \frac{1}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right)$  in the left-hand side is controlled similarly to the previous paragraph.

□

**Proposition 7.4.** Let  $\mathbf{h} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^p$ ,  $\boldsymbol{\psi} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$  be locally Lipschitz functions. If  $\mathbf{G} \in \mathbb{R}^{n \times p}$  has iid  $N(0, 1)$  entries then

$$\begin{aligned} &\mathbb{E} \left[ \frac{\left| \frac{p}{n} \|\boldsymbol{\psi}\|^2 - \frac{1}{n} \sum_{j=1}^p (\boldsymbol{\psi}^\top \mathbf{G} \mathbf{e}_j - \sum_{i=1}^n \frac{\partial \psi_i}{\partial g_{ij}})^2 \right|}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right] + \mathbb{E} \left[ \frac{\left| n \|\mathbf{h}\|^2 - \sum_{i=1}^n (\mathbf{g}_i^\top \mathbf{h} - \sum_{j=1}^p \frac{\partial h_j}{\partial g_{ij}})^2 \right|}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \right] \\ &\leq C_{45} \left( \sqrt{n+p} (1 + \Xi^{1/2}) + \Xi \right) \text{ where } \Xi = \mathbb{E} \left[ \frac{1}{\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n} \sum_{i=1}^n \sum_{j=1}^p \left( \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2 \right) \right]. \end{aligned}$$

*Proof of Proposition 7.4.* We first focus on the first term in the left-hand side. Theorem 7.1 in [3] provides that of  $\boldsymbol{\rho} : \mathbb{R}^{n \times p}$  is locally Lipschitz with  $\|\boldsymbol{\rho}\| \leq 1$  then

$$\mathbb{E} \left| p \|\boldsymbol{\rho}\|^2 - \sum_{j=1}^p \left( \boldsymbol{\rho}^\top \mathbf{G} \mathbf{e}_j - \sum_{i=1}^n \frac{\partial \rho_i}{\partial g_{ij}} \right)^2 \right| \leq C_{46} \sqrt{p} \left( 1 + \sum_{ij} \left\| \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}} \right\|^2 \right)^{1/2} + C_{47} \sum_{ij} \left\| \frac{\partial \boldsymbol{\rho}}{\partial g_{ij}} \right\|^2. \quad (47)$$

Let  $\boldsymbol{\rho} = n^{-1/2} \boldsymbol{\psi} / \|\mathbf{f}\|$  as in (45). Inequality (46) lets us bound from above the right-hand side of the previous display by the right-hand side of Proposition 7.4. In the left-hand side,  $p \|\boldsymbol{\rho}\|^2 = \frac{p}{n} \|\boldsymbol{\psi}\|^2 / (\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)$  as desired. For the left-hand side, using some algebra in [3, Section 7], for any random vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$  by the triangle and Cauchy-Schwarz inequalities we have

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

so that  $|p \|\boldsymbol{\rho}\|^2 - \|\mathbf{a}\|^2| \leq \frac{3}{2} |p \|\boldsymbol{\rho}\|^2 - \|\mathbf{b}\|^2| + 3\|\mathbf{a} - \mathbf{b}\|^2 + 2\|\mathbf{a} - \mathbf{b}\| \sqrt{p \|\boldsymbol{\rho}\|^2}$ . Applying this to  $b_j = \boldsymbol{\rho}^\top \mathbf{G} \mathbf{e}_j - \sum_{i=1}^n \frac{\partial \rho_i}{\partial g_{ij}}$  we use (47) to bound  $|p \|\boldsymbol{\rho}\|^2 - \|\mathbf{b}\|^2|$  and  $\|\boldsymbol{\rho}\| \leq 1$  to bound  $\sqrt{p \|\boldsymbol{\rho}\|^2} \leq \sqrt{p}$ . It remains to specify  $\mathbf{a}$  so that  $|p \|\boldsymbol{\rho}\|^2 - \|\mathbf{a}\|^2|$  coincides with the first term in the left-hand side of Proposition 7.4 and bound  $\|\mathbf{a} - \mathbf{b}\|$ . Consequently, we set

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

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

$$\sum_{ij} \left( \frac{\partial(D^{-1})}{\partial g_{ij}} \right)^2 = \frac{1}{D^6} \sum_{ij} \left( \mathbf{h}^\top \frac{\partial \mathbf{h}}{\partial g_{ij}} + \frac{\boldsymbol{\psi}^\top}{\sqrt{n}} \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right)^2 \leq \frac{2}{D^4} \sum_{ij} \left\| \frac{\partial \mathbf{h}}{\partial g_{ij}} \right\|^2 + \frac{1}{n} \left\| \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} \right\|^2. \quad (48)$$

543 using again the Cauchy-Schwarz inequality and  $\max\{\|\mathbf{h}\|^2, \|\boldsymbol{\psi}\|^2/n\} \leq D^2$ . We obtain  $\|\mathbf{a} - \mathbf{b}\|^2 \leq$   
544  $D^{-2} \sum_{ij} \|\frac{\partial \mathbf{h}}{\partial g_{ij}}\|^2 + \frac{1}{n} \|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2$  which completes the proof for the first term in the left-hand side of  
545 Proposition 7.4. For the second term in the left-hand side, the proof is similar with by exchanging the  
546 role of  $n$  and  $p$  in (47) and applying (47) to  $\mathbf{h}/D$  instead of  $\boldsymbol{\psi}/(\sqrt{n}D)$ .  $\square$

547 **Proposition 7.6.** *Let Assumption 1.1 be fulfilled. Then*

$$\mathbb{E}[\{(\|\mathbf{h}\|^2 + \|\boldsymbol{\psi}\|^2/n)^{-\frac{1}{2}} \|\boldsymbol{\varepsilon}\|^{-1} \xi_{VI}\}^2] \leq C_{48}(\gamma, \mu) \quad \text{for} \quad \xi_{VI} = \boldsymbol{\varepsilon}^\top (\mathbf{G}\mathbf{h} - \text{tr}[\mathbf{A}]\boldsymbol{\psi}). \quad (39)$$

548 *Proof of Proposition 7.6.* Apply (44) with  $\boldsymbol{\rho} = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$  and  $\boldsymbol{\eta} = \mathbf{h}/D$  where  $D = (\|\mathbf{h}\|^2 +$   
549  $\|\boldsymbol{\psi}\|^2/n)^{1/2}$  as in the previous proof (this scalar  $D$  is not related to the diagonal matrix  $\mathbf{D} =$   
550  $\text{diag}\{\psi'(\mathbf{r})\}$ ). Since  $\boldsymbol{\varepsilon}$  has 0 derivative with respect to  $\mathbf{G}$  we find

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

551 The right-hand side is bounded from above by  $C_{49}(\gamma, \mu)$  thanks to (46) and (36). For the second term  
552 above we use product rule and (23),

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

553 To complete the proof we need to bound from above the expectation of the square of the second  
554 and third terms colored in purple are bounded by  $C_{50}(\gamma, \mu)$ . Since  $\|\mathbf{h}\| \leq D$ , the second term is  
555 bounded from above by  $\|\mathbf{A}\|_{op} \|\mathbf{G}\|_{op}$  since  $|\psi'| \leq 1$  and  $\mathbb{E}[\|\mathbf{A}\|_{op}^2 \|\mathbf{G}\|_{op}^2] \leq C_{51}(\gamma, \mu)$  thanks to  
556  $\|\mathbf{A}\|_{op} \leq 1/(n\mu)$  and [10, Theorem II.13]. For the third term, we use the Cauchy-Schwarz inequality  
557  $(\sum_{ij} \frac{\varepsilon_i h_j}{\|\boldsymbol{\varepsilon}\|})^2 \leq \|\mathbf{h}\|^2 \sum_{ij} (\frac{\partial(D^{-1})}{\partial g_{ij}})^2$ , (48) and (36).  $\square$

## 558 9 Proof of differentiability results

559 **Theorem 2.1.** *Let Assumption 1.1 be fulfilled. For almost every  $(\mathbf{y}, \mathbf{X})$  the map  $(\mathbf{y}, \mathbf{X}) \mapsto \hat{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{X})$*   
560 *is differentiable at  $(\mathbf{y}, \mathbf{X})$  and there exists a matrix  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  with  $\|\boldsymbol{\Sigma}^{1/2} \hat{\mathbf{A}} \boldsymbol{\Sigma}^{1/2}\|_{op} \leq (n\mu)^{-1}$  s.t.*

$$\begin{aligned} (\partial/\partial y_i) \hat{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{X}) &= \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i \psi'(r_i), \\ (\partial/\partial x_{ij}) \hat{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{X}) &= \hat{\mathbf{A}} \mathbf{e}_j \psi(r_i) - \hat{\mathbf{A}} \mathbf{X}^\top \mathbf{e}_i \psi'(r_i) \hat{\boldsymbol{\beta}}_j, \end{aligned} \quad \text{where } r_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \quad (5)$$

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

$$\begin{aligned} \text{df} &= \text{tr}[\mathbf{X}(\partial/\partial \mathbf{y}) \hat{\boldsymbol{\beta}}] = \text{tr}[\mathbf{X} \hat{\mathbf{A}} \mathbf{X} \text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\}], \\ \mathbf{V} &= \text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\}(\mathbf{I}_n - \mathbf{X}(\partial/\partial \mathbf{y}) \hat{\boldsymbol{\beta}}) = \text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\} - \text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\} \mathbf{X} \hat{\mathbf{A}} \mathbf{X} \text{diag}\{\boldsymbol{\psi}'(\mathbf{r})\}. \end{aligned} \quad (6) \quad (7)$$

562 satisfy  $0 \leq \text{df} \leq n$  and  $0 \leq \text{tr}[\mathbf{V}] \leq n$ .

563 The first part of the following proof is similar to the argument using the KKT conditions in [3]. After  
564 (51), the argument is novel and lets us derive the convenient formula (5) and the existence of matrix  
565  $\hat{\mathbf{A}}$  which plays a central role in the contractions (23)-(27).

566 *Proof of Theorem 2.1.*  $\mathbf{X}_t = \mathbf{X} + t\mathbf{U}$  and  $\mathbf{y}_t = \mathbf{y} + t\mathbf{v}$  with  $t \in \mathbb{R}$  where  $\mathbf{U} \in \mathbb{R}^{n \times p}$  and  
567  $\mathbf{v} \in \mathbb{R}^n$  are fixed. Let  $\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}(\mathbf{y}_t, \mathbf{X}_t)$  and  $\hat{\mathbf{r}}_t = \mathbf{y}_t - \mathbf{X}_t \hat{\boldsymbol{\beta}}_t$  and  $\hat{\boldsymbol{\psi}}(\mathbf{y}_t, \mathbf{X}_t) = \boldsymbol{\psi}(\hat{\mathbf{r}}_t)$ .  
568 By convention, without arguments  $\hat{\boldsymbol{\beta}}, \boldsymbol{\psi}$  refer to  $(\mathbf{y}, \mathbf{X})$  which is  $(\mathbf{y}_t, \mathbf{X}_t)$  at  $t = 0$ . By the KKT  
569 conditions,  $\mathbf{X}^\top \hat{\boldsymbol{\psi}} \in n \text{dg}(\hat{\boldsymbol{\beta}})$  and  $\mathbf{X}_t^\top \hat{\boldsymbol{\psi}}_t \in n \text{dg}(\hat{\boldsymbol{\beta}}_t)$ , by strong convexity of  $g$ , we have

$$n\mu \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}})\|^2 \leq (\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}})^\top (\mathbf{X}_t^\top \hat{\boldsymbol{\psi}}_t - \mathbf{X}^\top \hat{\boldsymbol{\psi}}). \quad (49)$$

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

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

572 Adding up the above two displays we have

$$n\mu\|\Sigma^{1/2}(\hat{\beta}_t - \hat{\beta})\|^2 + \|\hat{\psi}_t - \hat{\psi}\|^2 \leq (\hat{\beta}_t - \hat{\beta})^\top (\mathbf{X}_t^\top \hat{\psi}_t - \mathbf{X}^\top \hat{\psi}) + (\hat{\psi}_t - \hat{\psi})^\top (\hat{\mathbf{r}}_t - \hat{\mathbf{r}}). \quad (50)$$

573 By  $\mathbf{X}_t^\top \hat{\psi}_t - \mathbf{X}^\top \hat{\psi} = (\mathbf{X}_t - \mathbf{X})^\top \hat{\psi} + \mathbf{X}_t^\top (\hat{\psi}_t - \hat{\psi})$  and  $\mathbf{X}_t (\hat{\beta}_t - \hat{\beta}) + \hat{\mathbf{r}}_t - \hat{\mathbf{r}} = \mathbf{y}_t - \mathbf{y} - (\mathbf{X}_t - \mathbf{X})^\top \hat{\beta}$ ,  
574 we have

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

575 By the Cauchy-Schwartz inequality, the above implies

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

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

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

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

581 Taking the limit  $t \rightarrow 0^+$  in (49) and using the chain rule, where the derivatives exist we have

$$\begin{aligned} & n\mu\|\Sigma^{1/2}(\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U}))\|_2^2 \\ & \leq \left(\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U})\right)^\top \left(\mathbf{U}^\top \hat{\psi} + \mathbf{X}^\top \text{diag}(\hat{\psi}')(-\mathbf{U}\hat{\beta} - \mathbf{X}\frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U}) + (I_n - \mathbf{X}\frac{\partial\hat{\beta}}{\partial\mathbf{y}})\mathbf{v})\right) \\ & = \left(\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U})\right)^\top B(\mathbf{U}, \mathbf{v}) - \left\|\text{diag}(\hat{\psi}')^{\frac{1}{2}}\mathbf{X}\left(\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U})\right)\right\|_2^2 \end{aligned} \quad (51)$$

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

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

584 where  $(u_{ij})_{i=1,\dots,n,j=1,\dots,p}$  are the entries of  $\mathbf{U}$ . By the Cauchy-Schwartz inequality, (51) provides  
585 us the following two main ingredients:

$$\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U}) = 0 \text{ for all } (\mathbf{U}, \mathbf{v}) \text{ such that } B(\mathbf{U}, \mathbf{v}) = 0, \quad (52)$$

586

$$\left\|\Sigma^{1/2}\left(\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U})\right)\right\|_2 \leq \mu^{-1}n^{-1}\|\Sigma^{-1/2}B(\mathbf{U}, \mathbf{v})\|_2. \quad (53)$$

587 Since both  $\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U})$  and  $B(\mathbf{U}, \mathbf{v})$  are linear in  $(\mathbf{U}, \mathbf{v}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$  into  $\mathbb{R}^p$ , Proposition 9.1  
588 implies that there exists a matrix  $\hat{\mathbf{A}} \in \mathbb{R}^{p \times p}$  such that  $\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U}) = \hat{\mathbf{A}}B(\mathbf{U}, \mathbf{v})$  for all  $(\mathbf{U}, \mathbf{v})$ ,  
589 and by (53),  $\hat{\mathbf{A}}$  can be chosen such that  $\|\Sigma^{1/2}\hat{\mathbf{A}}\Sigma^{1/2}\|_{op} \leq (n\mu)^{-1}$  thanks to the operator norm  
590 identity in Proposition 9.1. With  $(\mathbf{U}, \mathbf{v}) = (\mathbf{e}_i \mathbf{e}_j^\top, \mathbf{0})$  for  $(i, j) \in [n] \times [p]$  and  $(\mathbf{U}, \mathbf{v}) = (\mathbf{0}, \mathbf{e}_k)$  for  
591  $k \in [n]$ , we obtain the stated formulae for  $(\partial x_{ij}/\partial)\hat{\beta}$  and  $(\partial y_k/\partial)\hat{\beta}$  in (5).

592 Now we show that both  $\text{tr}[\mathbf{V}] := \text{tr}[\mathbf{D} - \mathbf{D}\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}]$  and  $\hat{\text{df}} := \text{tr}[\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}]$  are in  $[0, n]$   
593 where  $\mathbf{D} := \text{diag}\{\psi'(\mathbf{r})\}$ . Using the symmetric part of  $\hat{\mathbf{A}}$  defined as  $\tilde{\mathbf{A}} := (\hat{\mathbf{A}} + \hat{\mathbf{A}}^\top)/2$  we have  
594  $\text{tr}[\mathbf{V}] = \text{tr}[\mathbf{D} - \mathbf{D}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top \mathbf{D}]$  and  $\hat{\text{df}} = \text{tr}[\mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top \mathbf{D}^{1/2}]$  by property of the trace. In (51),  
595 take  $\mathbf{U} = \mathbf{0}$  so that  $\frac{\partial\hat{\beta}}{\partial\mathbf{y}}\mathbf{v} + \frac{\partial\hat{\beta}}{\partial\mathbf{X}}(\mathbf{U}) = \hat{\mathbf{A}}B(\mathbf{U}, \mathbf{v}) = \hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v}$  and we have with  $\mathbf{G} = \mathbf{X}\Sigma^{-1/2}$

$$(1 + \frac{n\mu}{\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2})\|\mathbf{D}^{1/2}\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v}\|^2 \leq n\mu\|\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v}\|^2 + \|\mathbf{D}^{1/2}\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v}\|^2 \quad (54)$$

$$\leq \mathbf{v}^\top \mathbf{D}\mathbf{X}\hat{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v} = \mathbf{v}^\top \mathbf{D}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top \mathbf{D}\mathbf{v} \quad (55)$$

for all  $\mathbf{v}$ . This implies the positive semi-definite property of the symmetric matrix  $\mathbf{D}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}$ , and thus  $\hat{\mathbf{d}}\mathbf{f} \geq 0$  and  $\text{tr}[\mathbf{V}] \leq \text{tr}[\mathbf{D}] \leq n$ . With  $\tilde{\mathbf{v}} = \mathbf{D}^{1/2}\mathbf{v}$ , it also implies  $(1 + n\mu/\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2)\|\mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}^{1/2}\tilde{\mathbf{v}}\|^2 \leq \tilde{\mathbf{v}}^\top\mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}^{1/2}\tilde{\mathbf{v}}$ , which implies by the Cauchy-Schwartz inequality  $(1 + n\mu/\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2)\|\mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}^{1/2}\|_{op} \leq 1$ . The same operator norm inequality with  $\hat{\mathbf{A}}$  replaced by  $\tilde{\mathbf{A}}$  thanks to the triangle inequality. Thus  $\hat{\mathbf{d}}\mathbf{f} \leq \text{tr}[\mathbf{D}](1 + n\mu/\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2)^{-1} \leq n$  as well as

$$\begin{aligned} \text{tr}[\mathbf{V}] &= \text{tr}[\mathbf{D}^{1/2}(\mathbf{I}_n - \mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}^{1/2})\mathbf{D}^{1/2}] \geq \text{tr}[\mathbf{D}](1 - (1 + n\mu/\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2)^{-1}) \\ &= \text{tr}[\mathbf{D}]/(\|\mathbf{D}^{1/2}\mathbf{G}\|_{op}^2/(n\mu) + 1) \\ &\geq \text{tr}[\mathbf{D}]/(\|\mathbf{G}\|_{op}^2/(n\mu) + 1) \\ &\geq 0 \end{aligned} \quad (56)$$

thanks to  $\psi' \in [0, 1]$ . Inequality (55) with  $\tilde{\mathbf{v}} = \mathbf{D}^{1/2}\mathbf{v}$  and  $\mathbf{M} = \mathbf{I}_n - \mathbf{D}^{1/2}\mathbf{X}\tilde{\mathbf{A}}\mathbf{X}^\top\mathbf{D}^{1/2}$  implies  $\|(\mathbf{M} - \mathbf{I}_n)\tilde{\mathbf{v}}\|^2 \leq \tilde{\mathbf{v}}^\top(\mathbf{I}_n - \mathbf{M})\tilde{\mathbf{v}}$ . As the left-hand side is  $\|\mathbf{M}\tilde{\mathbf{v}}\|^2 - 2\tilde{\mathbf{v}}^\top\mathbf{M}\tilde{\mathbf{v}} + \|\tilde{\mathbf{v}}\|^2$ , this yields  $\|\mathbf{M}\tilde{\mathbf{v}}\|^2 \leq \tilde{\mathbf{v}}^\top\mathbf{M}\tilde{\mathbf{v}} \leq \|\tilde{\mathbf{v}}\|\|\mathbf{M}\tilde{\mathbf{v}}\|$ . If  $\tilde{\mathbf{v}}$  has unit norm and is such that  $\|\mathbf{M}\tilde{\mathbf{v}}\| = \|\mathbf{M}\|_{op}$  this gives  $\|\mathbf{M}\|_{op} \leq 1$  so that  $\|\mathbf{V}\|_{op} = \|\mathbf{D}^{1/2}\mathbf{M}\mathbf{D}^{1/2}\|_{op} \leq \|\mathbf{D}\|_{op} \leq 1$ . This gives another proof of  $\text{tr}[\mathbf{V}] \leq n$ .  $\square$

*Proof of Remark 2.2.* The proof for the intercept term included is the same to that of Theorem 2.1. The only difference is that when computing the derivatives,

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

609

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

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

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

where  $\Psi' := \text{diag}(\hat{\psi}') - \hat{\psi}'\hat{\psi}'^\top/\mathbf{1}^\top\hat{\psi}'$ . By taking limit of  $t \rightarrow 0$  in Equation (50),

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

613

$\square$

**Proposition 9.1** (A lemma on linear transformations). *Let  $\mathbf{A}$  and  $\mathbf{B}$  be two real matrices with shape  $n$  by  $p$ . Assume that  $\mathbf{B}\mathbf{v} = \mathbf{0}$  for all  $\mathbf{v}$  such that  $\mathbf{A}\mathbf{v} = \mathbf{0}$  with  $\mathbf{v} \in \mathbb{R}^p$ . Then the matrix  $\mathbf{C} := \mathbf{B}\mathbf{A}^+$  where  $\mathbf{A}^+$  is the Moore-Penrose pseudoinverse of  $\mathbf{A}$  satisfies  $\mathbf{B} = \mathbf{C}\mathbf{A}$  and  $\|\mathbf{C}\|_{op} = \max_{\mathbf{u} \in \mathbb{R}^n: \mathbf{A}\mathbf{u} \neq \mathbf{0}} \{\|\mathbf{B}\mathbf{u}\|_2/\|\mathbf{A}\mathbf{u}\|_2\}$ .*

618 *Proof.* Let  $r$  be the rank of  $\mathbf{A}$ . We let  $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$  be the SVD of  $\mathbf{A}$ , where  $\mathbf{V}$  has orthonormal  
619 columns  $\mathbf{v}_1, \dots, \mathbf{v}_p$  with the first  $r$  columns spanning the row space of  $\mathbf{A}$ , and the last  $p - r$  columns  
620 spanning the nullspace of  $\mathbf{A}$ . Let  $\mathbf{u}_i$  denote the  $i$ -th column of  $\mathbf{U}$ . Let

$$\mathbf{C} := \mathbf{B}\mathbf{A}^+ := \sum_{i \in [r]} d_i^{-1} \mathbf{B}\mathbf{v}_i \mathbf{u}_i^\top$$

621 where  $\mathbf{A}^+$  is the Moore-Penrose pseudoinverse of  $\mathbf{A}$ . Notice that  $\mathbf{A}^+ \mathbf{A} \mathbf{v} = \sum_{i \in [r]} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{v} =$   
622  $P_{\text{row}(\mathbf{A})} \mathbf{v}$  project  $\mathbf{v} \in \mathbb{R}^p$  onto the row space of  $\mathbf{A}$ . So  $\mathbf{B}\mathbf{A}^+ \mathbf{A} \mathbf{v} = \mathbf{B}\mathbf{v}$  if  $\mathbf{v} \in \text{row}(\mathbf{A})$ , and  
623  $\mathbf{B}\mathbf{A}^+ \mathbf{A} \mathbf{v} = \mathbf{0}$  if  $\mathbf{v} \in \text{Ker}(\mathbf{A})$ . By the assumption that  $\mathbf{B}\mathbf{v} = \mathbf{0}$  for all  $\mathbf{v}$  such that  $\mathbf{A}\mathbf{v} = \mathbf{0}$ , we have  
624  $\mathbf{B}\mathbf{A}^+ \mathbf{A} \mathbf{v} = \mathbf{B}\mathbf{v}$  holds for all  $\mathbf{v} \in \mathbb{R}^p = \text{row}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A})$ .

625 For  $\|\mathbf{B}\mathbf{A}^+\|_{op}$ , we notice that  $\mathbf{A}^+$  maps any  $\mathbf{u} \in \text{col}(\mathbf{A})^\perp$  to  $\mathbf{0}$ . The ratio  $\|\mathbf{B}\mathbf{A}^+ \mathbf{u}\|_2 / \|\mathbf{u}\|_2$  for  
626  $\mathbf{u} \in \mathbb{R}^n$  is maximized only when  $\mathbf{u} \in \text{col}(\mathbf{A})$ : Otherwise, we can replace  $\mathbf{u}$  with the projection of  $\mathbf{u}$   
627 onto  $\text{col}(\mathbf{A})$ , denoted by  $\mathbf{A}\mathbf{v} := P_{\text{col}(\mathbf{A})} \mathbf{u}$ , and we will have a ratio with the same numerator, but a  
628 smaller denominator and thus a larger ratio:

$$\frac{\|\mathbf{B}\mathbf{A}^+ \mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \frac{\|\mathbf{B}\mathbf{A}^+ (\mathbf{A}\mathbf{v} + \mathbf{u} - \mathbf{A}\mathbf{v})\|_2}{\|\mathbf{A}\mathbf{v} + \mathbf{u} - \mathbf{A}\mathbf{v}\|_2} \leq \frac{\|\mathbf{B}\mathbf{A}^+ \mathbf{A}\mathbf{v}\|_2}{\|\mathbf{A}\mathbf{v}\|_2} = \frac{\|\mathbf{B}\mathbf{v}\|_2}{\|\mathbf{A}\mathbf{v}\|_2}.$$

629 This implies  $\|\mathbf{B}\mathbf{A}^+\|_{op} = \max_{\mathbf{v} \in \mathbb{R}^n} \frac{\|\mathbf{B}\mathbf{v}\|_2}{\|\mathbf{A}\mathbf{v}\|_2}$ . □

630 **10 Additional Figures (anisotropic Gaussian design)**

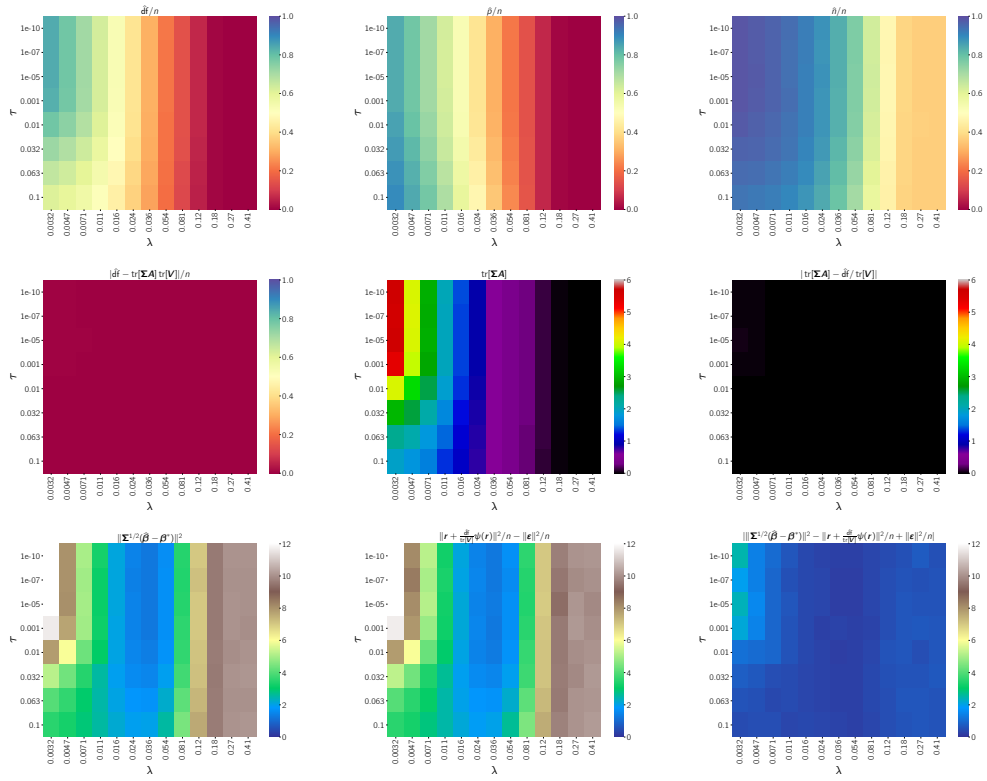


Figure 4: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with  $\Lambda = 0.054n^{1/2}$  and  $(\lambda, \tau)$  where  $\lambda \in [0.0032, 0.41]$  and  $\tau \in [10^{-10}, 0.1]$ . Each cell is the average over 100 repetitions. See the simulation setup in Section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



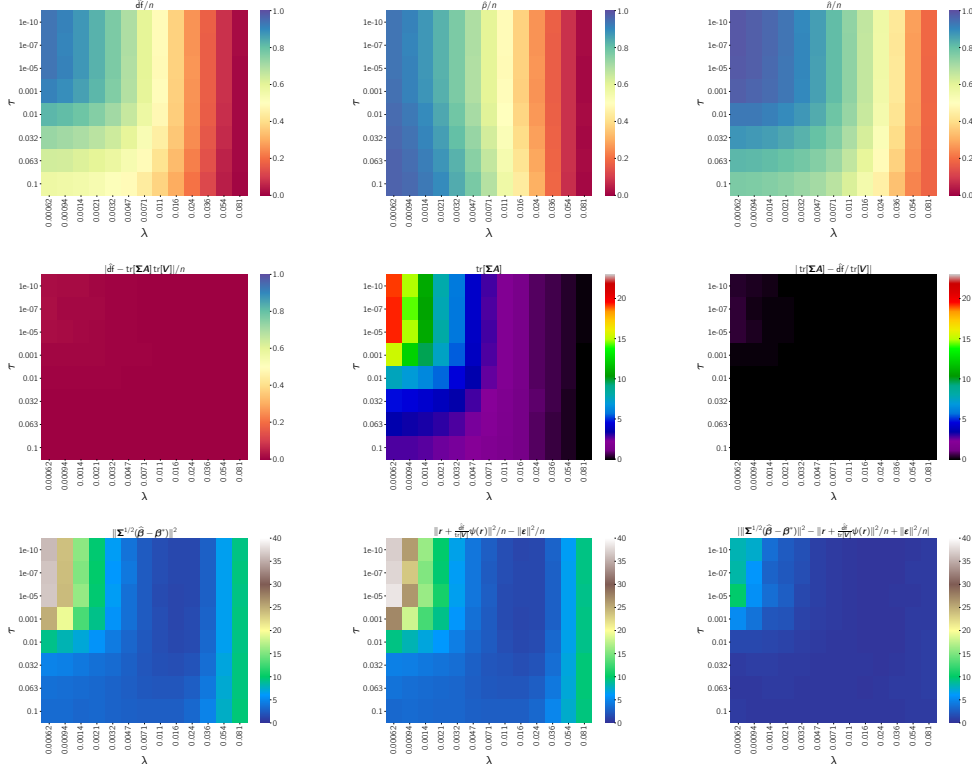


Figure 5: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with  $\Lambda = 0.024n^{1/2}$  and  $(\lambda, \tau)$  where  $\lambda \in [0.00062, 0.081]$  and  $\tau \in [10^{-10}, 0.1]$ . Each cell is the average over 50 repetitions. See the simulation setup in Section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

## 631 11 Additional Figures (non-Gaussian, Rademacher design)

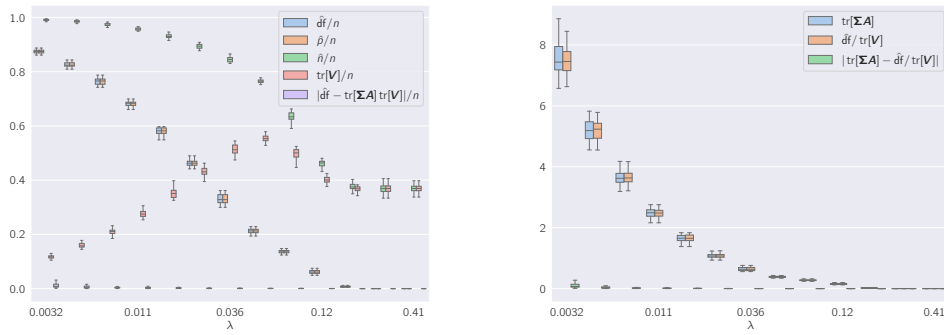


Figure 6: Boxplots for  $\hat{df}$ ,  $\hat{\beta}$ ,  $\hat{h}$ ,  $\text{tr}[\mathbf{V}]$ ,  $\text{tr}[\hat{\Sigma}\hat{\mathbf{A}}]$  and  $|\text{tr}[\hat{\Sigma}\hat{\mathbf{A}}] - \hat{df}/\text{tr}[\mathbf{V}]|$  in Huber Elastic-Net regression with  $\tau = 10^{-10}$  and  $\lambda \in [0.0032, 0.41]$ . The data are generated with  $\mathbf{X}$  having iid entries taking value  $\pm 1$  each with probability 0.5 (so that  $\Sigma = \mathbf{I}_p$ ). Each box contains 30 data points.

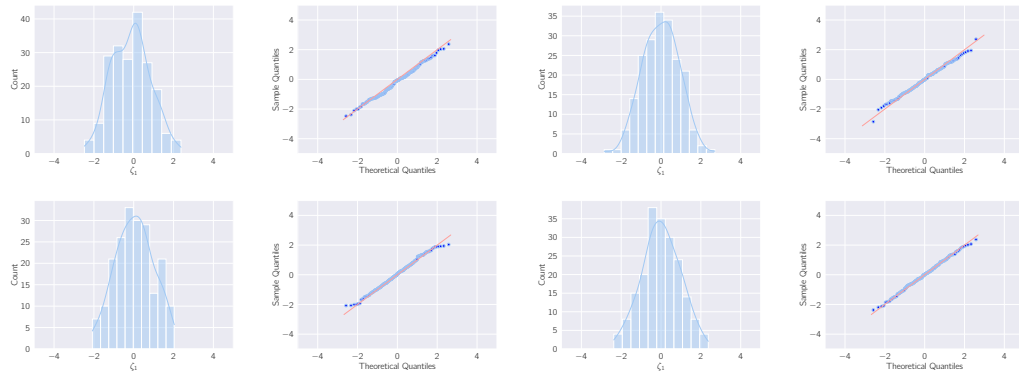


Figure 7: Histogram and QQ-plot for  $\zeta_1$  in (13) under Huber Elastic-Net regression for different choices of tuning parameters  $(\lambda, \tau)$ . Left Top:  $(0.036, 10^{-10})$ , Right Top:  $(0.054, 0.01)$ , Left Bottom:  $(0.036, 0.01)$ , Right Bottom:  $(0.024, 0.1)$ . Each figure contains 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.