

362 **Notation.** For vectors in \mathbb{R}^q or \mathbb{R}^n , the Euclidean norm is $\|\cdot\|$ and $\|\cdot\|_q$ is the ℓ_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 γ only are denoted $C_0(\gamma), C_1(\gamma), \dots$ and constant that depend on γ, μ 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 Ω , its indicator function is denoted by I_Ω 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, ψ_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 ψ 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 ε_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 ψ 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 Ξ 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 $\mathbb{E}[\|\mathbf{G}\|_{op}^2]$, 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 ξ_{IV} and ξ_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 ξ_V and ξ_{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 Ω 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 Ω , 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 ηI_Ω . 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 Ω 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 Ω be the event in Corollary 4.2. Using the bound on the operator norm of \mathbf{G} in Ω , 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 Ω 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

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

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

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

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

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

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

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

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

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

9 Proof of differentiability results

Theorem 2.1. *Let Assumption 1.1 be fulfilled. For almost every (\mathbf{y}, \mathbf{X}) the map $(\mathbf{y}, \mathbf{X}) \mapsto \hat{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{X})$ 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.*

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

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

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

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

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

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 $\mathbf{v} \in \mathbb{R}^n$ 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 $\hat{\boldsymbol{\psi}}(\mathbf{y}_t, \mathbf{X}_t) = \psi(\hat{\mathbf{r}}_t)$. 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 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

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

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

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

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 $\tilde{\mathbf{A}}$ replaced by $\hat{\mathbf{A}}$ 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

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

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)**

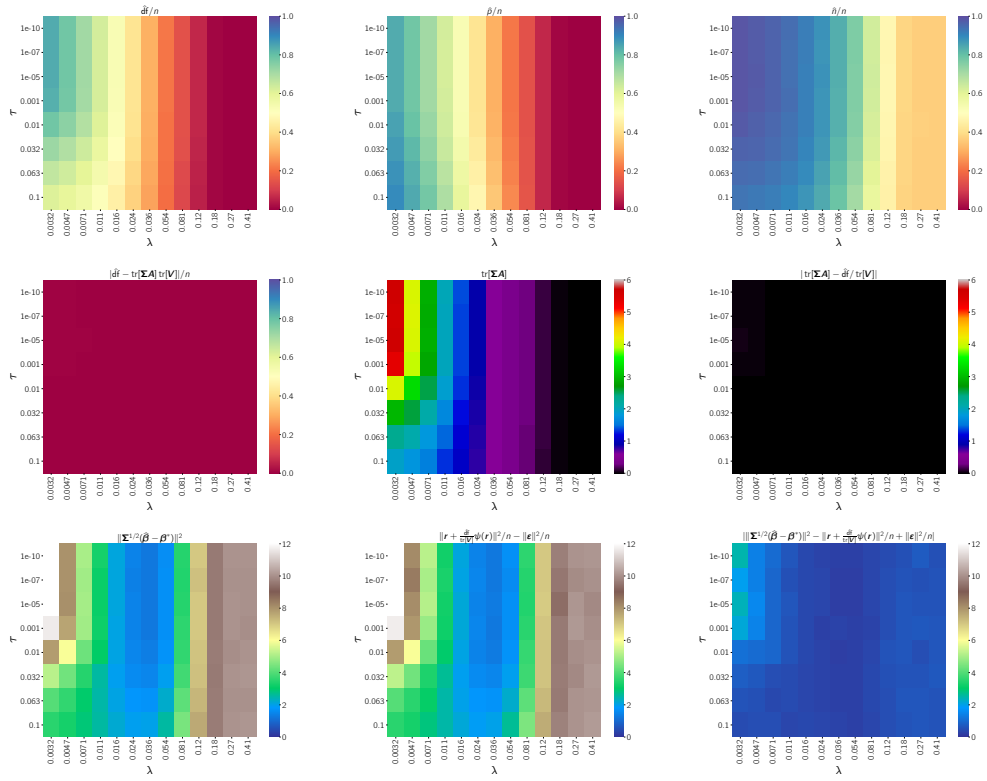


Figure 4: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.054n^{1/2}$ and (λ, τ) where $\lambda \in [0.0032, 0.41]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is the average over 100 repetitions. See the simulation setup in Section 6 in the paper for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

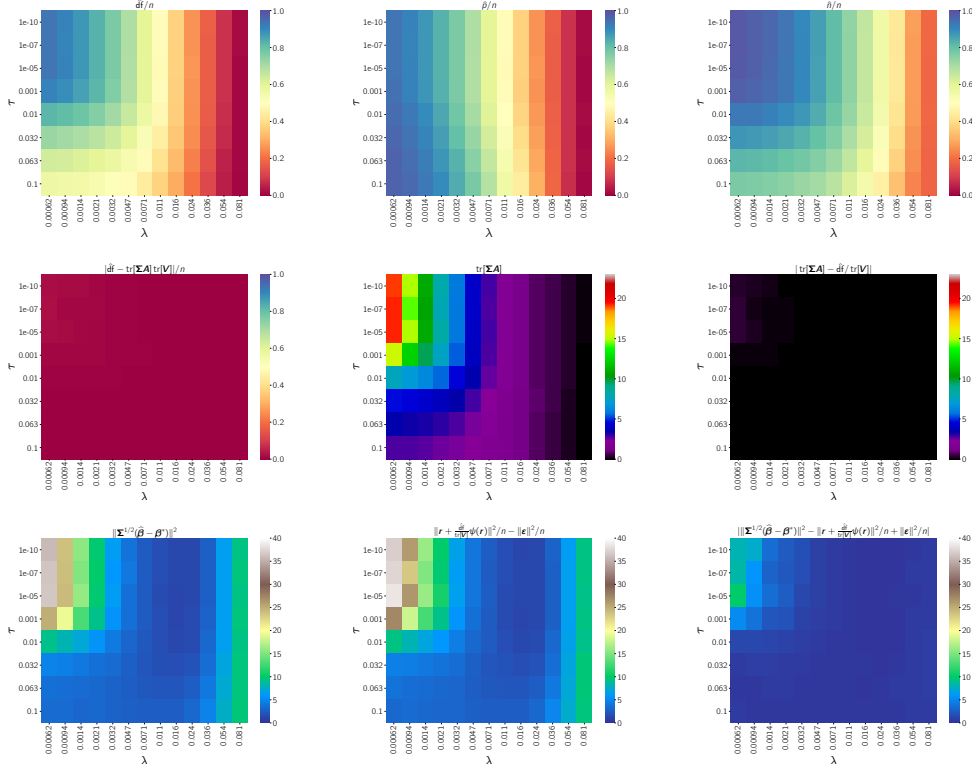


Figure 5: Heatmaps for the Huber loss and Elastic-Net penalty on a grid of tuning parameters with $\Lambda = 0.024n^{1/2}$ and (λ, τ) where $\lambda \in [0.00062, 0.081]$ and $\tau \in [10^{-10}, 0.1]$. Each cell is the average over 50 repetitions. See the simulation setup in Section 6 in the paper for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

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

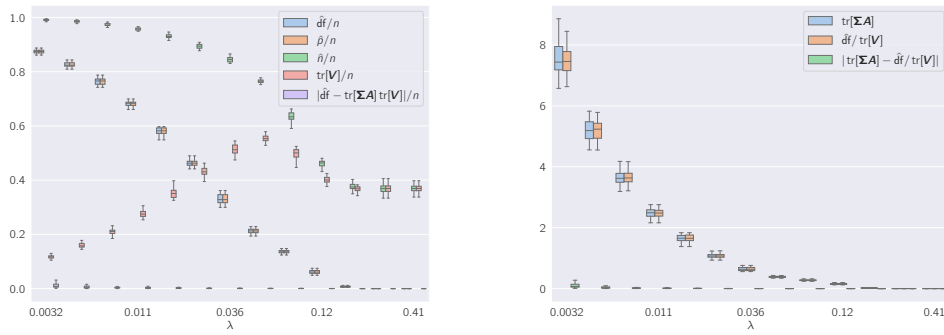


Figure 6: Boxplots for \hat{df} , $\hat{\beta}$, \hat{h} , $\text{tr}[\mathbf{V}]$, $\text{tr}[\hat{\Sigma}\hat{\mathbf{A}}]$ and $|\text{tr}[\hat{\Sigma}\hat{\mathbf{A}}] - \hat{df}/\text{tr}[\mathbf{V}]|$ in Huber Elastic-Net regression with $\tau = 10^{-10}$ and $\lambda \in [0.0032, 0.41]$. The data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

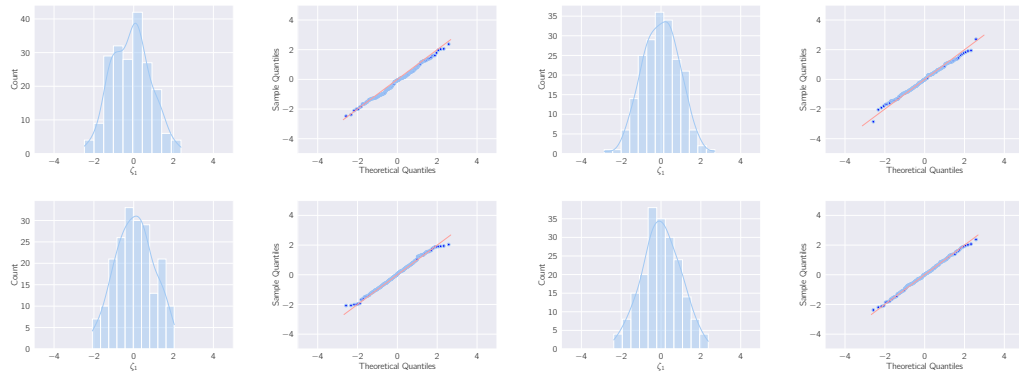


Figure 7: Histogram and QQ-plot for ζ_1 in (13) under Huber Elastic-Net regression for different choices of tuning parameters (λ, τ) . Left Top: $(0.036, 10^{-10})$, Right Top: $(0.054, 0.01)$, Left Bottom: $(0.036, 0.01)$, Right Bottom: $(0.024, 0.1)$. Each figure contains 100 data points generated with Rademacher design matrix (each entry has value ± 1 with probability 0.5) and iid ε_i from the t -distribution with 2 degrees of freedom.