# Supplement to "Precise Asymptotics of Bagging Regularized M-estimators"

Takuya Koriyama, Pratik Patil, Jin-Hong Du, Kai Tan, and Pierre C. Bellec.

This document serves as a supplement to the paper titled "Precise Asymptotics of Bagging Regularized M-estimators". Below, we provide an outline of the supplement along with a summary of the notation used in the main paper and the supplement.

TABLE 4
*Outline of the supplement.*

| Section | Content |
|---|---|
| Appendix A | Sign of solutions $(\eta_H, \eta_G)$ to System 1b |
| Appendix B | Proofs of Theorems 1, 2 and 5 from Section 3 |
| Appendix C | Proof of Proposition 7, 8 and other miscellaneous details from Section 4 |
| Appendix D | Details of arguments in Remark 7 from Section 5 |
| Appendix E | Proof of Theorem 9 and Conjecture 10 from Section 6 |
| Appendix F | Additional numerical illustrations and experimental details |

TABLE 5
*Summary of some of the specific notation used in the paper.*

| Notation | Description |
|---|---|
| $(\boldsymbol{x}_i, y_i)$, $i \in [n]$ | Observation vector with the feature vector in $\mathbb{R}^p$ and the response variable in $\mathbb{R}$ |
| $\boldsymbol{X}, \boldsymbol{y}$ | Feature matrix in $\mathbb{R}^{n \times p}$ and the response vector in $\mathbb{R}^n$ |
| $\boldsymbol{\theta}$ | Signal vector in the linear model in $\mathbb{R}^p$ with entries drawn i.i.d. from the distribution $F_\theta$ |
| $\boldsymbol{z}$ | Noise vector in the linear model in $\mathbb{R}^n$ with entries drawn i.i.d. from the distribution $F_z$ |
| $k$, $M$ | Size of each subsample and the total number of subsamples (bags) used in the ensemble |
| $I_m$, $m \in [M]$ | Index set (subset of $[n]$) of the $m$-th subsample, with $\lvert I_m \rvert = k$ |
| $(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$, $m \in [M]$ | Feature matrix in $\mathbb{R}^{k \times p}$ and response vector in $\mathbb{R}^k$ of the subsampled dataset indexed by $I_m$ |
| $(\boldsymbol{X}_{I_m \cap I_\ell}, \boldsymbol{y}_{I_m \cap I_\ell})$, $m \neq \ell \in [M]$ | Feature matrix in $\mathbb{R}^{\lvert I_m \cap I_\ell \rvert \times p}$ and response vector in $\mathbb{R}^{\lvert I_m \cap I_\ell \rvert}$ of the overlapped dataset corresponding to the overlap between the subsampled datasets indexed by $I_m$ and $I_\ell$ |
| loss, reg, $\lambda$ | Loss function, regularization function, and regularization level used for the regularized M-estimator |
| $\widehat{\boldsymbol{\theta}}_m$, $\widetilde{\boldsymbol{\theta}}_M$ | Component regularized M-estimator fitted on the $m$-th subsample and the ensemble estimator |
| $R_M$, $\mathcal{R}_M$ | Squared prediction risk of the ensemble estimator and its asymptotic limit |
| $\mathcal{R}_1$, $\mathcal{R}_\infty$ | Asymptotic risk of the non-ensemble estimator ($M = 1$) and the full-ensemble estimator ($M = \infty$) |
| $\delta$, $c$ | Inverse data aspect ratio $n/p$ and subsample ratio $k/n$ |
| $(\alpha, \beta, \kappa, \nu)$ | Parameters characterizing the ensemble risk asymptotics with $M = 1$ (System 1a) |
| $G$, $H$ | Standard normal random variables in the ensemble risk asymptotics with $M = 1$ (System 1a) |
| $\Theta$, $Z$ | Random variables drawn according to distributions $F_\theta$ and $F_z$, respectively (System 1a) |
| $(\eta_G, \eta_H)$ | Correlation parameters in the ensemble risk asymptotics with $M = \infty$ (System 1b) |
| $G, \widetilde{G}, H, \widetilde{H}$ | Random variables appearing in the ensemble risk asymptotics with $M = \infty$ (System 1b) |
| $(a, \tau)$ | Parameters in alternate formulation of the ensemble risk asymptotics with $M = 1$ (System 2) |
| $\xi$ | Parameter in alternate formulation of the ensemble risk asymptotics with $M = \infty$ (System 2) |

## APPENDIX A: CORRELATION SIGNS

PROPOSITION 11. *The signs of the solution* $(\eta_G^\star, \eta_H^\star)$ *to System 1b are characterized as*

$$\mathrm{sign}(\eta_G^\star) = \mathrm{sign}(F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(0)), \qquad \mathrm{sign}(\eta_H^\star) = \mathrm{sign}(F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(0))$$

*where* $\mathrm{sign}(x) := \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$.

PROOF. The claim immediately follows from the fact that $\eta_G^\star$ and $\eta_H^\star$ satisfies the fixed-point equations $\eta_G - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G) = 0$ and $\eta_H - F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H) = 0$, respectively, and the maps $\eta_G \mapsto \eta_G - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_G)$ and $\eta_H \mapsto \eta_H - F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta_H)$ are nondecreasing since $F_{\mathsf{reg}} \circ F_{\mathsf{loss}}$ and $F_{\mathsf{loss}} \circ F_{\mathsf{reg}}$ are $(c \wedge \widetilde{c})$-Lipschitz with $c \wedge \widetilde{c} \le 1$ (Theorem 1). □

Recall the property of $F_{\mathsf{loss}}$ and $F_{\mathsf{reg}}$ in Theorem 1. Since $F_{\mathsf{loss}}$ and $F_{\mathsf{reg}}$ in System 1b are non-decreasing, combined with Proposition 11, we get the following simple sufficient condition which determines the sign of $(\eta_H^\star, \eta_G^\star)$:

$$F_{\mathsf{loss}}(0) \text{ and } F_{\mathsf{reg}}(0) \text{ are non-negative} \Rightarrow \eta_H^\star \text{ and } \eta_G^\star \text{ are non-negative.}$$

Here $F_{\mathsf{loss}}(0)$ and $F_{\mathsf{reg}}(0)$ can be written as follows:

$$F_{\mathsf{loss}}(0) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}\Big[\mathbb{E}\big[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \kappa) \mid Z\big] \cdot \mathbb{E}\big[\mathsf{env}'_{\widetilde{\mathsf{loss}}}(Z + \widetilde{\alpha} G; \widetilde{\kappa}) \mid Z\big]\Big],$$

$$F_{\mathsf{reg}}(0) = \frac{1}{\nu\widetilde{\nu}\alpha\widetilde{\alpha}} \cdot \mathbb{E}\Big[\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu} H; \tfrac{1}{\nu}) \mid \Theta\big] \cdot \mathbb{E}\big[\mathsf{env}'_{\widetilde{\mathsf{reg}}}(\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}} H; \tfrac{1}{\widetilde{\nu}}) \mid \Theta\big]\Big].$$

In particular, if the same loss and regularizer are used, i.e., $\mathsf{loss} = \widetilde{\mathsf{loss}}$ and $\mathsf{reg} = \widetilde{\mathsf{reg}}$, and the subsample sizes are the same, i.e., $k = \widetilde{k}$, then the solutions satisfying System 1b are same, i.e., $(\alpha, \beta, \kappa, \nu) = (\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\kappa}, \widetilde{\nu})$, so that it is easy to see from the above formula that $F_{\mathsf{loss}}(0) \ge 0$ and $F_{\mathsf{reg}}(0) \ge 0$. This means that the solutions $(\eta_G^\star, \eta_H^\star)$ are non-negative when the same loss, regularizer, and subsample size are used.

Another case such that the solutions $(\eta_G^\star, \eta_H^\star)$ are positive is when $\mathsf{loss}$ and $\widetilde{\mathsf{loss}}$ are least squares and $\mathsf{reg}$ and $\widetilde{\mathsf{reg}}$ are ridge (but possibly different regularization parameters). This is because $\mathsf{env}'_f; (x;\tau)$ is linear in $x$ for any squared loss (and regularizer) of the form $f(\cdot) = \lambda(\cdot)^2$ so that $F_{\mathsf{loss}}(0) = C_1 \mathbb{E}[Z^2] \ge 0$ and $F_{\mathsf{reg}}(0) = C_2 \mathbb{E}[\Theta^2] \ge 0$ for some positive constants $C_1, C_2$.

REMARK 8 (Negative estimator error correlation). Let us take $\mathsf{reg}, \widetilde{\mathsf{reg}}$ as the indicator functions

$$\mathsf{reg}(x) = \Pi_{(-\infty,-t]}(x) := \begin{cases} +\infty & x > -t \\ 0 & x \le -t \end{cases} \quad \text{and} \quad \widetilde{\mathsf{reg}}(x) = \Pi_{[\widetilde{t},+\infty)}(x) := \begin{cases} +\infty & x < \widetilde{t} \\ 0 & x \ge \widetilde{t} \end{cases}$$

where $t, \widetilde{t} \ge 0$ are non-negative constants. Note that the two sets, $(-\infty, -t]$ and $[\widetilde{t}, +\infty)$, are disjoint. Noting $\mathsf{prox}_{\mathsf{reg}}(x) = \min\{x, -t\}$ and $\mathsf{prox}_{\widetilde{\mathsf{reg}}}(x) = \max\{x, \widetilde{t}\}$, we have

$$F_{\mathsf{reg}}(\eta_H) = \frac{1}{\nu\widetilde{\nu}\alpha\widetilde{\alpha}} \mathbb{E}\big[\big(\min\{\Theta + \tfrac{\beta}{\nu} H, -t\} - \Theta\big) \cdot \big(\max\{\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}} \widetilde{H}, \widetilde{t}\} - \Theta\big)\big].$$

Thus, if $\Theta$ is included in the closed set $[-t, \widetilde{t}]$ with probability 1, then we have $F_{\mathsf{reg}}(\eta_H) \le 0$ for all $\eta_H$ so that $\eta_G^\star = F_{\mathsf{reg}}(\eta_H^\star)$ is non-positive. This is intuitive as we will see in Theorem 2 that $\eta_G$ characterizes the limiting behavior of the correlations between two estimators trained on $\mathsf{reg}$ and $\widetilde{\mathsf{reg}}$.

## APPENDIX B: PROOFS FOR RESULTS IN SECTION 3

### B.1. Proof of Theorem 1.

*Part 1 .* We redefine $F_{\mathsf{loss}}$ and $F_{\mathsf{reg}}$ for convenience:

$$F_{\mathsf{loss}}(\eta_{_G}) = \tfrac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa) \cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})],$$

$$F_{\mathsf{reg}}(\eta_{_H}) = \tfrac{1}{\alpha\widetilde{\alpha}} \cdot \mathbb{E}\Big[\Big(\tfrac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{1}{\nu}\Big) - \tfrac{\beta}{\nu}H\Big) \cdot \Big(\tfrac{1}{\widetilde{\nu}} \cdot \mathsf{env}'_{\widetilde{\mathsf{reg}}}\Big(\tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H} + \Theta; \tfrac{1}{\widetilde{\nu}}\Big) - \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}\Big)\Big].$$

By the Cauchy–Schwarz inequality, using (4b) in System 1a, we have

$$\begin{aligned}
|F_{\mathsf{loss}}(\eta_{_G})| &\le \tfrac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]^{1/2} \cdot \mathbb{E}[\mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G} + Z; \widetilde{\kappa})^2]^{1/2} \\
&= \tfrac{c\widetilde{c}\delta}{\beta\widetilde{\beta}} \cdot \sqrt{\tfrac{\beta^2}{c\delta}} \cdot \sqrt{\tfrac{\widetilde{\beta}^2}{\widetilde{c}\delta}} \\
&= \sqrt{c\widetilde{c}}.
\end{aligned}$$

for all $\eta_{_G} \in [-1,1]$. By the same argument, the Cauchy–Schwarz inequality and (4a) in System 1a yield $|F_{\mathsf{loss}}(\eta_{_H})| \le 1$ for all $\eta_{_H} \in [-1,1]$.

*Part 2 .* Let us prove the differentiability and contraction of compositions. We will use the following lemma to argue the differentiability of $F_{\mathsf{loss}}$ and $F_{\mathsf{reg}}$.

LEMMA 12. *Let $G$ and $Z$ be independent $\mathcal{N}(0,1)$ random variables. Then, for any Lipschitz functions $(f, \widetilde{f})$ with bounded second moment $\mathbb{E}[f(G)^2], \mathbb{E}[\widetilde{f}(G)^2] < +\infty$, the map*

$$\varphi : [-1,1] \to \mathbb{R}, \quad \eta \mapsto \mathbb{E}[f(G)\widetilde{f}(\eta G + \sqrt{1-\eta^2}Z)]$$

*has the derivative*

$$\varphi'(\eta) = \mathbb{E}[f'(G)\widetilde{f}'(\eta G + \sqrt{1-\eta^2}Z)].$$

PROOF. Since $\widetilde{f}$ is Lipschitz and $\mathcal{N}(0,1)$ has no point mass, $\widetilde{f}$ is differentiable at $G \sim \mathcal{N}(0,1)$ with probability 1. By the dominated convergence theorem, we have

$$\varphi'(\eta) = \mathbb{E}\big[f(G)\widetilde{f}'\big(\eta G + \sqrt{1-\eta^2}Z\big)\big(G - \frac{\eta}{\sqrt{1-\eta^2}}Z\big)\big].$$

Let us define $A = \eta G + \sqrt{1-\eta^2}Z$ and $B = \sqrt{1-\eta^2}G - \eta Z$ so that $(A,B)$ are independent Gaussian $\mathcal{N}(0,1)$ and $\varphi'(\eta) = (1-\eta^2)^{-1/2}\mathbb{E}[f(\eta A + \sqrt{1-\eta^2}B)\widetilde{f}'(A)B]$. Using Stein's formula for $B$ conditionally on $A$, we are left with

$$\varphi'(\eta) = \mathbb{E}[\widetilde{f}'(A)f'(\eta A + \sqrt{1-\eta^2}B)] = \mathbb{E}[\widetilde{f}'(\eta G + \sqrt{1-\eta^2}Z)f'(G)],$$

where we have used $(A, \eta A + \sqrt{1-\eta^2}B) \overset{\mathrm{d}}{=} (\eta G + \sqrt{1-\eta^2}Z, G)$. (Here and throughout $\overset{\mathrm{d}}{=}$ refers to equality in distribution.) □

Notice that $\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)$ and $\tfrac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{1}{\nu}\Big) - \tfrac{\beta}{\nu}H$ have finite second moments, thanks to the existence of the solution to System 1a. Thus, applying Lemma 12 with $(f(x), \widetilde{f}(x)) = (\mathsf{env}'_{\mathsf{loss}}(\alpha x + Z; \kappa), \mathsf{env}'_{\mathsf{loss}}(\widetilde{\alpha}x + Z; \widetilde{\kappa}))$ and $(f(x), \widetilde{f}(x)) = (\tfrac{\beta}{\nu}x -$

$\frac{1}{\nu}\mathsf{env}_{\mathsf{reg}}(\frac{\beta}{\nu}x+\Theta;\frac{1}{\nu}), \frac{\widetilde{\beta}}{\widetilde{\nu}}x-\frac{1}{\widetilde{\nu}}\mathsf{env}_{\mathsf{reg}}(\frac{\widetilde{\beta}}{\widetilde{\nu}}x+\Theta;\frac{1}{\widetilde{\nu}}))$, we find that $F_{\mathsf{loss}}$ and $F_{\mathsf{reg}}$ are differentiable and the derivatives are given by:

$$F'_{\mathsf{loss}}(\eta_G) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\cdot\mathbb{E}[\alpha\mathsf{env}''_{\mathsf{loss}}(\alpha G+Z;\kappa)\cdot\widetilde{\alpha}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G}+Z;\widetilde{\kappa})]$$

$$F'_{\mathsf{reg}}(\eta_H) = \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\frac{\beta}{\nu}\Big(1-\frac{1}{\nu}\cdot\mathsf{env}''_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H+\Theta;\frac{1}{\nu}\Big)\Big)\cdot\frac{\widetilde{\beta}}{\widetilde{\nu}}\Big(1-\frac{1}{\widetilde{\nu}}\cdot\mathsf{env}''_{\widetilde{\mathsf{reg}}}\Big(\frac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}+\Theta;\frac{1}{\widetilde{\nu}}\Big)\Big)\Big]$$

for all $\eta_G,\eta_H\in[-1,1]$. Note that the non-expansiveness of the proximal operator implies that the map $x\mapsto\mathsf{env}'_f(x;\tau)=\tau^{-1}(x-\mathsf{prox}_f(x;\tau))$ is $\tau^{-1}$-Lipschitz and non-decreasing for any convex function $f$. Thus, $F'_{\mathsf{loss}}(\eta_G)$ and $F'_{\mathsf{reg}}(\eta_H)$ are non-negative and uniformly bounded from above as follows:

$$\begin{aligned}
0\le F'_{\mathsf{loss}}(\eta_G) &\le \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa}\cdot\mathbb{E}[\widetilde{\alpha}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G}+Z;\widetilde{\kappa})] && (0\le\mathsf{env}''_{\mathsf{loss}}(\alpha G+Z;\kappa)\le\kappa^{-1})\\
&= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa}\cdot\mathbb{E}[\widetilde{G}\cdot\mathsf{env}'_{\widetilde{\mathsf{loss}}}(\widetilde{\alpha}\widetilde{G}+Z;\widetilde{\kappa})] && \text{(by Stein's lemma)}\\
&= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{\alpha}{\kappa}\cdot\frac{\widetilde{\nu}\widetilde{\alpha}}{\widetilde{c}\delta} = c\cdot\frac{\alpha\widetilde{\alpha}\widetilde{\nu}}{\beta\widetilde{\beta}\kappa} && \text{(using (4d) in System 1a);}
\end{aligned}$$

$$\begin{aligned}
0\le F'_{\mathsf{reg}}(\eta_H) &\le \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\frac{\beta}{\nu}\Big(1-\frac{1}{\nu}\cdot\mathsf{env}''_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H+\Theta;\frac{1}{\nu}\Big)\Big)\Big]\cdot\frac{\widetilde{\beta}}{\widetilde{\nu}} && (0\le\mathsf{env}''_{\widetilde{\mathsf{reg}}}(\frac{\widetilde{\beta}}{\widetilde{\nu}}H+\Theta;\frac{1}{\widetilde{\nu}})\le\widetilde{\nu})\\
&= \frac{1}{\alpha\widetilde{\alpha}}\frac{\widetilde{\beta}}{\widetilde{\nu}}\mathbb{E}\Big[H\Big(\frac{\beta}{\nu}H-\frac{1}{\nu}\cdot\mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu}H+\Theta;\frac{1}{\nu}\Big)\Big)\Big] && \text{(by Stein's lemma)}\\
&= \frac{1}{\alpha\widetilde{\alpha}}\frac{\widetilde{\beta}}{\widetilde{\nu}}\cdot\beta\kappa = \frac{\beta\widetilde{\beta}\kappa}{\alpha\widetilde{\alpha}\widetilde{\nu}} && \text{(using (4c) in System 1a).}
\end{aligned}$$

Thus, by the chain rule, noting that $(\beta\widetilde{\beta}\kappa)/(\alpha\widetilde{\alpha}\widetilde{\nu})$ is cancelled out, we have

$$0\le(F_{\mathsf{loss}}\circ F_{\mathsf{reg}})'(\eta_H)\le c,\quad 0\le(F_{\mathsf{reg}}\circ F_{\mathsf{loss}})'(\eta_G)\le c.$$

By switching the role of $(\alpha,\beta,\kappa,\nu,c)$ and $(\widetilde{\alpha},\widetilde{\beta},\widetilde{\kappa},\widetilde{\nu},\widetilde{c})$, it also holds that

$$0\le(F_{\mathsf{loss}}\circ F_{\mathsf{reg}})'(\eta_H)\le \widetilde{c},\quad 0\le(F_{\mathsf{reg}}\circ F_{\mathsf{loss}})'(\eta_G)\le \widetilde{c}.$$

Thus, by taking the minimum of $(c,\widetilde{c})$, we find that the compositions

$$\eta_H\mapsto F_{\mathsf{loss}}\circ F_{\mathsf{reg}}(\eta_H),\quad \eta_G\mapsto F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\eta_G)$$

are $(c\wedge\widetilde{c})$-Lipschitz.

*Part 3.* Let us show the uniqueness and existence of the solution to the nonlinear system:

$$\eta_H = F_{\mathsf{loss}}(\eta_G),\quad \eta_G = F_{\mathsf{reg}}(\eta_H).$$

We have shown in the previous paragraph that the map $F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\eta_H):[-1,1]\to[-1,1]$ are non-decreasing, differentiable, and 1-Lipschitz. Thus, by Brouwer's fixed-point theorem, there exists some $\eta_G^\star\in[-1,1]$ such that

$$\eta_G^\star = F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\eta_G^\star)$$

Letting $\eta_H^\star = F_{\mathsf{loss}}(\eta_H^\star)\in[-\sqrt{c\widetilde{c}},\sqrt{c\widetilde{c}}]$, we find that the pair $(\eta_G^\star,\eta_H^\star)$ satisfies the nonlinear system. Next, we show the uniqueness. Suppose $(\eta_G,\eta_H)$ and $(\widetilde{\eta}_G,\widetilde{\eta}_H)$ satisfy the system. Then, $\eta_G$ and $\widetilde{\eta}_G$ are the solution to the fixed-point equation $\eta = F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\eta)$:

$$\eta_G = F_{\mathsf{reg}}(\eta_H) = F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\eta_G),\quad \widetilde{\eta}_G = F_{\mathsf{reg}}(\widetilde{\eta}_H) = F_{\mathsf{reg}}\circ F_{\mathsf{loss}}(\widetilde{\eta}_G).$$

Suppose $\eta_{G} \neq \widetilde{\eta}_{G}$. This implies that the map $\eta \mapsto \eta - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta)$ is constant over a nonempty closed interval of $[-1,1]$, and hence we can find some $\eta^{\star}_{G} \in (-1,1)$ such that $1-(F_{\mathsf{reg}} \circ F_{\mathsf{loss}})'(\eta^{\star}_{G}) = 0$. Recalling the argument in the previous paragraph, we must have $F'_{\mathsf{loss}}(\eta^{\star}_{G}) = c\frac{\alpha\widetilde{\alpha}\widetilde{\nu}}{\beta\widetilde{\beta}\kappa}$, and this equality case gives

$$\mathbb{E}\Big[\mathsf{env}''_{\widetilde{\mathsf{loss}}}\big(\widetilde{\alpha}\eta^{\star}_{G} G + \sqrt{1-(\eta^{\star}_{G})^2}G_0 + Z;\widetilde{\kappa}\big)\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(\alpha G + Z;\kappa)\big)\Big] = 0.$$

for independent standard normals $(G, G_0)$. Since $\mathsf{env}''_f(\cdot;\kappa) \in [0,\kappa^{-1}]$ for any convex function $f$, the integrand is non-negative with probability $1$. Thus, we must have

$$\iint_{-\infty}^{\infty} \mathsf{env}''_{\widetilde{\mathsf{loss}}}\big(\widetilde{\alpha}\eta^{\star}_{G} x + \widetilde{\alpha}\sqrt{1-(\eta^{\star}_{G})^2}y + Z;\widetilde{\kappa}\big)\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(\alpha x + Z;\kappa)\big)\phi(x)\phi(y)dxdy = 0$$

with probability $1$ with respect to $Z$, where $\phi$ is the pdf of $\mathcal{N}(0,1)$. By the change of variable $(x,y) \mapsto (u,v) = (ax+Z, \widetilde{\alpha}\eta^{\star}_{G} x + \widetilde{\alpha}\sqrt{1-(\eta^{\star}_{G})^2}y + Z)$, since $(u,v)$ also spans $\mathbb{R}\times\mathbb{R}$ thanks to $|\eta^{\star}_{G}| < 1$, the previous display reads to

$$\iint_{-\infty}^{\infty} \mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(u;\kappa)\big)\phi\Big(\frac{u-Z}{\alpha}\Big)\phi\Big(\frac{v - Z - \eta^{\star}_{G}(u-Z)}{\alpha\sqrt{1-|\eta^{\star}_{G}|^2}}\Big)dudv = 0.$$

Let us decouple the integration $\iint(\cdot)dudv$ with a lower bound of the form $\int(\cdot)du \cdot \int(\cdot)dv$. Using the inequality $\exp(-(x+y)^2) \geq \exp(-2x^2 - 2y^2)$ for the density $\phi$ of $\mathcal{N}(0,1)$, we find that there exists some positive (deterministic) constants $(c, C)$ that depend on $(\alpha, \eta^{\star}_{G})$ only such that

$$\phi\Big(\frac{u-Z}{\alpha}\Big)\phi\Big(\frac{v - Z - \eta^{\star}_{G}(u-Z)}{\alpha\sqrt{1-|\eta^{\star}_{G}|^2}}\Big) \geq C\cdot\phi(c\cdot Z)\phi(c\cdot u)\phi(c\cdot v)$$

for all $u, v \in \mathbb{R}$ and any realization of $Z$. Combining this lower estimate with the previous display, we obtain

$$\phi(cZ)\cdot\int_{-\infty}^{\infty}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})\phi(cv)dv \cdot \int_{-\infty}^{\infty}\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(u;\kappa)\big)\phi(cu)du = 0.$$

with probability $1$ with respect to $Z$. Since $\phi(cZ)$ is always strictly positive, we must have

$$\int_{-\infty}^{\infty}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})\phi(cv)dv = 0 \quad \text{or} \quad \int_{-\infty}^{\infty}\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(u;\kappa)\big)\phi(cu)du = 0.$$

Suppose $\int_{-\infty}^{\infty}\mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})\phi(cv)dv = 0$ holds. This implies $\int_K \mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})dv = 0$ for any compact set $K \subset \mathbb{R}$ due to $\min_{v\in K}\phi(cv) > 0$ and the non-negativity of $\mathsf{env}''$. Since $x \mapsto \mathsf{env}'_{\widetilde{\mathsf{loss}}}(x;\kappa)$ is Lipschitz, it is absolutely continuous, and hence it has the integral form $\mathsf{env}'_{\widetilde{\mathsf{loss}}}(x;\widetilde{\kappa}) = \mathsf{env}'_{\widetilde{\mathsf{loss}}}(0;\widetilde{\kappa}) + \int_0^x \mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})dv$. Since we know $\int_0^x \mathsf{env}''_{\widetilde{\mathsf{loss}}}(v;\widetilde{\kappa})dv = 0$, we find that $x \mapsto \mathsf{env}'_{\widetilde{\mathsf{loss}}}(x;\widetilde{\kappa})$ is constant. However, this cannot happen since (4d) in System 1a then gives $\widetilde{\nu}\widetilde{\alpha} = 0$: a contradiction with the positiveness of $(\widetilde{\alpha},\widetilde{\beta},\widetilde{\kappa},\widetilde{\nu})$. Therefore, the other case $\int_{-\infty}^{\infty}\big(\kappa^{-1} - \mathsf{env}''_{\mathsf{loss}}(u;\kappa)\big)\phi(cu)du = 0$ must hold. However, by the same argument integrating the constant derivative over $[0,x]$, this in turn gives $\mathsf{env}'_{\mathsf{loss}}(x;\kappa) = \mathsf{env}'_{\mathsf{loss}}(0;\kappa) + \kappa^{-1}x$. Noting $\mathsf{env}'_{\mathsf{loss}}(x;\kappa) = (x - \mathsf{prox}_{\mathsf{loss}}(x;\kappa))/\kappa$, this means that the map $x \mapsto \mathsf{prox}_{\mathsf{loss}}(x;\kappa)$ is constant. Denoting by $p_0$ the constant, we have that for all real $u$,

$$\kappa^{-1}(u - p_0) = \kappa^{-1}(u - \mathsf{prox}_{\mathsf{loss}}(u;\kappa)) \in \partial\mathsf{loss}(p_0),$$

which means $\partial\mathsf{loss}(p_0) = \mathbb{R}$. For any $p \in \mathbb{R}\setminus\{p_0\}$, the definition of the subdifferential gives $\mathsf{loss}(p) \geq \mathsf{loss}(p_0) + \sup_{u\in\partial\mathsf{loss}(p_0)} u(p - p_0) = +\infty$. This contradicts the assumption that $\mathsf{loss}(p) < +\infty$ at two distinct $p \in \mathbb{R}$.

**B.2. Proof of Theorem 2.** In this section, for a vector $\boldsymbol{w}$, the norm $\|\boldsymbol{w}\|$ indicates the $\ell_2$ norm unless specified otherwise. By the assumption $0 \in \operatorname{argmin}_x \mathsf{loss}(x) \cap \operatorname{argmin}_x \mathsf{reg}(x)$, taking $\mathsf{loss}_{\text{new}}(x) = \mathsf{loss}(x) - \mathsf{loss}(0)$ and $\mathsf{reg}_{\text{new}}(x) = \mathsf{reg}(x) - \mathsf{reg}(0)$ if necessary, we assume without loss of generality that loss and reg are non-negative and have 0 as a minimizer.

By the change of variable $\boldsymbol{b} \mapsto \boldsymbol{h} = (\boldsymbol{b} - \boldsymbol{\theta})/\sqrt{p}$, denoting $\boldsymbol{G} = \sqrt{p}\boldsymbol{X}$ so that $\boldsymbol{G}$ has i.i.d. $\mathcal{N}(0,1)$ entries, the regularized M-estimator $\widehat{\boldsymbol{\theta}}$ of interest and the residual vector $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}$ can be written as

$$\widehat{\boldsymbol{\theta}} = \sqrt{p}\widehat{\boldsymbol{h}} + \boldsymbol{\theta}, \quad \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}$$

where

$$\widehat{\boldsymbol{h}} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) \quad \text{with} \quad \mathsf{obj}(\boldsymbol{h}) := \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \sum_{j\in[p]} \mathsf{reg}(\sqrt{p}h_j + \theta_j). \tag{32}$$

Throughout this section, we denote $\boldsymbol{\psi} = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) \in \mathbb{R}^n$ where $\boldsymbol{e}_i$ are canonical basis of $\mathbb{R}^n$. Let $\widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{h}}$ be the corresponding notation for another. Then, our goal is to show

$$\widehat{\boldsymbol{h}}^\top \widetilde{\boldsymbol{h}} \xrightarrow{\mathrm{p}} \alpha\widetilde{\alpha}\eta_{_G} \quad \text{and} \quad \boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/p \xrightarrow{\mathrm{p}} \beta\widetilde{\beta}\eta_{_H}.$$

B.2.1. *Smoothing by adding diminishing ridge penalty.* For some positive and diminishing scalar $\mu = \mu_n \to 0$ to be specified later, we define the smoothed regularized M-estimator $\widehat{\boldsymbol{h}}_\mu$ as

$$\widehat{\boldsymbol{h}}_\mu \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 \tag{33}$$

where $\mathsf{obj}(\boldsymbol{h})$ is the objective function for the original regularized M-estimator $\widehat{\boldsymbol{h}}$ in (32). We denote by $\boldsymbol{\psi}_\mu = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}_\mu)$ the smoothed version of $\boldsymbol{\psi}$. This strategy of adding and additive ridge penalty for the mathematical analysis as a first step, and then arguing by approximation ($\mu \to 0$) to study the original $\widehat{\boldsymbol{h}}_0 = \widehat{\boldsymbol{h}}$ is ubiquitous; see, for instance, [64, 23], [73, Appendix B.3], [16]. Here, we use the following lemma to show that (33) approximates well the original estimator for any sequence $\mu = \mu_n$ indexed by $n$ and converging to 0.

LEMMA 13 (Ridge smoothing). *Let $\boldsymbol{h}_\mu$ be the smoothed M-estimator and $\widehat{\boldsymbol{h}}$ be the original M-estimator.*

1. *(Monotonicity) We have $\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|_2^2 \le \|\widehat{\boldsymbol{h}}\|_2^2 - \|\widehat{\boldsymbol{h}}_\mu\|_2^2$ for any $\mu > 0$.*
2. *(Convergence in $\|\cdot\|_2$) As $n, p \to +\infty$ with $n/p \to \delta$, $|I|/n \to c$, and $\mu = \mu_n$ for any $\mu_n \to 0$, we have*

$$\|\widehat{\boldsymbol{h}}_\mu\|_2^2 \xrightarrow{\mathrm{p}} \alpha^2, \quad \frac{\|\boldsymbol{\psi}_\mu\|_2^2}{p} \xrightarrow{\mathrm{p}} \beta^2, \quad \|\widehat{\boldsymbol{h}}_\mu - \widehat{\boldsymbol{h}}\|_2^2 = o_{\mathbb{P}}(1), \quad \frac{\|\boldsymbol{\psi}_\mu - \boldsymbol{\psi}\|_2^2}{p} = o_{\mathbb{P}}(1),$$

*where $\alpha$ and $\beta$ are solutions to System 1a.*

PROOF. By the strong convexity of the ridge term $p\mu^2/2\|\boldsymbol{h}\|^2$, the smoothed one $\widehat{\boldsymbol{h}}_\mu$ also minimizes the convex function: $\boldsymbol{h} \mapsto \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 - \frac{p\mu}{2}\|\boldsymbol{h} - \widehat{\boldsymbol{h}}_\mu\|^2$. By the optimality of $\widehat{\boldsymbol{h}}_\mu$ and $\widehat{\boldsymbol{h}}$, we have

$$\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}_\mu\|^2 \le \mathsf{obj}(\widehat{\boldsymbol{h}}) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}\|^2 - \frac{p\mu}{2}\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|^2 \le \mathsf{obj}(\widehat{\boldsymbol{h}}_\mu) + \frac{p\mu}{2}\|\widehat{\boldsymbol{h}}\|^2 - \frac{p\mu}{2}\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}_\mu\|^2$$

so that subtracting $\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu)$ from the both sides and dividing by $p\mu/2>0$, we obtain the first claim.

Recall $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ and $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2$. Then, by Lemma 13-(1) and the Lipschitz condition of $\mathsf{loss}'$, it suffices to show the convergence $\|\widehat{\boldsymbol{h}}_\mu\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ for the smoothed estimator $\widehat{\boldsymbol{h}}_\mu$. Note that $\widehat{\boldsymbol{h}}_\mu$ minimizes the function below

$$\boldsymbol{h} \mapsto \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \sum_{j\in[p]} \mathsf{reg}_j^{\mu_n}(\sqrt{p}h_j) \quad \text{where} \quad \mathsf{reg}_j^{\mu}(x) := \mathsf{reg}(x+\theta_j) + \mu\frac{x^2}{2}.$$

Now we suppose that for any standard normal $\boldsymbol{g} = (g_j)_{j=1}^p \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)$, the convergence of the Moreau envelope

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) - \mathsf{reg}_j^{\mu_n}(0) \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH+\Theta;\tau) - \mathsf{reg}(\Theta)]$$

holds for all $c\in\mathbb{R}$ and $\tau>0$. Then, by [95, Theorem 3.1], we have $\|\widehat{\boldsymbol{h}}_\mu\|^2 \xrightarrow{\mathrm{p}} \alpha^2$ and complete the proof. By the weak law of large numbers, the above display holds with $\mu=0$ (see [95, Lemma 4.1] for details):

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) - \mathsf{reg}_j^{\mu=0}(0) \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH+\Theta;\tau) - \mathsf{reg}(\Theta)].$$

Then, noting $\mathsf{reg}_j^{\mu=0}(0) = \mathsf{reg}(\theta_j) = \mathsf{reg}_j^{\mu}(0)$ for any $\mu\ge 0$, it suffices to show

$$\frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) - \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) = o_{\mathbb{P}}(1).$$

For each $j$, the monotonicity $\mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) \le \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau)$ holds since the objective function is monotone in the sense of $\mathsf{reg}_j^{\mu=0}(x) \le \mathsf{reg}_j^{\mu}(x)$ for all $x\in\mathbb{R}$ and all $\mu\ge 0$. On the other hand, by the optimality of $\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau)$ and $\mathsf{prox}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau)$, we find

$$\begin{aligned}
\mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) &\le \frac{1}{2\tau}(cg_j - \mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2 + \mathsf{reg}_j^{\mu_n}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau)) \\
&= \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) + \frac{\mu_n}{2}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2
\end{aligned}$$

for each $j\in[p]$. Thus, it holds that

$$0 \le \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu_n}}(cg_j;\tau) - \frac{1}{p}\sum_{j\in[p]} \mathsf{env}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) \le \frac{\mu_n}{2}\cdot\frac{1}{p}\sum_{j\in[p]}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2,$$

where $\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau) = \mathsf{prox}_{\mathsf{reg}}(cg_j+\theta_j;\tau) - \theta_j$ by the definition of $\mathsf{reg}_j^{\mu}$. Here, under Assumption D-(1), the expectation $\mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(cg_j+\theta_j;\tau)-\theta_j)^2]$ under $g_j\sim\mathcal{N}(0,1)$ and $\theta_j\sim\Theta$ is finite for all $c\in\mathbb{R}$ and $\tau>9$ (cf. [95, equation (123)]). Therefore, by the weak law of large numbers, we have $\frac{1}{p}\sum_{j\in[p]}(\mathsf{prox}_{\mathsf{reg}_j^{\mu=0}}(cg_j;\tau))^2 \xrightarrow{\mathrm{p}} \mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(cH+\Theta;\tau)-\Theta)^2] < +\infty$. This means that for any $\mu=\mu_n\to 0$, the RHS of the previous display is $o_{\mathbb{P}}(1)$. $\square$

B.2.2. *Bounding norm of regularized M-estimators.* For some positive scalar $K$ to be specified later, we add another regularization term; we define $\widehat{\boldsymbol{h}}_{\mu,K}$ as

$$\widehat{\boldsymbol{h}}_{\mu,K} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \underbrace{\frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2 - K}{2}\Big)}_{\text{additional term}} \quad \text{with} \qquad \widehat{\lambda} := \mathsf{obj}(\mathbf{0})$$

where $\mathsf{F}:\mathbb{R}\to\mathbb{R}$ is convex, non-negative, and non-decreasing with $\lim_{x\to+\infty}\mathsf{F}(x) = +\infty$, as well as differentiable with $\mathsf{F}'(u) = 0$ if $u \le 0$ and $\mathsf{F}'(u) = 1$ if $u \ge 1$. For instance, we may take $\mathsf{F}$ as an integral of the smoothed step function as follows:

$$\mathsf{F}(x) := \int_{-\infty}^{x} f(u)\,\mathrm{d}u \quad \text{with} \quad f(u) := \begin{cases} 1 & u \ge 1 \\ 3u^2 - 2u & u \in (0,1) \\ 0 & u \le 0 \end{cases}$$

Note in passing that the regularization parameter $\widehat{\lambda} = \mathsf{obj}(\mathbf{0}) = \sum_{i\in I}\mathsf{loss}(z_i) + \sum_{j\in[p]}\mathsf{reg}(\theta_j)$ is independent of the design matrix $\boldsymbol{G}$ and non-negative since $\mathsf{loss}$ and $\mathsf{reg}$ are non-negative.

Now we claim that for sufficiently large $K > 0$ this modified estimator $\widehat{\boldsymbol{h}}_{\mu,K}$ coincides with the smoothed estimator $\widehat{\boldsymbol{h}}_{\mu}$. Furthermore, thanks to the additional regularizer $\frac{\widehat{\lambda}}{2}\mathsf{F}\big(\frac{\|\boldsymbol{h}\|^2-K}{2}\big)$, the norm of $\boldsymbol{h}_{\mu,K}$ and $\boldsymbol{\psi}_{\mu,K}$ are suitably bounded as follows:

LEMMA 14. *The following convergences hold.*

1. *If we set $K = 2\alpha^2$ where $\alpha > 0$ is the solution to System 1a, we have*
$$\mathbb{P}(\widehat{\boldsymbol{h}}_{\mu,K} = \widehat{\boldsymbol{h}}_{\mu}) \xrightarrow{\mathrm{p}} 1.$$

2. *For any $K \ge 0$, there exists a positive constant $C_K$ that only depends on $K$ such that*
$$\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \le C_K, \quad \|\boldsymbol{\psi}_{\mu,K}\|^2 \le C_K(1 + \|\mathsf{loss}'\|_{\mathrm{Lip}}^2)(\|\mathsf{loss}'(\boldsymbol{z})\|^2 + \|\boldsymbol{G}\|_{\mathrm{op}}^2).$$
*Thus, the constant $C^* = 1.1C_K(1+\|\mathsf{loss}'\|_{\mathrm{Lip}}^2)(\mathbb{E}[\mathsf{loss}'(Z)^2] + (1+\delta^{-1/2}))^2$ satisfies*
$$\mathbb{P}\Big(\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \vee (n^{-1}\|\boldsymbol{\psi}_{\mu,K}\|^2) \vee \sup_{m\ge1}\mathbb{E}\Big[\|\widehat{\boldsymbol{h}}_{\mu,K}\|^{2m} \vee (n^{-1}\|\boldsymbol{\psi}_{\mu,K}\|^2)^m \mid \boldsymbol{z}, \boldsymbol{\theta}\Big]^{1/m} \le C^*\Big) \to 1.$$

PROOF. Let us consider the event $\Omega := \{\|\boldsymbol{h}\|^2 \le K\}$ with $K = 2\alpha^2$, which holds with high probability, since $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$. Combined with the monotonicity $\|\boldsymbol{h}_\mu\|^2 \le \|\boldsymbol{h}\|^2$ by Lemma 13-(1), we have $\|\widehat{\boldsymbol{h}}_\mu\|^2 \le K$ under the event $\Omega$. Since $\mathsf{F}'(u) = 0$ for all $u \le 0$, combined with the Karush-Kuhn-Tucker (KKT) condition $-p\mu\widehat{\boldsymbol{h}}_\mu \in \partial\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu)$ for the smoothed estimator $\widehat{\boldsymbol{h}}_\mu$, we observe that $\widehat{\boldsymbol{h}}_\mu$ satisfies the KKT conditions for $\widehat{\boldsymbol{h}}_{\mu,K}$ under the event $\Omega$:

$$-p\mu\widehat{\boldsymbol{h}}_\mu - \widehat{\lambda}\mathsf{F}'\Big(\frac{\|\widehat{\boldsymbol{h}}_\mu\|^2 - K}{2}\Big)\widehat{\boldsymbol{h}}_\mu = -p\mu\widehat{\boldsymbol{h}}_\mu - 0\cdot\widehat{\boldsymbol{h}}_\mu \in \partial\mathsf{obj}(\widehat{\boldsymbol{h}}_\mu).$$

This implies $\mathbb{P}(\widehat{\boldsymbol{h}}_\mu = \widehat{\boldsymbol{h}}_{\mu,K}) \ge \mathbb{P}(\Omega) \to 1$.

By the non-negativity of $(\mathsf{obj}(\cdot), p\mu\|\cdot\|^2, \frac{\widehat{\lambda}}{2}\mathsf{F}(\cdot))$ and the optimality of $\widehat{\boldsymbol{h}}_{\mu,K}$, it holds that

$$0 \le \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 - K}{2}\Big) \le \mathsf{obj}(\widehat{\boldsymbol{h}}_{\mu,K}) + \frac{p\mu}{2}\|\boldsymbol{h}_{\mu,K}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}_{\mu,K}\|^2 - K}{2}\Big)$$

$$\leq \mathsf{obj}(\mathbf{0}) + \frac{p\mu}{2}\|\mathbf{0}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\mathbf{0}\|^2 - K}{2}\Big)$$
$$= \widehat{\lambda} + 0 + 0.$$

When $\widehat{\lambda} = 0$ then all inequalities above holds with equality, which means that $\mathbf{0}$ minimizes the objective function $\mathsf{obj}(\boldsymbol{h}) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2-K}{2}\Big)$ for $\widehat{\boldsymbol{h}}_{\mu,K}$. Since this objective function is strongly convex due to the ridge term, the minimizer is unique. This means $\widehat{\boldsymbol{h}}_{\mu,K} = \mathbf{0}$ when $\widehat{\lambda} = 0$. On the other hand, if $\widehat{\lambda} > 0$ then dividing the above display by $\widehat{\lambda} > 0$ we are left with $\mathsf{F}\Big(\frac{\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2-K}{2}\Big) \leq 2$. Since $\mathsf{F}$ is non-decreasing and coercive on the positive side, i.e., $\lim_{x\to+\infty}\mathsf{F}(x) \to +\infty$, this gives $\|\widehat{\boldsymbol{h}}_{\mu,K}\|^2 \leq C(K)$ for a constant $C(K) > 0$ depending on $K$ only. Combined with the Lipschitz condition of $\mathsf{loss}'$, the norm of $\boldsymbol{\psi}_{\mu,K} = \sum_{i\in I}\boldsymbol{e}_i\mathsf{loss}'(z_i - \boldsymbol{g}_i^\top\boldsymbol{h}_{\mu,K})$ is bounded as

$$\|\boldsymbol{\psi}_{\mu,K}\| \leq \|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'(\boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}_{\mu,K}) - \mathsf{loss}'(\boldsymbol{z})\| \leq \|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'\|_{\mathrm{Lip}}\|\boldsymbol{G}\|_{\mathrm{op}}\|\widehat{\boldsymbol{h}}_{\mu,K}\|.$$

Since $\mathsf{loss}'(z_i)^2$ has a finite second moment by Assumption D-(1), the weak law of large numbers gives $n^{-1}\|\mathsf{loss}'(\boldsymbol{z})\|^2 \xrightarrow{\mathrm{p}} \mathbb{E}[\mathsf{loss}'(Z)^2] < +\infty$. Since $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries, we also have $\|\boldsymbol{G}\|_{op}^2/n \xrightarrow{\mathrm{p}} (1+\delta^{-1/2})^2$ by standard results of the maximal singular value of a Gaussian matrix, e.g., [34, Theorem II.13]. This completes the proof. □

B.2.3. *Derivative formulae for strongly convex regularizer.*

LEMMA 15 ([17]). *Let $\widehat{\boldsymbol{h}} \in \mathbb{R}^p$ be the regularized M-estimator of the form*

$$\widehat{\boldsymbol{h}} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top\boldsymbol{h}) + \mathsf{R}(\boldsymbol{h}),$$

*where $I \subset [n]$ is a subset independent of $(\boldsymbol{g}_i)_{i\in[n]}$, $\mathsf{loss}: \mathbb{R}\to\mathbb{R}$ is a convex and differentiable function with Lipschitz derivative, $\mathsf{R}: \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ is a $\tau$-strongly convex regularizer for some $\tau > 0$, and $\mathsf{R}$ is independent of $(\boldsymbol{g}_i)$. Then, there exists a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ satisfying $\|\boldsymbol{A}\|_{\mathrm{op}} \leq \tau^{-1}$ and $\mathrm{tr}[\boldsymbol{A}] \geq 0$ such that $\widehat{\boldsymbol{h}} \in \mathbb{R}^p$ and $\boldsymbol{\psi} = \sum_{i\in I}\boldsymbol{e}_i\mathsf{loss}'(z_i - \boldsymbol{g}_i^\top\widehat{\boldsymbol{h}}) \in \mathbb{R}^n$ are both differentiable as functions of the design $(\boldsymbol{G}) = (g_{ij})$ with the derivative given by*

(34)
$$\forall i \in [n], \forall j \in [p], \quad \frac{\partial\widehat{\boldsymbol{h}}}{\partial g_{ij}} = \boldsymbol{A}\Big(\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi} - \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{e}_i\boldsymbol{e}_j^\top\widehat{\boldsymbol{h}}\Big), \quad \frac{\partial\boldsymbol{\psi}}{\partial g_{ij}} = -\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi} - \boldsymbol{V}\boldsymbol{e}_i\boldsymbol{e}_j^\top\widehat{\boldsymbol{h}}.$$

*where $\boldsymbol{D}$ and $\boldsymbol{V}$ are $n\times n$ matrices defined by $\boldsymbol{D} = \sum_{i\in I}\boldsymbol{e}_i\boldsymbol{e}_i^\top\mathsf{loss}''(z_i - \boldsymbol{g}_i^\top\widehat{\boldsymbol{h}})$ and $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}$. Furthermore, the matrix $\boldsymbol{V}$ defined above is positive semidefinite with its operator norm bounded as $\|\boldsymbol{V}\|_{\mathrm{op}} \leq \|\mathsf{loss}'\|_{\mathrm{Lip}}$.*

Recall that the modified estimator $\widehat{\boldsymbol{h}}_{\mu,K}$ in Appendix B.2.2 minimizes the objective function $\sum_{i\in I}\mathsf{loss}(z_i - \boldsymbol{g}_i^\top\boldsymbol{h}) + \mathsf{R}(\boldsymbol{h})$ where $\mathsf{R}$ is the regularizer of the form

$$\mathsf{R}(\boldsymbol{h}) := \sum_{j\in[p]}\mathsf{reg}(\sqrt{p}h_j + \theta_j) + \frac{p\mu}{2}\|\boldsymbol{h}\|^2 + \frac{\widehat{\lambda}}{2}\mathsf{F}\Big(\frac{\|\boldsymbol{h}\|^2 - K}{2}\Big),$$

which is $(p\mu)$-strongly convex. Thus, we can apply Lemma 15; there exists a matrix $\boldsymbol{A}_{\mu,K} \in \mathbb{R}^{p\times p}$ such that

$$\|\boldsymbol{A}_{\mu,K}\|_{\mathrm{op}} \le (p\mu)^{-1}, \quad \mathrm{tr}[\boldsymbol{A}_{\mu,K}] \ge 0, \tag{35}$$

and $\widehat{\boldsymbol{h}}_{\mu,K}$ and $\boldsymbol{\psi}_{\mu,K} = \sum_{i\in I} \boldsymbol{e}_i \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}}_{\mu,K})$ are differentiable with respect to the design matrix $\boldsymbol{G} = (g_{ij})$ as in (34) with

$$\boldsymbol{D}_{\mu,K} = \sum_{i\in I} \boldsymbol{e}_i\boldsymbol{e}_i^\top \mathsf{loss}''(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}}_{\mu,K}) \qquad \boldsymbol{V}_{\mu,K} = \boldsymbol{D}_{\mu,K} - \boldsymbol{D}_{\mu,K}\boldsymbol{G}\boldsymbol{A}_{\mu,K}\boldsymbol{G}^\top\boldsymbol{D}_{\mu,K}.$$

Now we claim that the trace of $\boldsymbol{V}_{\mu,K}$ and $\boldsymbol{A}_{\mu,K}$ are empirical quantities that converge to the remaining solution $\nu$ and $\kappa$:

LEMMA 16. *For any $\mu \to 0$ such that $\mu^{-1} = O(n^{1/8})$, we have*

$$\mathrm{tr}[\boldsymbol{V}_{\mu,K}]/p \xrightarrow{\mathrm{p}} \nu \quad \text{and} \quad \mathrm{tr}[\boldsymbol{A}_{\mu,K}] \xrightarrow{\mathrm{p}} \kappa.$$

PROOF. See Appendix B.5. □

B.2.4. *Proof of Theorem 2.* Below we will take the diminishing constant $\mu_n \to 0$ as in Lemma 16. Dropping the dependence on $(\mu, K)$ for simplicity, we denote $\boldsymbol{h} = \widehat{\boldsymbol{h}}_{\mu,K}$, $\boldsymbol{\psi} = \boldsymbol{\psi}_{\mu,K}$, and $\boldsymbol{V} = \boldsymbol{V}_{\mu,K}$, $\boldsymbol{A} = \boldsymbol{A}_{\mu,K}$. In the same way, we use the notation $(\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{A}})$ for the other estimator. Then, by the convergence in $\|\cdot\|_2$ from Lemma 13 and Lemma 14, it suffices to show the convergence of correlation for the modified ones:

$$\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) \xrightarrow{\mathrm{p}} \eta_{\scriptscriptstyle G} \quad \text{and} \quad \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_{\scriptscriptstyle H}.$$

First, we will argue by the Gaussian Poincaré inequality that $\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha})$ and $\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta})$ concentrate on random quantities that are independent of the scaled design matrix $\boldsymbol{G} = \sqrt{p}\boldsymbol{X}$. Throughout this section, we denote by $\bar{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I}]$ the conditional expectation with respect to the design matrix $\boldsymbol{G}$ given signal $\boldsymbol{\theta}$, noise $\boldsymbol{z}$ and subsample index $I, \widetilde{I}$. Since $\boldsymbol{G}$ is independent of $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$, the conditional distribution of $\boldsymbol{G}$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is still that of a matrix with i.i.d. $\mathcal{N}(0,1)$ entries.

LEMMA 17. *For any $\mu_n > 0$ such that $\mu_n^{-1} = o(n^{1/2})$, we have*

$$\bar{\mathbb{E}}\big[(\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \bar{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}])^2\big] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu^{-2}) = o_{\mathbb{P}}(1).$$
$$\bar{\mathbb{E}}\big[(\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}])^2\big] = \mathcal{O}_{\mathbb{P}}(n\mu^{-2}) = o_{\mathbb{P}}(n^2).$$

PROOF. By the Gaussian Poincaré inequality, noting $\partial_{ij}(\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}) = \widetilde{\boldsymbol{h}}^\top\partial_{ij}\boldsymbol{h} + \boldsymbol{h}^\top\partial_{ij}\widetilde{\boldsymbol{h}}$ with $\partial_{ij} = \frac{\partial}{\partial g_{ij}}$, the conditional variance of $\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}$ is bounded from above as

$$\bar{\mathbb{E}}\Big[(\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \bar{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}])^2\Big] \le \sum_{ij}\bar{\mathbb{E}}\Big[(\partial_{ij}(\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}))^2\Big] \le 2\sum_{ij}\bar{\mathbb{E}}\Big[(\widetilde{\boldsymbol{h}}^\top\partial_{ij}\boldsymbol{h})^2\Big] + 2\sum_{ij}\bar{\mathbb{E}}\Big[(\boldsymbol{h}^\top\partial_{ij}\widetilde{\boldsymbol{h}})^2\Big],$$

where we denote by $\sum_{ij} = \sum_{i=1}^n\sum_{j=1}^p$ for brevity. By the derivative formula (34), the first term on the RHS is bounded as

$$\begin{aligned}\sum_{ij}\Big(\widetilde{\boldsymbol{h}}^\top\partial_{ij}\boldsymbol{h}\Big)^2 &= \sum_{i,j}\Big(\widetilde{\boldsymbol{h}}^\top\boldsymbol{A}(\boldsymbol{e}_j\psi_i - \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{e}_i h_j)\Big)^2 \\ &\le 2\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{h}}\|^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{G}\|_{\mathrm{op}}^2\|\boldsymbol{D}\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{h}}\|^2\|\boldsymbol{h}\|^2.\end{aligned}$$

Using the upper bound $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ from (35) and the moment bound in Lemma 14-(2), the conditional expectation (with respect to $\bar{\mathbb{E}}$) of the RHS is $\mathcal{O}_{\mathbb{P}}(n^{-1}\mu_n^{-2})$. By symmetry, we also get $\bar{\mathbb{E}}[\sum_{ij}\bar{\mathbb{E}}[(\boldsymbol{h}^\top\partial_{ij}\widetilde{\boldsymbol{h}})^2]] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu_n^{-2})$.

We use a similar argument for the variance of $\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}$. The Gaussian Poincaré gives $\bar{\mathbb{E}}[(\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}} - \bar{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}])^2] \le 2\sum_{ij}\bar{\mathbb{E}}[(\widetilde{\boldsymbol{\psi}}^\top\partial_{ij}\boldsymbol{\psi})^2 + (\boldsymbol{\psi}^\top\partial_{ij}\widetilde{\boldsymbol{\psi}})^2]$, where

$$\begin{aligned}\sum_{ij}\Big(\widetilde{\boldsymbol{\psi}}^\top\partial_{ij}\boldsymbol{\psi}\Big)^2 &= \sum_{ij}\Big(\widetilde{\boldsymbol{\psi}}^\top(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i - \boldsymbol{V}\boldsymbol{e}_i\boldsymbol{w}_j)\Big)^2 \\ &\lesssim \|\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}^\top\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{\psi}}\|^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{V}\|_{\mathrm{op}}^2\|\widetilde{\boldsymbol{\psi}}\|^2\|\boldsymbol{h}\|^2.\end{aligned}$$

by the derivative formula. Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ and the moment bound in Lemma 14-(2), the conditional expectation $\bar{\mathbb{E}}[\cdot]$ of the RHS is $\mathcal{O}_{\mathbb{P}}(n\mu_n^{-2} + n) = \mathcal{O}_{\mathbb{P}}(n\mu_n^{-2})$. ☐

Let us define $\widehat{\eta}_{G}$ and $\widehat{\eta}_{H}$ as the "truncated" values of the conditional expectation of $\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}$ and $\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}$, respectively:

$$\widehat{\eta}_{G} := \Pi_{[-1,1]}\big(\bar{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]/(\alpha\widetilde{\alpha})\big), \quad \widehat{\eta}_{H} := \Pi_{[-1,1]}\big(\bar{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta})\big)$$

where $\Pi_{[-1,1]}(x) := \max(\min(x,1),-1)$ is the projection map onto $[-1,1]$. Here, we emphasize that $(\widehat{\eta}_{H},\widehat{\eta}_{G})$ is independent of the design matrix $\boldsymbol{G}$ since $\boldsymbol{G}$ is integrated out. Furthermore, the absolute values of $\widehat{\eta_{G}}$ and $\widehat{\eta_{H}}$ are less than 1 due to the truncation, and hence the two matrices

$$\begin{pmatrix}1 & \widehat{\eta}_{H}\\ \widehat{\eta}_{H} & 1\end{pmatrix} \text{ and } \begin{pmatrix}1 & \widehat{\eta}_{G}\\ \widehat{\eta}_{G} & 1\end{pmatrix}$$

are both positive semidefinite. By the concentration from Lemma 17 and the convergence $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2 > 0$, noting the truncation $x \mapsto \Pi_{[-1,1]}(x)$ is continuous, we find that the random variables $(\widehat{\eta}_{H},\widehat{\eta}_{G})$ defined above still capture the correlations, that is,

$$\begin{aligned}\widehat{\eta}_{H} &= \Pi_{[-1,1]}\Big(\frac{\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}}{\|\boldsymbol{\psi}\|\|\widetilde{\boldsymbol{\psi}}\|}\Big) + o_{\mathbb{P}}(1) = \frac{\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}}{\|\boldsymbol{\psi}\|\|\widetilde{\boldsymbol{\psi}}\|} + o_{\mathbb{P}}(1).\\ \widehat{\eta}_{G} &= \Pi_{[-1,1]}\Big(\frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big) + o_{\mathbb{P}}(1) = \frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|} + o_{\mathbb{P}}(1).\end{aligned}$$

where the second equation follows from the fact that the correlations are less than 1 in absolute values by the Cauchy–Schwarz inequality.

Now, we use the multivariate normal approximation below to motivate System 1b.

LEMMA 18 (Proposition 5.1 in [15]). *Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{I}_q)$ and let $\boldsymbol{F}: \mathbb{R}^q \to \mathbb{R}^{q\times M}$ be a locally Lipschitz function with $M \le q$. Then there exists a standard normal vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}_M, \boldsymbol{I}_M)$ such that*

$$\mathbb{E}\Big[\Big\|\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{z} - \sum_{l\in[q]}\frac{\partial\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{e}_l}{\partial z_l} - \Big\{\boldsymbol{F}(\boldsymbol{z})^\top\boldsymbol{F}(\boldsymbol{z})\Big\}^{1/2}\boldsymbol{w}\Big\|^2\Big] \le C_3\sum_{l\in[q]}\mathbb{E}\Big[\Big\|\frac{\partial\boldsymbol{F}(z)}{\partial z_l}\Big\|_F^2\Big],$$

*where $\{\cdot\}^{1/2}$ is the square root of the positive semidefinite matrix.*

For each $j \in [p]$, applying Lemma 18 with $\boldsymbol{F} = \left[\frac{\boldsymbol{\psi}}{\sqrt{p}\beta}\ \frac{\widetilde{\boldsymbol{\psi}}}{\sqrt{p}\widetilde{\beta}}\right] \in \mathbb{R}^{n\times 2}$ and $\boldsymbol{z} = \boldsymbol{G}\boldsymbol{e}_j \in \mathbb{R}^n$, using the derivative formula (34), we find that there exists a random vector $\boldsymbol{w}_j \in \mathbb{R}^2$ such that

$$\boldsymbol{w}_j \mid (\boldsymbol{\theta}, \boldsymbol{\epsilon}, \boldsymbol{G}^{-j}, I, \widetilde{I}) \overset{\mathrm{d}}{=} \mathcal{N}(\mathbf{0}_2, \boldsymbol{I}_2)$$

and

$$\overline{\mathbb{E}}\Big[\Big\|\frac{1}{\sqrt{p}}\begin{pmatrix}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j + \boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j)/\beta \\ (\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]\widetilde{h}_j + \widetilde{\boldsymbol{\psi}}^\top \widetilde{\boldsymbol{D}}\boldsymbol{G}\widetilde{\boldsymbol{A}}\boldsymbol{e}_j)/\widetilde{\beta}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2)\end{pmatrix}^{1/2} \boldsymbol{w}_j\Big\|^2\Big]$$

$$\le C_4 \sum_{i\in[n]} \overline{\mathbb{E}}\Big[\frac{1}{p\beta^2}\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2 + \frac{1}{p\widetilde{\beta}^2}\|\frac{\partial \widetilde{\boldsymbol{\psi}}}{\partial g_{ij}}\|^2\Big].$$

Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu_n)^{-1}$ from (35) and Lemma 14, we have

$$\overline{\mathbb{E}}\Big[\|\boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}),$$

$$\sum_{ij} \overline{\mathbb{E}}\Big[\|\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}}\|^2\Big] \le \overline{\mathbb{E}}\Big[2\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{V}\|_{\mathrm{F}}^2\|\boldsymbol{h}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(n\mu^{-2}).$$

Thus, summing over $j \in [p]$ in the previous display, noting $\mu^{-2} = o(n)$, we are left with

$$\sum_{j\in[p]}\Big\|\frac{1}{\sqrt{p}}\begin{pmatrix}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j)/\beta \\ (\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]\widetilde{h}_j)/\widetilde{\beta}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2)\end{pmatrix}^{1/2} \boldsymbol{w}_j\Big\|^2 = o_{\mathbb{P}}(n).$$

By the (1/2)-Hölder continuity for the matrix square root: $\|\boldsymbol{M}^{1/2} - \boldsymbol{N}^{1/2}\|_{\mathrm{op}} \le \|\boldsymbol{M} - \boldsymbol{N}\|_{\mathrm{op}}^{1/2}$ for positive semidefinite matrices $\boldsymbol{M}, \boldsymbol{N}$, combined with the convergence $\|\boldsymbol{\psi}\|^2/p \to \beta^2$, $\|\widetilde{\boldsymbol{\psi}}\|^2/p \to \widetilde{\beta}^2$ and the concentration $\widehat{\eta}_H = \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) + o_{\mathbb{P}}(1)$, we have

$$\Big\|\begin{pmatrix}|\|\boldsymbol{\psi}\|^2/(p\beta^2) & \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \\ \boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) & \|\widetilde{\boldsymbol{\psi}}\|^2/(p\widetilde{\beta}^2)\end{pmatrix}^{1/2} - \begin{pmatrix}1 & \widehat{\eta}_H \\ \widehat{\eta}_H & 1\end{pmatrix}^{1/2}\Big\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Combined with the previous display, we get

$$\sum_{j\in[p]}\Big\|\frac{1}{\sqrt{p}}\begin{pmatrix}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\boldsymbol{V}]h_j)/\beta \\ (\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + \mathrm{tr}[\widetilde{\boldsymbol{V}}]\widetilde{h}_j)/\widetilde{\beta}\end{pmatrix} - \begin{pmatrix}1 & \widehat{\eta}_H \\ \widehat{\eta}_H & 1\end{pmatrix}^{1/2} \boldsymbol{w}_j\Big\|^2 = o_{\mathbb{P}}(n) + o_{\mathbb{P}}(1)\sum_{j\in[p]}\|\boldsymbol{w}_j\|^2 = o_{\mathbb{P}}(n),$$

where the last equation follows from $\sum_{j\in[p]} \overline{\mathbb{E}}[\|\boldsymbol{w}_j\|^2] = 2p$, which follows from the fact that the marginal distribution of $\boldsymbol{w}_j$ given $(\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I})$ is $\mathcal{N}(0,1)$ (note, however, that we do not establish or take for granted that $(\boldsymbol{w}_j)_{j\in[p]}$ are i.i.d.). Using $\mathrm{tr}[\boldsymbol{V}]/p \xrightarrow{\mathrm{p}} \nu$ (Lemma 16) and $\|\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(1)$ from Lemma 14, we can also replace $\mathrm{tr}[\boldsymbol{V}]h_j$ by $p\kappa h_j$. As a consequence, we get

(36)
$$\sum_{j\in[p]}\Big\|\begin{pmatrix}\frac{1}{\sqrt{p}\beta}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\nu h_j) \\ \frac{1}{\sqrt{p}\widetilde{\beta}}(\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G}\boldsymbol{e}_j^\top + p\widetilde{\nu}\widetilde{h}_j)\end{pmatrix} - \begin{pmatrix}\widehat{w}_j \\ \widetilde{w}_j\end{pmatrix}\Big\|^2 = o_{\mathbb{P}}(n) \text{ where } \begin{pmatrix}\widehat{w}_j \\ \widetilde{w}_j\end{pmatrix} := \begin{pmatrix}1 & \widehat{\eta}_H \\ \widehat{\eta}_H & 1\end{pmatrix}^{1/2} \boldsymbol{w}_j.$$

Here, we emphasize that the conditional distribution of $(\widehat{w}_j, \widetilde{w}_j)$ is given by

$$\begin{pmatrix}\widehat{w}_j \\ \widetilde{w}_j\end{pmatrix} \mid (\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}^{-j}, I, \widetilde{I}) \overset{\mathrm{d}}{=} \mathcal{N}\Big(\mathbf{0}_2, \begin{pmatrix}1 & \widehat{\eta}_H \\ \widehat{\eta}_H & 1\end{pmatrix}\Big)$$

since the conditional distribution of $\boldsymbol{w}_j$ given $(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}^{-j}, I, \widetilde{I})$ is standard normal $\mathcal{N}(0_2, I_2)$ and $\widehat{\eta}_H$ is $\sigma(\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I})$-measurable. For each $j \in [p]$, let us define $\Xi_j$ and $\widetilde{\Xi}_j$ as

$$\Xi_j = \frac{\boldsymbol{\psi}^\top \boldsymbol{G} \boldsymbol{e}_j^\top + p\nu h_j}{\sqrt{p}\beta} - \widehat{w}_j, \quad \widetilde{\Xi}_j = \frac{\widetilde{\boldsymbol{\psi}}^\top \boldsymbol{G} \boldsymbol{e}_j^\top + p\widetilde{\nu}\widetilde{h}_j}{\sqrt{p}\widetilde{\beta}} - \widetilde{w}_j$$

so that the bound (36) reads $\sum_{j\in[p]}[\Xi_j^2 + \widetilde{\Xi}_j^2] = o_{\mathbb{P}}(n)$. By the KKT conditions $\boldsymbol{G}^\top \boldsymbol{\psi} \in \sqrt{p}\partial \mathsf{reg}(\sqrt{p}\boldsymbol{h} + \boldsymbol{\theta})$ and $\boldsymbol{G}^\top \widetilde{\boldsymbol{\psi}} \in \sqrt{p}\partial \widetilde{\mathsf{reg}}(\sqrt{p}\widetilde{\boldsymbol{h}} + \boldsymbol{\theta})$ for $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}$, respectively, $\sqrt{p}h_j$ and $\sqrt{p}\widetilde{h}_j$ can be written as

$$\begin{pmatrix} \sqrt{p}h_j \\ \sqrt{p}\widetilde{h}_j \end{pmatrix} = \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + \frac{\beta}{\nu}(\widehat{w}_j + \Xi_j); \nu^{-1}) - \theta_j \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + \frac{\widetilde{\beta}}{\widetilde{\nu}}(\widetilde{w}_j + \widetilde{\Xi}_j); \widetilde{\nu}^{-1}) - \theta_j \end{pmatrix}.$$

Now we claim that the bound $\sum_{j\in[p]} \Xi_j^2 = o_{\mathbb{P}}(n)$ also holds in the conditional expectation $\overline{\mathbb{E}}$, i.e., $\overline{\mathbb{E}}[\sum_{j\in[p]} \Xi_j^2] = o_{\mathbb{P}}(n)$. To this end, it suffices to show $\overline{\mathbb{E}}[(p^{-1}\sum_{j\in[p]} \Xi_j^2)^2] \le C$ with high probability for a constant $C$. Using the upper estimate

$$0 \le \frac{1}{p}\sum_{j\in[p]} \Xi_j^2 \le 3\Big(\frac{\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2}{\beta^2 p^2} + \frac{\nu^2}{\beta^2}\|\boldsymbol{h}\|^2 + \frac{1}{p}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)$$

and the moment bound form Lemma 14, we get

$$\overline{\mathbb{E}}\Big[\Big(\frac{1}{p}\sum_{j\in[p]}\Xi_j^2\Big)^2\Big] \le C_5\Big(1 + \overline{\mathbb{E}}\Big[\Big(\frac{1}{p}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)^2\Big]\Big)$$

with high probability. For the second term, expanding the square of the summation,

$$\overline{\mathbb{E}}\Big[\Big(p^{-1}\sum_{j\in[p]}(\widehat{w}_j)^2\Big)^2\Big] = \frac{1}{p^2}\sum_{i,j}\overline{\mathbb{E}}[(\widehat{w}_i)^2(\widehat{w}_j)^2] \le \frac{1}{p^2}\sum_{i,j}\sqrt{\overline{\mathbb{E}}[(\widehat{w}_i)^4]}\sqrt{\overline{\mathbb{E}}[(\widehat{w}_j)^4]} = 6.$$

where we have used $\overline{\mathbb{E}}[(\widehat{w}_j)^4] = 6$ for all $j$, which follows from the fact that the marginal law of $\widehat{w}_j$ is $\mathcal{N}(0,1)$. This gives $\overline{\mathbb{E}}[(\frac{1}{p}\sum_{j\in[p]}\Xi_j^2)^2] \le C$ with high probability for a constant $C$, and hence we get the estimate $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\Xi_j^2] = o_{\mathbb{P}}(1)$. The same argument yields $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\widetilde{\Xi}_j^2] = o_{\mathbb{P}}(1)$.

Combined with the proximal representation of $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$ using $(\Xi_j, \widetilde{\Xi}_j)$, since $\mathsf{prox}_f(\cdot)$ is 1-Lipschitz for any convex function $f$, the upper bounds of $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\Xi_j^2]$ and $\overline{\mathbb{E}}[\frac{1}{p}\sum_{j\in[p]}\widetilde{\Xi}_j^2]$ yield the following simple proximal approximation of $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$:

$$\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}\Big[\Big\|\begin{pmatrix} \sqrt{p}h_j \\ \sqrt{p}\widetilde{h}_j \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)\widehat{w}_j; \nu^{-1}) - \theta_j \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{w}_j; \widetilde{\nu}^{-1}) - \theta_j \end{pmatrix}\Big\|^2\Big] = o_{\mathbb{P}}(1).$$

Noting $\overline{\mathbb{E}}[\|\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(1)$, this lets us approximate $\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]/\alpha\widetilde{\alpha}$ by the inner product of proximal operators:

$$\begin{aligned}\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}} + o_{\mathbb{P}}(1) &= \frac{1}{p}\sum_{j\in[p]}\frac{1}{\alpha\widetilde{\alpha}}\overline{\mathbb{E}}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)\widehat{w}_j; \nu^{-1}) - \theta_j\Big)\Big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{w}_j; \widetilde{\nu}^{-1}) - \theta_j\Big)\Big] \\ &= \frac{1}{p}\sum_{j\in[p]} F_{\mathsf{reg}}(\widehat{\eta}_H; \theta_j),\end{aligned}$$

where for the second inequality, we used the fact that the marginal law of $(\widehat{w}_j, \widetilde{w}_j)$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is jointly Gaussian, with zero mean, unit variance, and correlation $\widehat{\eta}_{_H}$, and where $F_{\mathsf{reg}}(\cdot;\theta_j) : [-1,1] \to \mathbb{R}$ is the function defined by

$$F_{\mathsf{reg}}(\eta;\theta_j) = \iint \frac{\left(\mathsf{prox}_{\mathsf{reg}}(\theta_j + \frac{\beta}{\nu}x; \frac{1}{\nu}) - \theta_j\right)\left(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + \frac{\widetilde{\beta}}{\widetilde{\nu}}(\eta x + \sqrt{1-\eta^2}y); \frac{1}{\widetilde{\nu}}) - \theta_j\right)}{\alpha\widetilde{\alpha}} \varphi(x)\varphi(y)\,\mathrm{d}x\,\mathrm{d}y$$

with $\varphi$ being the standard normal probability function and $\iint = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}$. Notice that for each $\eta \in [-1,1]$, the sequence $(F_{\mathsf{reg}}(\eta;\theta_j))_{j=1}^p$ are i.i.d. random variables with mean $\mathbb{E}[F_{\mathsf{reg}}(\eta;\theta_j)] = F_{\mathsf{reg}}(\eta)$. Furthermore, by the Jensen's inequality and the Cauchy–Schwarz inequality, the expectation of the absolute value is finite:

$$\begin{aligned}
\mathbb{E}\Big[|F_{\mathsf{reg}}(\eta;\theta_j)|\Big] &\le \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\big|\big(\mathsf{prox}_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu}H; \tfrac{1}{\nu}) - \Theta\big)\cdot\big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}; \tfrac{1}{\widetilde{\nu}}) - \Theta\big)\big|\Big] \\
&\le \frac{1}{\alpha\widetilde{\alpha}}\mathbb{E}\Big[\big(\mathsf{prox}_{\mathsf{reg}}(\Theta + \tfrac{\beta}{\nu}H; \tfrac{1}{\nu}) - \Theta\big)^2\Big]^{1/2}\cdot\mathbb{E}\Big[\big(\mathsf{prox}_{\widetilde{\mathsf{reg}}}(\Theta + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{H}; \tfrac{1}{\widetilde{\nu}}) - \Theta\big)^2\Big]^{1/2} \\
&= 1 \qquad \text{(by (4a) in System 1a)}.
\end{aligned}$$

Thus, the weak law of large numbers gives $p^{-1}\sum_{j\in[p]} F_{\mathsf{reg}}(\eta;\theta_j) \xrightarrow{\mathrm{p}} F_{\mathsf{reg}}(\eta)$ for each $\eta \in [-1,1]$. Next, we claim that this convergence holds uniformly over $\eta \in [-1,1]$.

LEMMA 19. *Let $I$ be a bounded and closed interval of $\mathbb{R}$ and let $(f_n)_{n\ge1} : I \to \mathbb{R}$ be a sequence of non-decreasing functions such that $f_n(x) \xrightarrow{\mathrm{p}} f(x)$ pointwise for some function $f : I \to \mathbb{R}$. If $f$ is non-decreasing and continuous, then the uniform convergence* $\sup_{x\in I}|f_n(x) - f(x)| \xrightarrow{\mathrm{p}} 0$ *holds.*

Note that Lemma 19 is a probabilistic analogue of a similar statement for deterministic functions: if a sequence of real-valued monotone functions (on $\mathbb{R}$) converges pointwise to a continuous function on a compact set $I \subset \mathbb{R}$, then the convergence is uniform on the set $I$. The probabilistic version is known but in a more general setting, so to keep the treatment self-contained we give a basic proof below which is similar to proof of the Glivenko-Cantelli theorem (cf. [98, Theorem 19.1]).

PROOF. Let us write $I = [a,b]$. Since $f$ is continuous and $I$ is compact, $f$ is uniformly continuous on $I$. For any $\epsilon > 0$, there exists some $\delta_\epsilon > 0$ such that $|f(x) - f(y)| < \epsilon/2$ for all $x, y \in I$ such that $|x - y| \le \delta_\epsilon$. Now for sufficiently large integer $k = k_\epsilon \in \mathbb{N}$ such that $(b-a)/k < \delta_\epsilon$, let us take equally spaced grids $(x_i)_{i=0}^k$ over $[a,b]$ such that $a = x_0 < x_1 < \cdots < x_k = b$ and $(x_i - x_{i-1}) = (b-a)/k$ for all $i \in \{0,\dots k\}$. Let $\Omega_\epsilon = \cap_{i=0}^k\{|f_n(x_i) - f(x_i)| \le \epsilon/2\}$ be the event under which $f_n$ and $f$ are sufficiently close at the finite grids. Note that this event holds with probability converging to 1, thanks to the pointwise convergence $f_n(x_i) \xrightarrow{\mathrm{p}} f(x_i)$ at finitely many $x_i$. Then, since $f_n$ is non-decreasing while $f$ does not move more than $\epsilon/2$ in $[x_{i-1}, x_i]$, for all $i \in \{0,1,\dots,k\}$ and for all $x \in [x_{i-1}, x_i]$ we have

$$\begin{aligned}
f_n(x_{i-1}) \le f_n(x) \le f_n(x_i), \\
-\epsilon/2 - f(x_{i-1}) \le -f(x) \le -f(x_i) + \epsilon/2.
\end{aligned}$$

In the event $\Omega_\epsilon$, summing the two lines it holds that $-\epsilon \le f_n(x) - f(x) \le \epsilon$ for all $i \in \{0,1,\dots,k\}$ and for all $x \in [x_{i-1}, x_i]$. $\square$

Recall that we have shown in the proof of Theorem 1 that $F_{\mathsf{reg}}$ is differentiable with a non-negative derivative. By the same argument, the map $\eta \mapsto F_{\mathsf{reg}}(\eta;\theta_j)$ is differentiable with a non-negative derivative. Thus, applying Lemma 19 with $f_n(\cdot) = p^{-1}\sum_{j\in[p]} F_{\mathsf{reg}}(\cdot;\theta_j)$, $f(\cdot) = F_{\mathsf{reg}}(\cdot)$ and $I = [-1,1]$, we get the uniform convergence:

$$\sup_{\eta\in[-1,1]} \Big|\frac{1}{p}\sum_{j\in[p]} F_{\mathsf{reg}}(\eta;\theta_j) - F_{\mathsf{reg}}(\eta)\Big| \xrightarrow{\mathrm{p}} 0.$$

Combined with $\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}} = \frac{1}{p}\sum_{j\in[p]} \widehat{F}_{\mathsf{reg}}(\widehat{\eta}_{{}_H};\theta_j) + o_{\mathbb{P}}(1)$ and $\widehat{\eta}_{{}_H} \in [-1,1]$, we are left with

$$\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}} = \frac{1}{p}\sum_{j\in[p]} F_{\mathsf{reg}}(\widehat{\eta}_{{}_H};\theta_j) + o_{\mathbb{P}}(1) = F_{\mathsf{reg}}(\widehat{\eta}_{{}_H}) + o_{\mathbb{P}}(1).$$

Recall the definition $\widehat{\eta}_{{}_G} = \Pi_{[-1,1]}(\frac{\overline{\mathbb{E}}[\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}]}{\alpha\widetilde{\alpha}})$ where $\Pi_{[-1,1]}$ is the projection onto $[-1,1]$. By the continuity of the projection map and the bound $\sup_{\eta\in[-1,1]}|F_{\mathsf{reg}}(\eta)| \le 1$ (see Theorem 1), the above display yields

$$\widehat{\eta}_{{}_G} = \Pi_{[-1,1]}(F_{\mathsf{reg}}(\widehat{\eta}_{{}_H})) + o_{\mathbb{P}}(1) = F_{\mathsf{reg}}(\widehat{\eta}_{{}_H}) + o_{\mathbb{P}}(1).$$

Next, let us show $\widehat{\eta}_{{}_H} = F_{\mathsf{loss}}(\widehat{\eta}_{{}_G}) + o_{\mathbb{P}}(1)$ using the same argument. Using Lemma 18 with $\boldsymbol{F} = [\boldsymbol{h}/\alpha \mid \widetilde{\boldsymbol{h}}/\widetilde{\alpha}] \in \mathbb{R}^{p\times 2}$ and $\boldsymbol{z} = \boldsymbol{g}_i$, there exists a random vector $\boldsymbol{u}_i \in \mathbb{R}^2$ with conditional distribution

$$\boldsymbol{u}_i \mid \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{G}_{-i}, I, \widetilde{I} \stackrel{\mathrm{d}}{=} \mathcal{N}(\boldsymbol{0}_2, \boldsymbol{I}_2)$$

and

$$\overline{\mathbb{E}}\Big[\Big\|\begin{pmatrix}(\boldsymbol{g}_i^\top\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\psi_i + \boldsymbol{h}^\top\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{e}_i)/\alpha\\ (\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\psi}_i + \widetilde{\boldsymbol{h}}^\top\widetilde{\boldsymbol{A}}\boldsymbol{G}^\top\widetilde{\boldsymbol{D}}\boldsymbol{e}_i)/\widetilde{\alpha}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha}\\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2}\boldsymbol{u}_i\Big\|^2\Big] \le C_6\sum_{j=1}^p \overline{\mathbb{E}}\frac{1}{\alpha^2}\Big\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\Big\|^2 + \frac{1}{\widetilde{\alpha}^2}\Big\|\frac{\partial\widetilde{\boldsymbol{h}}}{\partial g_{ij}}\Big\|^2. \tag{37}$$

Using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ in (35) Lemma 14, noting $\mu^{-2} = o(n)$, it follows that

$$\overline{\mathbb{E}}\big[\|\boldsymbol{h}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|^2\big] = \mathcal{O}_{\mathbb{P}}(n^{-1}\mu^{-2}) = o_{\mathbb{P}}(1)$$

$$\sum_{i\in[n]}\sum_{j\in[p]}\overline{\mathbb{E}}\Big[\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\|^2\Big] \le 2\overline{\mathbb{E}}\Big[\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + \|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2\|\boldsymbol{h}\|^2\Big] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}) = o_{\mathbb{P}}(n)$$

so that summing over $i\in[n]$ the inequality (37), we get

$$\sum_{i\in[n]}\Big\|\begin{pmatrix}(\boldsymbol{g}_i^\top\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\psi_i)/\alpha\\ (\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\widetilde{\psi}_i)/\widetilde{\alpha}\end{pmatrix} - \begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha}\\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2}\boldsymbol{u}_i\Big\|^2 = o_{\mathbb{P}}(n).$$

Appealing to the $(1/2)$-Hölder continuity for the matrix square root again, now using the two convergences $\|\boldsymbol{h}\|^2 \to \alpha^2$, $\|\widetilde{\boldsymbol{h}}\|^2 \xrightarrow{\mathrm{p}} \widetilde{\alpha}^2$ and the concentration $\widehat{\eta}_{{}_G} = \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) + o_{\mathbb{P}}(1)$, we get

$$\Big\|\begin{pmatrix}\|\boldsymbol{h}\|^2/\alpha^2 & \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}/\alpha\widetilde{\alpha}\\ \widetilde{\boldsymbol{h}}^\top\boldsymbol{h}/\alpha\widetilde{\alpha} & \|\widetilde{\boldsymbol{h}}\|^2/\widetilde{\alpha}^2\end{pmatrix}^{1/2} - \begin{pmatrix}1 & \widehat{\eta}_{{}_G}\\ \widehat{\eta}_{{}_G} & 1\end{pmatrix}^{1/2}\Big\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Combined with the convergence $\operatorname{tr}[\boldsymbol{A}] \xrightarrow{\mathrm{p}} \kappa, \operatorname{tr}[\widetilde{\boldsymbol{A}}] \xrightarrow{\mathrm{p}} \widetilde{\kappa}$, noting that $\sum_{i\in[n]} \|\boldsymbol{u}_i\|^2$ and $\|\boldsymbol{\psi}\|^2 + \|\widetilde{\boldsymbol{\psi}}\|^2$ are both $\mathcal{O}_{\mathbb{P}}(n)$, we are left with

$$\frac{1}{n}\sum_{i\in[n]}\left\|\begin{pmatrix}\boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i\\ \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i\end{pmatrix} - \begin{pmatrix}\alpha\widehat{u}_i\\ \widetilde{\alpha}\widetilde{u}_i\end{pmatrix}\right\|^2 = o_{\mathbb{P}}(1) \quad \text{where} \quad \begin{pmatrix}\widehat{u}_i\\ \widetilde{u}_i\end{pmatrix} = \begin{pmatrix}1 & \widehat{\eta}_{G}\\ \widehat{\eta}_{G} & 1\end{pmatrix}^{1/2}\boldsymbol{u}_i.$$

By the same argument that we used to bound $\Xi_j$ and $\widetilde{\Xi}_j$, using Lemma 14 and the fact that the marginal law of $\widehat{u}_i$ (and $\widetilde{u}_i$) is $\mathcal{N}(0,1)$, we can show that the conditional expectation $\overline{\mathbb{E}}$ of the square of LHS is bounded from above by a constant $C$ with high probability. Thus, the above approximation also holds in the conditional expectation $\overline{\mathbb{E}}$:

$$\frac{1}{n}\sum_{i\in[n]}\overline{\mathbb{E}}\Big[\left\|\begin{pmatrix}\boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i\\ \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i\end{pmatrix} - \begin{pmatrix}\alpha\widehat{u}_i\\ \widetilde{\alpha}\widetilde{u}_i\end{pmatrix}\right\|^2\Big] = o_{\mathbb{P}}(1).$$

Let us define $\Xi^i := \boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i - \alpha\widehat{u}_i$ and $\widetilde{\Xi}^i := \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \widetilde{\kappa}\widetilde{\psi}_i - \widetilde{\alpha}\widetilde{u}_i$ so that the above display reads $\sum_{i\in[n]} \overline{\mathbb{E}}[(\Xi^i)^2 + (\widetilde{\Xi}^i)^2] = o(n)$. Since $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ and $\widetilde{\psi}_i = \widetilde{\mathsf{loss}}'(\widetilde{z}_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}})$ for all $i \in I\cap\widetilde{I}$, the residuals can be written as

$$z_i - \boldsymbol{g}_i^\top \boldsymbol{h} = \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i - \Xi^i;\kappa), \quad z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} = \mathsf{prox}_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i - \widetilde{\Xi}^i;\widetilde{\kappa})$$

for all $i\in I\cap\widetilde{I}$. Since the map $x \mapsto \mathsf{env}'_f(x;\tau) = f' \circ \mathsf{prox}_f(x;\tau)$ is a composition of Lipschitz functions if $f$ is convex and differentiable with Lipschitz derivative, the moment bound $\sum_{i\in[n]} \overline{\mathbb{E}}[(\Xi^i)^2 + (\widetilde{\Xi}^i)^2] = o(n)$ lets us approximate $\psi_i$ and $\widetilde{\psi}_i$ by $\mathsf{env}'_{\mathsf{loss}}$ and $\mathsf{env}'_{\widetilde{\mathsf{loss}}}$ as follows:

$$\frac{1}{n}\sum_{i\in I\cap\widetilde{I}}\overline{\mathbb{E}}\left\|\begin{pmatrix}\psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\\ \widetilde{\psi}_i - \mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i;\widetilde{\kappa})\end{pmatrix}\right\|^2 = o_{\mathbb{P}}(1) \quad \text{with} \quad \begin{pmatrix}\widehat{u}_i\\ \widetilde{u}_i\end{pmatrix} | \boldsymbol{z},\boldsymbol{\theta},\boldsymbol{G}_{-i}, I, \widetilde{I} \stackrel{\mathrm{d}}{=} \mathcal{N}(0_2, \begin{pmatrix}1 & \widehat{\eta}_{G}\\ \widehat{\eta}_{G} & 1\end{pmatrix})$$

Noting $\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$ by Lemma 14, this lets us approximate $\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]/(p\beta\widetilde{\beta})$ by the inner product of $(\mathsf{env}'_{\mathsf{loss}}, \mathsf{env}'_{\widetilde{\mathsf{loss}}})$:

$$\begin{aligned}\frac{\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]}{p\beta\widetilde{\beta}} &= \frac{|I\cap\widetilde{I}|}{p\beta\widetilde{\beta}}\frac{1}{|I\cap\widetilde{I}|}\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]\\ &= \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\frac{1}{|I\cap\widetilde{I}|}\sum_{i\in I\cap\widetilde{I}}\overline{\mathbb{E}}[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\cdot \mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i - \widetilde{\alpha}\widetilde{u}_i;\widetilde{\kappa})] + o_{\mathbb{P}}(1).\end{aligned}$$

Since the marginal distribution of $(u_i, \widetilde{u}_i)$ given $(\boldsymbol{\theta}, \boldsymbol{z}, I, \widetilde{I})$ is centered normal with unit variance and correlation $\widehat{\eta}_{G}$, the above display reads

$$\frac{\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\widetilde{\boldsymbol{\psi}}]}{p\beta\widetilde{\beta}} = \frac{1}{|I\cap\widetilde{I}|}\sum_{i\in I\cap\widetilde{I}} F_{\mathsf{loss}}(\widehat{\eta}_{G}; z_i) + o_{\mathbb{P}}(1)$$

where $F_{\mathsf{loss}}(\eta; z_i) : [-1,1] \to \mathbb{R}$ is the function defined as:

$$F_{\mathsf{loss}}(\eta; z_i) = \frac{c\widetilde{c}\delta}{\beta\widetilde{\beta}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\varphi(x)\varphi(y)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x;\kappa)\mathsf{env}'_{\widetilde{\mathsf{loss}}}(z_i + \widetilde{\alpha}(\eta x + \sqrt{1-\eta^2}y);\widetilde{\kappa})\,\mathrm{d}x\,\mathrm{d}y.$$

By the same argument for $F_{\mathsf{reg}}(\eta;\theta_j)$, the sequence $(F_{\mathsf{loss}}(\eta; z_i))_{i\in[n]}$ are i.i.d. random variables with mean $\mathbb{E}[F_{\mathsf{loss}}(\eta; z_i)] = F_{\mathsf{loss}}(\eta)$ and the expectation of the absolute value $|F_{\mathsf{loss}}(\eta; z_i)|$ is finite. Thus, by the weak law of large numbers, we have $\frac{1}{m}\sum_{i\in[m]} F_{\mathsf{loss}}(\eta; z_i) \xrightarrow{\mathrm{p}}$

$\mathbb{E}[F_{\mathsf{loss}}(\eta; z_1)] = F_{\mathsf{loss}}(\eta)$ for any deterministic integer $m = m_n \to +\infty$. Then, we have that for any $\epsilon > 0$, denoting by $\sum_K$ the sum over all possible value $K$ taken by $I \cap \widetilde{I}$,

$$\begin{aligned}
&\mathbb{P}\Big(|\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \\
&= \sum_K \mathbb{P}\Big(|\frac{1}{|K|} \sum_{i \in K} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon,\ I \cap \widetilde{I} = K\Big) \qquad \text{(by additivity of disjoint events)} \\
&= \sum_K \mathbb{P}\Big(|\frac{1}{|K|} \sum_{i \in K} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \mathbb{P}(I \cap \widetilde{I} = K) \qquad \text{(by independence)} \\
&= \sum_{m \geq 0} \mathbb{P}\Big(|\frac{1}{m} \sum_{i \in [M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \mathbb{P}(|I \cap \widetilde{I}| = m) \quad \text{(since } (z_i)_{i \in K} \overset{\mathrm{d}}{=} (z_i)_{i \in [m]} \text{ for } m = |K|\text{).}
\end{aligned}$$

We now split the sum over $m \geq 0$ into two as follows:

$$\begin{aligned}
&\mathbb{P}\Big(|\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \\
&= \Big(\sum_{m \leq n c \widetilde{c}/2} + \sum_{m > n c \widetilde{c}/2}\Big) \mathbb{P}\Big(|\frac{1}{m} \sum_{i \in [M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \mathbb{P}\Big(|I \cap \widetilde{I}| = m\Big) \\
&\leq \mathbb{P}(|I \cap \widetilde{I}| \leq n c \widetilde{c}/2) + \sup_{m > n c \widetilde{c}/2} \mathbb{P}\Big(|\frac{1}{m} \sum_{i \in [M]} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta)| > \epsilon\Big) \mathbb{P}(|I \cap \widetilde{I}| > n c \widetilde{c}/2)
\end{aligned}$$

where $\mathbb{P}(|I \cap \widetilde{I}| > n c \widetilde{c}/2) \leq 1$ in the rightmost term. The second term converges to 0 by the weak law of large numbers, and the first by Chebyshev's inequality applied to the hypergeometric distribution.

This gives $\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) \xrightarrow{\mathrm{p}} F_{\mathsf{loss}}(\eta)$ pointwise for any $\eta \in [-1, 1]$. By the same argument we used for $F_{\mathsf{reg}}$ and $F_{\mathsf{reg}}(\cdot; \theta_j)$, $F_{\mathsf{loss}}$ and $F_{\mathsf{loss}}(\cdot; z_i)$ are non-decreasing and continuous. Thus, we can apply Lemma 19 with $f_n(\cdot) = \frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\cdot; z_i)$, $f(\cdot) = F_{\mathsf{loss}}(\cdot)$ and obtain the uniform convergence:

$$\sup_{\eta \in [-1,1]} \Big| \frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\eta; z_i) - F_{\mathsf{loss}}(\eta) \Big| = o_{\mathbb{P}}(1).$$

Combined with $\frac{\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]}{p \beta \widetilde{\beta}} = \frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} F_{\mathsf{loss}}(\widehat{\eta}_G; z_i)$ and $\widehat{\eta}_G \in [-1, 1]$, we are left with

$$\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]/(p \beta \widetilde{\beta}) = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1).$$

Recalling $\widehat{\eta}_H = \Pi_{[-1,1]}(\bar{\mathbb{E}}[\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}]/(p \beta \widetilde{\beta}))$ where $\Pi_{[-1,1]}$ is the projection onto $[-1, 1]$, by the continuity of the projection map and $|F_{\mathsf{loss}}(\eta)| \leq \sqrt{c \widetilde{c}} \leq 1$ for all $\eta \in [-1, 1]$ (see Theorem 1), we finally obtain

$$\widehat{\eta}_H = \Pi_{[-1,1]}(F_{\mathsf{loss}}(\widehat{\eta}_G)) + o_{\mathbb{P}}(1) = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1).$$

In summary, we have shown that $\widehat{\eta}_G = F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1)$ and $\widehat{\eta}_H = F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1)$. By the continuity of $F_{\mathsf{reg}}$ and $F_{\mathsf{loss}}$, it holds that

$$\widehat{\eta}_G = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\widehat{\eta}_G) + o_{\mathbb{P}}(1) \quad \text{and} \quad \widehat{\eta}_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1).$$

Let $T : [-1,1] \to \mathbb{R}$ be the map $T : \eta \mapsto \eta - F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta)$ so that the above result (for $\widehat{\eta}_G$) reads to $T(\widehat{\eta}_G) = o_{\mathbb{P}}(1)$. Since $T$ is continuous in $[-1,1]$ and has a unique root $\eta_G$ by Theorem 1, for any $\epsilon > 0$ let $C(\epsilon) = \min_{u \in [-1,1] \setminus [\eta_G^\star - \epsilon, \eta_G^\star + \epsilon]} T(u)$ and note $C(\epsilon) > 0$ by compactness. Hence

$$\mathbb{P}(|\widehat{\eta}_G - \eta_G^\star| \geq \epsilon) \leq \mathbb{P}(T(\widehat{\eta}_G) \geq C(\epsilon)) \to 0$$

due to $T(\widehat{\eta}_G) = o_{\mathbb{P}}(1)$. By the same argument, combining $\widehat{\eta}_H = F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\widehat{\eta}_H) + o_{\mathbb{P}}(1)$ and the fact that the map $\eta \mapsto \eta - F_{\mathsf{loss}} \circ F_{\mathsf{reg}}(\eta)$ is continuous and has a unique root $\eta_H$ by Theorem 1, we obtain $\widehat{\eta}_H \xrightarrow{\mathrm{p}} \eta_H$.

Recalling $\widehat{\eta}_G = \boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) + o_{\mathbb{P}}(1)$ and $\widehat{\eta}_H = \boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) + o_{\mathbb{P}}(1)$, this gives $\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha\widetilde{\alpha}) \xrightarrow{\mathrm{p}} \eta_G$ and $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_H$, thereby completing the proof.

Finally, let us show the proximal approximation of estimators $(\sqrt{p}h_j, \sqrt{p}\widetilde{h}_j)$ and residuals $(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}, z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}})$. Now that we have established $\boldsymbol{\psi}^\top \widetilde{\boldsymbol{\psi}}/(p\beta\widetilde{\beta}) \xrightarrow{\mathrm{p}} \eta_H$ and $\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}/(\alpha, \widetilde{\alpha}) \to^p \eta_G$, using the above argument again with $\widehat{\eta}_H$ replaced by $\eta_H$ and $\widehat{\eta}_G$ by $\eta_G$, we obtain the following result.

COROLLARY 20. *Denote by $\overline{\mathbb{E}}[\cdot | \boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I}]$ be the conditional expectation. Then, we have the following (conditional) proximal representation of the error vectors and residuals:*

$$\frac{1}{p}\sum_{j\in[p]} \overline{\mathbb{E}}\Big[\Big\| \begin{pmatrix} \sqrt{p}h_j + \theta_j \\ \sqrt{p}\widetilde{h}_j + \theta_j \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)H_j; \nu^{-1}) \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{H}_j; \widetilde{\nu}^{-1}) \end{pmatrix} \Big\|^2\Big] = o_{\mathbb{P}}(1),$$

$$\frac{1}{|I \cap \widetilde{I}|}\sum_{i\in I\cap\widetilde{I}} \overline{\mathbb{E}}\Big[\Big\| \begin{pmatrix} z_i - \boldsymbol{g}_i^\top \boldsymbol{h} \\ z_i - \boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{loss}}(z_i + \alpha G_i; \kappa) \\ \mathsf{prox}_{\widetilde{\mathsf{loss}}}(z_i + \alpha \widetilde{G}_i; \widetilde{\kappa}) \end{pmatrix} \Big\|^2\Big] = o_{\mathbb{P}}(1),$$

*where $(H_j, \widetilde{H}_j)_{j\in[p]}$ and $(G_i, \widetilde{G}_i)_{i\in I\cap\widetilde{I}}$ are jointly normals such that:*

$$\forall j \in [p] \quad \begin{pmatrix} H_j \\ \widetilde{H}_j \end{pmatrix} | \boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I} \stackrel{\mathrm{d}}{=} \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_H \\ \eta_H & 1 \end{bmatrix} \right),$$

$$\forall i \in I \cap \widetilde{I}, \quad \begin{pmatrix} G_i \\ \widetilde{G}_i \end{pmatrix} | \boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I} \stackrel{\mathrm{d}}{=} \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta_G \\ \eta_G & 1 \end{bmatrix} \right).$$

Since $|X_n| = o_{\mathbb{P}}(1)$ is equivalent to $\mathbb{E}[1 \wedge |X_n|] = o(1)$ for any random variable $X_n$, the conditional proximal representation of the error vectors $(\boldsymbol{h}, \widetilde{\boldsymbol{h}})$ in Corollary 20 reads

$$\mathbb{E}\Big[1 \wedge \frac{1}{p}\sum_{j\in[p]} \overline{\mathbb{E}}[\mathsf{A}_j]\Big] = o(1) \text{ where } \mathsf{A}_j := \Big\| \begin{pmatrix} \sqrt{p}h_j + \theta_j \\ \sqrt{p}\widetilde{h}_j + \theta_j \end{pmatrix} - \begin{pmatrix} \mathsf{prox}_{\mathsf{reg}}(\theta_j + (\beta/\nu)H_j; \nu^{-1}) \\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j + (\widetilde{\beta}/\widetilde{\nu})\widetilde{H}_j; \widetilde{\nu}^{-1}) \end{pmatrix} \Big\|^2$$

Here the expectation $\mathbb{E}$ is with respect to $(\boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I})$ (recall the definition $\overline{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot | \boldsymbol{z}, \boldsymbol{\theta}, I, \widetilde{I}]$). Note that the integrand is bounded from below as

$$1 \wedge \frac{1}{p}\sum_{j\in[p]} \overline{\mathbb{E}}[\mathsf{A}_j] \geq \frac{1}{p}\sum_{j\in[p]} \Big(1 \wedge \overline{\mathbb{E}}[\mathsf{A}_j]\Big) \geq \frac{1}{p}\sum_{j\in[p]} \overline{\mathbb{E}}[1 \wedge \mathsf{A}_j],$$

where the second inequality follows from the Jensen's inequality $\overline{\mathbb{E}}[f(X)] \leq f(\overline{\mathbb{E}}[X])$ applied with the concave function $f(x) = 1 \wedge x$ and $X = \mathsf{A}_j$. Taking the expectation of the above display and using the tower property, we are left with

$$\frac{1}{p}\sum_{j\in[p]}\mathbb{E}[1\wedge \mathsf{A}_j]=o(1).$$

Since the marginal distribution of the integrand $(1\wedge \mathsf{A}_j)$ is the same for all $j\in[p]$ by symmetry, the LHS equals to $\mathbb{E}[1\wedge \mathsf{A}_{j'}]$ for any $j'\in[p]$. This gives $\max_{j\in[p]}\mathbb{E}[1\wedge \mathsf{A}_j]=o(1)$ and completes the proof of the desired joint approximation of estimators $(\sqrt{p}h_j,\sqrt{p}\widetilde{h}_j)$. The joint approximation of residual follows from the same argument, so we omit the proof.

**B.3. Proof of Theorem 3.** The proof strategy is the same as Appendix B.2.4. We first claim the average $p^{-1}\sum_{j\in[p]}\Phi(\cdots)$ concentrates on its conditional expectation:

$$\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j+\theta_j\\ \sqrt{p}\widetilde{h}_j+\theta_j\\ \theta_j\end{pmatrix}=\frac{1}{p}\sum_{j\in[p]}\bar{\mathbb{E}}\Phi\begin{pmatrix}\sqrt{p}h_j+\theta_j\\ \sqrt{p}\widetilde{h}_j+\theta_j\\ \theta_j\end{pmatrix}+o_{\mathbb{P}}(1),\tag{38}$$

where recall $\bar{\mathbb{E}}[\cdot]=\mathbb{E}[\cdot|\boldsymbol{\theta},\boldsymbol{z},I,\widetilde{I}]$. Note that the marginal distribution of $\boldsymbol{\theta}=(\theta_j)_{j\in[p]}$ is bounded in the second moment by the additional assumption we imposed for this theorem. Furthermore, Lemma 14-(2) implies that there exists a deterministic constant $C$ which depends on $(\delta,\|\mathsf{loss}'\|_{\mathrm{Lip}},\|\widetilde{\mathsf{loss}}'\|_{\mathrm{Lip}},\alpha,\widetilde{\alpha},\mathbb{E}[\mathsf{loss}'(Z)^2],\mathbb{E}[\widetilde{\mathsf{loss}}'(Z)^2])$ only such that $\|\boldsymbol{h}\|^2\vee\|\widetilde{\boldsymbol{h}}\|^2\le C$ with probability 1. Then, combined with Corollary 20, using the pseudo-Lipschitz property of the test function $\Phi$, we have

$$\frac{1}{p}\sum_{j\in[p]}\bar{\mathbb{E}}\Phi\begin{pmatrix}\sqrt{p}h_j+\theta_j\\ \sqrt{p}\widetilde{h}_j+\theta_j\\ \theta_j\end{pmatrix}=\frac{1}{p}\sum_{j\in[p]}\bar{\mathbb{E}}\Phi\begin{pmatrix}\mathsf{prox}_{\mathsf{reg}}(\theta_j+(\beta/\nu)H_j;\nu^{-1})\\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j+(\widetilde{\beta}/\widetilde{\nu})\widetilde{H}_j;\widetilde{\nu}^{-1})\\ \theta_j\end{pmatrix}+o_{\mathbb{P}}(1),$$

By the law of large numbers,

$$\frac{1}{p}\sum_{j\in[p]}\bar{\mathbb{E}}\Phi\begin{pmatrix}\mathsf{prox}_{\mathsf{reg}}(\theta_j+(\beta/\nu)H_j;\nu^{-1})\\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\theta_j+(\widetilde{\beta}/\widetilde{\nu})\widetilde{H}_j;\widetilde{\nu}^{-1})\\ \theta_j\end{pmatrix}\xrightarrow{\mathrm{p}}\mathbb{E}\Phi\begin{pmatrix}\mathsf{prox}_{\mathsf{reg}}(\Theta+(\beta/\nu)H;\nu^{-1})\\ \mathsf{prox}_{\widetilde{\mathsf{reg}}}(\Theta+(\widetilde{\beta}/\widetilde{\nu})\widetilde{H};\widetilde{\nu}^{-1})\\ \Theta\end{pmatrix},$$

where the integrand on the RHS is bounded in $L_1$ by Assumption D-(1), the pseudo-Lipschitz continuity of order 2 of $\Phi$, and the additional assumption that $\mathbb{E}[\Theta^2]<+\infty$ where $\Theta\sim F_\theta$.

Thus, it remains to show the claim (38). Recall that we have proved this result for the specific test function $\Phi(a,b,c)=(a-c)(b-c)$ in Lemma 17 by the Gaussian Poincaré inequality. Here, we use the same proof strategy. One minor thing we need to be careful of is that we do not assume the differentiability of the test function $\Phi$, so instead of calculating the derivative of $p^{-1}\sum_{j\in[p]}\Phi(\cdots)$, we prove its locally Lipschitz property. Specifically, we claim that for all $\boldsymbol{G},\boldsymbol{G}'\in\mathbb{R}^{n\times p}$, letting $\boldsymbol{h}=\boldsymbol{h}(\boldsymbol{G})$ be the estimator fitted on the design matrix $\boldsymbol{G}$,

$$\begin{aligned}&\left|\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j(\boldsymbol{G})+\theta_j\\ \sqrt{p}\widetilde{h}_j(\boldsymbol{G})+\theta_j\\ \theta_j\end{pmatrix}-\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j(\boldsymbol{G}')+\theta_j\\ \sqrt{p}\widetilde{h}_j(\boldsymbol{G}')+\theta_j\\ \theta_j\end{pmatrix}\right|^2\\ &\le C_*(n\mu_n^2)^{-1}\Big(1+\frac{\|\boldsymbol{\theta}\|^2}{p}+\frac{\|\mathsf{loss}'(\boldsymbol{z})\|^2}{n}+\frac{\|\widetilde{\mathsf{loss}}'(\boldsymbol{z})\|^2}{n}+\frac{\|\boldsymbol{G}\|_{\mathrm{op}}^2}{n}\Big)\|\boldsymbol{G}-\boldsymbol{G}'\|_{\mathrm{op}}^2,\end{aligned}\tag{39}$$

where $C_*$ is a deterministic constant. If this inequality holds, then the squared Frobenius norm of the weak derivative $\|\nabla_{\boldsymbol{G}}\frac{1}{p}\sum_{j\in[p]}\Phi(\cdots)\|_F^2$ is bounded from above by

$$C_*(n\mu_n^2)^{-1}\Big(1+\frac{\|\boldsymbol{\theta}\|^2}{p}+\frac{\|\mathsf{loss}'(\boldsymbol{z})\|^2}{n}+\frac{\|\widetilde{\mathsf{loss}}'(\boldsymbol{z})\|^2}{n}+\frac{\|\boldsymbol{G}\|_{\mathrm{op}}^2}{n}\Big),$$

so that combined with the Gaussian Poincaré inequality, we have

$$\begin{aligned}
&\overline{\mathbb{E}}\left(\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j+\theta_j\\ \sqrt{p}\widetilde{h}_j+\theta_j\\ \theta_j\end{pmatrix}-\frac{1}{p}\sum_{j\in[p]}\overline{\mathbb{E}}\Phi\begin{pmatrix}\sqrt{p}h_j+\theta_j\\ \sqrt{p}\widetilde{h}_j+\theta_j\\ \theta_j\end{pmatrix}\right)^2\\
&\le C_*(n\mu_n^2)^{-1}\Big(1+\frac{\|\boldsymbol{\theta}\|^2}{p}+\frac{\|\mathsf{loss}'(\boldsymbol{z})\|^2}{n}+\frac{\|\widetilde{\mathsf{loss}}'(\boldsymbol{z})\|^2}{n}+\frac{\overline{\mathbb{E}}[\|\boldsymbol{G}\|_{\mathrm{op}}^2]}{n}\Big)\\
&=\mathcal{O}_{\mathbb{P}}(n\mu_n^2)^{-1}.
\end{aligned}$$

For the last equality, we have used $\|\boldsymbol{\theta}\|_2^2/p=\mathcal{O}_{\mathbb{P}}(1)$ and $\|f'(\boldsymbol{z})\|_2^2/n=\mathcal{O}_{\mathbb{P}}(1)$ by the law of large number with the assumptions $\mathbb{E}[\Theta^2]<+\infty$ and $\mathbb{E}[f'(Z)^2]<+\infty$ for $f\in\{\mathsf{loss},\widetilde{\mathsf{loss}}\}$. Finally, using $\mu_n^{-1}=o(n^{1/2})$, we complete the proof of (38). Thus, it suffices to show the locally Lipschitz property (39).

By the pseudo-Lipschitz continuity of order 2 of the test function $\Phi$, there exists a constant $C$ (which depends on $\Phi$ only) such that

$$\begin{aligned}
&\left|\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j(\boldsymbol{G})+\theta_j\\ \sqrt{p}\widetilde{h}_j(\boldsymbol{G})+\theta_j\\ \theta_j\end{pmatrix}-\frac{1}{p}\sum_{j\in[p]}\Phi\begin{pmatrix}\sqrt{p}h_j(\boldsymbol{G}')+\theta_j\\ \sqrt{p}\widetilde{h}_j(\boldsymbol{G}')+\theta_j\\ \theta_j\end{pmatrix}\right|\\
&\le\frac{C}{p}\sum_{j\in[p]}\Big(1+|\theta_j|+\sqrt{p}\big(|h_j(\boldsymbol{G})|+|h_j(\boldsymbol{G}')|+|\widetilde{h}_j(\boldsymbol{G})|+|\widetilde{h}_j(\boldsymbol{G}')|\big)\Big)\\
&\qquad\qquad\times\sqrt{p}\big(|h_j(\boldsymbol{G})-h_j(\boldsymbol{G}')|+|\widetilde{h}_j(\boldsymbol{G})-\widetilde{h}_j(\boldsymbol{G}')|\big)\\
&=C\sum_{j\in[p]}v_j(\boldsymbol{G},\boldsymbol{G}')\big(|h_j(\boldsymbol{G})-h_j(\boldsymbol{G}')|+|\widetilde{h}_j(\boldsymbol{G})-\widetilde{h}_j(\boldsymbol{G}')|\big)
\end{aligned}$$

$$\le C\|\boldsymbol{v}(\boldsymbol{G},\boldsymbol{G}')\|\big(\|\boldsymbol{h}(\boldsymbol{G})-\boldsymbol{h}(\boldsymbol{G}')\|+\|\widetilde{\boldsymbol{h}}(\boldsymbol{G})-\widetilde{\boldsymbol{h}}(\boldsymbol{G}')\|\big)\tag{40}$$

where $\boldsymbol{v}(\boldsymbol{G},\boldsymbol{G}')=(v_j(\boldsymbol{G},\boldsymbol{G}'))_{j\in[p]}\in\mathbb{R}^p$ is defined as

$$v_j(\boldsymbol{G},\boldsymbol{G}'):=p^{-1/2}(1+|\theta_j|)+|h_j(\boldsymbol{G})|+|h_j(\boldsymbol{G}')|+|\widetilde{h}_j(\boldsymbol{G})|+|\widetilde{h}_j(\boldsymbol{G}')|.$$

Let us bound $\|\boldsymbol{v}(\boldsymbol{G},\boldsymbol{G}')\|$. Now, from Lemma 14-(2), we can take a deterministic constant $C_*$, which depends on $(\delta,\|\mathsf{loss}'\|_{\mathrm{Lip}},\|\widetilde{\mathsf{loss}}'\|_{\mathrm{Lip}},\alpha,\widetilde{\alpha},\mathbb{E}[\mathsf{loss}'(Z)^2],\mathbb{E}[\widetilde{\mathsf{loss}}'(Z)^2])$ only, such that

$$\|\boldsymbol{h}(\boldsymbol{G})\|^2\vee\|\widetilde{\boldsymbol{h}}(\boldsymbol{G})\|^2\le C_*\text{ and }\begin{cases}\|\boldsymbol{\psi}(\boldsymbol{G})\|^2\le C_*(\|\mathsf{loss}'(\boldsymbol{z})\|^2+\|\boldsymbol{G}\|_{\mathrm{op}}^2)\\ \|\widetilde{\boldsymbol{\psi}}(\boldsymbol{G})\|^2\le C_*(\|\widetilde{\mathsf{loss}}'(\boldsymbol{z})\|^2+\|\boldsymbol{G}\|_{\mathrm{op}}^2)\end{cases}\tag{41}$$

uniformly for all $\boldsymbol{G}\in\mathbb{R}^{n\times p}$. Thus, the upper bound $\|\boldsymbol{h}(\boldsymbol{G})\|^2\vee\|\widetilde{\boldsymbol{h}}(\boldsymbol{G})\|^2\le C_*$ yields

$$\|\boldsymbol{v}(\boldsymbol{G},\boldsymbol{G}')\|^2\le C(1+\|\boldsymbol{\theta}\|^2/p).$$

for a constant $C$ depending on $C_*$ only.

Next, let us control $\|\boldsymbol{h}(\boldsymbol{G})-\boldsymbol{h}(\boldsymbol{G}')\|$. By [11, Proposition 4.4], we know that $\boldsymbol{G}\mapsto\boldsymbol{h}(\boldsymbol{G})$ is locally Lipschitz as follows:

$$\|\boldsymbol{h}(\boldsymbol{G})-\boldsymbol{h}(\boldsymbol{G}')\|^2\le(n\mu_n^2)^{-1}\|\boldsymbol{G}-\boldsymbol{G}'\|_{\mathrm{op}}^2(\frac{\|\boldsymbol{\psi}(\boldsymbol{G})\|_2^2}{n}+\|\boldsymbol{h}(\boldsymbol{G})\|_2^2).$$

Combined with (41), we obtain

$$\|\boldsymbol{h}(\boldsymbol{G}) - \boldsymbol{h}(\boldsymbol{G}')\|^2 \le C_*(n\mu_n^2)^{-1}\Big(1 + \frac{\|\mathsf{loss}'(\boldsymbol{z})\|^2}{n} + \frac{\|\boldsymbol{G}\|_{\mathrm{op}}^2}{n}\Big)\|\boldsymbol{G} - \boldsymbol{G}'\|_{\mathrm{op}}^2.$$

Similarly, applying [11, Proposition 4.4] for $\boldsymbol{G} \mapsto \widetilde{\boldsymbol{h}}(\boldsymbol{G})$ and using (41), we get the same upper bound for $\|\widetilde{\boldsymbol{h}}(\boldsymbol{G}) - \widetilde{\boldsymbol{h}}(\boldsymbol{G}')\|^2$ with $\mathsf{loss}'$ on the RHS replaced by $\widetilde{\mathsf{loss}}'$.

Finally, substituting these upper bounds of $\|\boldsymbol{v}(\boldsymbol{G},\boldsymbol{G}')\|^2$, $\|\boldsymbol{h}(\boldsymbol{G}) - \boldsymbol{h}(\boldsymbol{G}')\|^2$, and $\|\widetilde{\boldsymbol{h}}(\boldsymbol{G}) - \widetilde{\boldsymbol{h}}(\boldsymbol{G}')\|^2$ into (40), we get the desired locally Lipschitz property (39).

The convergence of $|I \cap \widetilde{I}|^{-1}\sum_{i\in I\cap\widetilde{I}} \Phi(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\theta}}_I, y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\theta}}_{\widetilde{I}}, z_i)$ follows from the same argument, so we omit the proof.

**B.4. Proof of Theorem 5.** We assume that reg and $\widetilde{\text{reg}}$ are $\mu$-strongly convex for a fixed $\mu > 0$. Then [17, Theorem 5.1] implies that

$$\mathrm{tr}[\boldsymbol{A}]\,\mathrm{tr}[\boldsymbol{V}] - \mathsf{df} = \mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

On the other hand, by the same argument in the proof of Lemma 16 with the diminishing ridge penalty $\mu_n$ replaced by the strongly convexity parameter $\mu > 0$, we have $\mathrm{tr}[\boldsymbol{V}]/p \xrightarrow{\mathrm{p}} \nu > 0$ and $\mathrm{tr}[\boldsymbol{A}] \xrightarrow{\mathrm{p}} \kappa$. Therefore, we get

$$\mathrm{tr}[\boldsymbol{A}] - \mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(n^{-1/2}) \quad \text{and} \quad \mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(1).$$

Note in passing that the same things hold for $\widetilde{\mathsf{df}}$ and $\mathrm{tr}[\widetilde{\boldsymbol{V}}]$. By the Cauchy–Schwarz inequality, the error term due to the replacement of $(\mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}], \widetilde{\mathsf{df}}/\,\mathrm{tr}[\widetilde{\boldsymbol{V}}])$ by $(\mathrm{tr}[\boldsymbol{A}], \mathrm{tr}[\widetilde{\boldsymbol{A}}])$ is

$$\Big|\sum_{i\in[n]}\Big(r_i + \frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]}\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)^\top\Big(\widetilde{r}_i + \frac{\widetilde{\mathsf{df}}}{\mathrm{tr}[\widetilde{\boldsymbol{V}}]}\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\Big) - \sum_{i\in[n]}\Big(r_i + \mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)^\top\Big(\widetilde{r}_i + \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\Big)\Big|$$

$$\le \sqrt{\sum_{i\in[n]}\Big(r_i + \frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]}\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)^2}\cdot\Big|\frac{\widetilde{\mathsf{df}}}{\mathrm{tr}[\widetilde{\boldsymbol{V}}]} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\Big|\|\widetilde{\boldsymbol{\psi}}\| + \sqrt{\sum_{i\in[n]}(\widetilde{r}_i + \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i)^2}\cdot\Big|\frac{\mathsf{df}}{\mathrm{tr}[\boldsymbol{V}]} - \mathrm{tr}[\boldsymbol{A}]\Big|\|\boldsymbol{\psi}\|$$

$$= (\|\boldsymbol{z}\| + \mathcal{O}_{\mathbb{P}}(\sqrt{n}))\cdot\mathcal{O}_{\mathbb{P}}(1)$$

where the last equality follows from the following fact: $\|\boldsymbol{r}\|^2 \le 2(\|\boldsymbol{z}\|^2 + \|\boldsymbol{G}\boldsymbol{h}\|^2)$, $\|\boldsymbol{G}\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(n)$, $\|\boldsymbol{\psi}\|^2 = \mathcal{O}_{\mathbb{P}}(n)$, and $\mathsf{df}/\,\mathrm{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(1)$. Therefore, it suffices to show

$$n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} + \|\boldsymbol{z}\|^2 - \sum_{i\in[n]}(z_i - \boldsymbol{g}_i^\top\boldsymbol{h} + \mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i)(z_i - \boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} + \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i) = \mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\| + \mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

By simple algebra, the LHS can be decomposed into three terms $\xi_1 + \widetilde{\xi}_1 + \xi_2$ with:

$$\xi_1 = \sum_{i\in[n]} z_i(\boldsymbol{g}_i^\top\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i), \qquad \widetilde{\xi}_1 = \sum_{i\in[n]} z_i(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i)$$

$$\xi_2 = n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in[n]}\big(\boldsymbol{g}_i^\top\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i\big)\big(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \mathrm{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in\widetilde{I}\}}\,\widetilde{\psi}_i\big)$$

For $(\xi_1, \widetilde{\xi}_1)$, the derivative formula (44) and the argument in the proof of [17, Proposition 18] yield

$$\overline{\mathbb{E}}\Big[\frac{|\xi_1|}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2}\cdot\|\boldsymbol{z}\|}\Big] + \overline{\mathbb{E}}\Big[\frac{|\widetilde{\xi}_1|}{\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2}\cdot\|\boldsymbol{z}\|}\Big] \le C(\mu,\delta).$$

Since the denominators are $\mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\|$, we have $\xi_1 + \widetilde{\xi}_1 = \mathcal{O}_{\mathbb{P}}(1)\|\boldsymbol{z}\|$.

Below we bound $\xi_2$ using the following moment inequality, which is a variant of [11, Theorem 7.2].

LEMMA 21. *Let $\boldsymbol{\rho}, \widetilde{\boldsymbol{\rho}} : \mathbb{R}^{K\times Q} \to \mathbb{R}^Q$ be locally Lipschitz functions and $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} : \mathbb{R}^{K\times Q} \to \mathbb{R}^L$ be locally Lipschitz functions. If $(\boldsymbol{z}_k)_{k\in[K]}$ are i.i.d. $\mathcal{N}(\mathbf{0}_Q, \boldsymbol{I}_Q)$, we have*

$$\mathbb{E}\Big|\frac{K\boldsymbol{\rho}^\top\widetilde{\boldsymbol{\rho}} - \sum_{k\in[K]}(\boldsymbol{z}_k^\top\boldsymbol{\rho} - \sum_{q\in Q}\frac{\partial\rho_q}{z_{kq}})(\boldsymbol{z}_k^\top\widetilde{\boldsymbol{\rho}} - \sum_{q\in Q}\frac{\partial\widetilde{\rho}_q}{z_{kq}})}{\{\|\boldsymbol{\rho}\|^2 + \|\boldsymbol{\zeta}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{\rho}}\|^2 + \|\widetilde{\boldsymbol{\zeta}}\|^2\}^{1/2}}\Big| \le C_7(\sqrt{K}(1+\sqrt{\mathbb{E}[\Xi+\widetilde{\Xi}]}) + \mathbb{E}[\Xi+\widetilde{\Xi}])$$

*where* $\Xi := \frac{1}{\|\boldsymbol{\rho}\|^2+\|\boldsymbol{\zeta}\|^2}\sum_{k\in[K]}\sum_{q\in[Q]}\Big(\big\|\frac{\partial\boldsymbol{\rho}}{\partial g_{kq}}\big\|^2 + \big\|\frac{\partial\boldsymbol{\zeta}}{\partial g_{kq}}\big\|^2\Big)$.

(The proof of Lemma 21 is given in Appendix B.4.1.) By Lemma 21 with $(\boldsymbol{\rho}, \widetilde{\boldsymbol{\rho}}) = (\boldsymbol{h}, \widetilde{\boldsymbol{h}})$, $(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}) = (\boldsymbol{\psi}/\sqrt{p}, \widetilde{\boldsymbol{\psi}}/\sqrt{p})$, and $K = [n]$, we get

$$\overline{\mathbb{E}}\Big|\frac{n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in n}(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}})(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial\widetilde{h}_j}{\partial g_{ij}})}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2}}\Big| \le C_8(\sqrt{n}(1+\sqrt{\overline{\mathbb{E}}[\Xi+\widetilde{\Xi}]}) + \overline{\mathbb{E}}[\Xi+\widetilde{\Xi}]).$$

where

$$\Xi = \frac{1}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2}\sum_{i\in[n]}\sum_{j\in[p]}\Big(\big\|\frac{\partial\boldsymbol{h}}{\partial g_{ij}}\big\|^2 + \frac{1}{p}\big\|\frac{\partial\boldsymbol{\psi}}{\partial g_{ij}}\big\|^2\Big) \tag{42}$$

Let us bound $\Xi$. By the derivative formula (34), we have

$$\sum_{ij}\|\partial_{ij}\boldsymbol{h}\|^2 \le 2\|\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2\|\boldsymbol{h}\|^2, \quad \sum_{ij}\|\partial_{ij}\boldsymbol{\psi}\|^2 \le 2\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_{\mathrm{F}}^2\|\boldsymbol{\psi}\|^2 + 2\|\boldsymbol{V}\|_{\mathrm{F}}^2\|\boldsymbol{h}\|^2$$

where $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}$ and $\boldsymbol{D} = \mathrm{diag}\{\mathsf{loss}''(z_i - \boldsymbol{g}_i^\top\boldsymbol{h})\}$. By $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\|\boldsymbol{D}\|_{\mathrm{op}} \le \|\mathsf{loss}'\|_{\mathrm{Lip}}$, $\Xi$ is bounded from above by

$$\Xi \le 2p\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}\|_F^2 + 2\|\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\|_F^2 + 2p^{-1}\|\boldsymbol{V}\|_{\mathrm{F}}^2 \le C(\delta, \|\mathsf{loss}'\|_{\mathrm{Lip}})\cdot\|\boldsymbol{A}\|_{\mathrm{op}}^2\|\boldsymbol{G}\|_{\mathrm{op}}^4 \tag{43}$$

By $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ in (35) and Lemma 14, we obtain $\overline{\mathbb{E}}[\Xi] \le C_9$. This gives

$$\overline{\mathbb{E}}\Big|\frac{n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in n}(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}})(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial\widetilde{h}_j}{\partial g_{ij}})}{\{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2\}^{1/2}\{\|\widetilde{\boldsymbol{h}}\|^2 + p^{-1}\|\widetilde{\boldsymbol{\psi}}\|^2\}^{1/2}}\Big| \le C_{10}\sqrt{n}$$

Since the denominator is $\mathcal{O}_{\mathbb{P}}(1)$, we have

$$n\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top\boldsymbol{h} - \sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}}\Big)\Big(\boldsymbol{g}_i^\top\widetilde{\boldsymbol{h}} - \sum_{j\in[p]}\frac{\partial\widetilde{h}_j}{\partial g_{ij}}\Big) = \mathcal{O}_{\mathbb{P}}(\sqrt{n}).$$

By the derivative formula (34), noting that $\frac{\partial h_j}{\partial g_{ij}} = 0$ for all $i \notin I$ and $j \in [p]$, it holds that

$$\sum_{j\in[p]}\frac{\partial h_j}{\partial g_{ij}} = \begin{cases} 0 & i \notin I \\ \mathrm{tr}[\boldsymbol{A}]\psi_i - \boldsymbol{h}^\top\boldsymbol{A}\boldsymbol{G}\boldsymbol{D}\boldsymbol{e}_i & i \in I \end{cases} \tag{44}$$

Combined with $\|\boldsymbol{h}^\top \boldsymbol{AGD}\|^2 = \mathcal{O}_{\mathbb{P}}(n^{-1})$, the Cauchy–Schwarz inequality leads to

$$\sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top \boldsymbol{h} - \sum_{j\in[p]} \frac{\partial h_j}{\partial g_{ij}}\Big)\Big(\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \sum_{j\in[p]} \frac{\partial \widetilde{h}_j}{\partial g_{ij}}\Big) = \sum_{i\in[n]}\Big(\boldsymbol{g}_i^\top \boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\,\mathbb{1}_{\{i\in I\}}\,\psi_i\Big)\Big(\boldsymbol{g}_i^\top \widetilde{\boldsymbol{h}} - \operatorname{tr}[\widetilde{\boldsymbol{A}}]\,\mathbb{1}_{\{i\in \widetilde{I}\}}\,\widetilde{\psi}_i\Big) + \mathcal{O}_{\mathbb{P}}(1).$$

This gives $\xi_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{n}) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(\sqrt{n})$ and completes the proof.

B.4.1. *Proof of Lemma 21.* Let us denote $\boldsymbol{f} = (\boldsymbol{\rho}^\top, \boldsymbol{\zeta}^\top)^\top \in \mathbb{R}^{Q+L}$ and $\widetilde{\boldsymbol{f}} = (\widetilde{\boldsymbol{\rho}}^\top, \widetilde{\boldsymbol{\zeta}}^\top)^\top \mathbb{R}^{Q\times L}$. Let us write the product as

$$\frac{\boldsymbol{\rho}^\top \widetilde{\boldsymbol{\rho}}}{\|\boldsymbol{f}\|\|\widetilde{\boldsymbol{f}}\|} = \Big\|\frac{1}{2}\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \Big\|\frac{1}{2}\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2$$

Note that $\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|}$ and $\frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}$ are bounded by $1$ in the standard Euclid norm $\|\cdot\|$. Applying the $\chi$-square type moment inequality [11, Theorem 7.2] to these terms respectively, we observe that the following two terms are bounded from above by $\sqrt{K}\{1+\mathbb{E}[\Xi+\widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi+\widetilde{\Xi}]$ up to some universal constant:

$$\mathbb{E}\Big|K\Big\|\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \sum_{k\in[K]}\Big(\boldsymbol{z}_k^\top\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} + \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big) - \sum_{q\in[Q]}\frac{\partial}{\partial z_{kq}}\Big(\frac{\rho_q}{\|\boldsymbol{f}\|} + \frac{\widetilde{\rho}_q}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big)^2\Big|$$

$$\mathbb{E}\Big|K\Big\|\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big\|^2 - \sum_{k\in[K]}\Big(\boldsymbol{z}_k^\top\Big(\frac{\boldsymbol{\rho}}{\|\boldsymbol{f}\|} - \frac{\widetilde{\boldsymbol{\rho}}}{\|\widetilde{\boldsymbol{f}}\|}\Big) - \sum_{q\in[Q]}\frac{\partial}{\partial z_{kq}}\Big(\frac{\rho_q}{\|\boldsymbol{f}\|} - \frac{\widetilde{\rho}_q}{\|\widetilde{\boldsymbol{f}}\|}\Big)\Big)^2\Big|$$

Thus, by the triangle inequality, we are left with the bound of cross terms:

$$\mathbb{E}\Big|K\frac{\boldsymbol{\rho}^\top\widetilde{\boldsymbol{\rho}}}{\|\boldsymbol{f}\|\|\widetilde{\boldsymbol{f}}\|} - \sum_{k\in[K]} b_k\widetilde{b}_k\Big| \lesssim \sqrt{K}\{1+\mathbb{E}[\Xi+\widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi+\widetilde{\Xi}] \tag{45}$$

where $b_k = \boldsymbol{z}_k^\top\boldsymbol{\rho}/\|\boldsymbol{f}\| - \sum_{q\in[Q]}(\partial/\partial z_{kq})(\rho_q/\|\boldsymbol{f}\|)$. Expanding the derivative of the second term, $b_k$ can be written as

$$b_k = \underbrace{\frac{\boldsymbol{z}_k^\top\boldsymbol{\rho} - \sum_{q\in[Q]}\frac{\partial\rho_q}{\partial z_{kq}}}{\|\boldsymbol{f}\|}}_{=:a_k} - \sum_{q\in[Q]}\rho_q\frac{\partial}{\partial z_{kq}}\frac{1}{\|\boldsymbol{f}\|} = a_k + \sum_{q\in[Q]}\rho_q\frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3}\frac{\partial\boldsymbol{f}}{\partial z_{kq}}$$

Thus, by multiple applications of Cauchy–Schwarz inequality, using $\|\boldsymbol{\rho}\|^2\le\|\boldsymbol{f}\|^2$, we find that the error $\|\boldsymbol{b}-\boldsymbol{a}\|^2 = \sum_{k\in[K]}(b_k-a_k)^2$ is bounded from above by $2\Xi$:

$$\|\boldsymbol{a}-\boldsymbol{b}\|^2 = \sum_{k\in K}\Big(\sum_{q\in[Q]}\rho_q\frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3}\frac{\partial\boldsymbol{f}}{\partial z_{kq}}\Big)^2 \le \sum_k\|\boldsymbol{\rho}\|^2\sum_q\Big(\frac{\boldsymbol{f}^\top}{\|\boldsymbol{f}\|^3}\frac{\partial\boldsymbol{f}}{\partial z_{kq}}\Big)^2 \le \frac{1}{\|\boldsymbol{f}\|^2}\sum_{k,q}\Big\|\frac{\partial\boldsymbol{f}}{\partial z_{kq}}\Big\|^2 \le 2\Xi.$$

The same argument gives $\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2\le 2\widetilde{\Xi}$.

Now we claim the following deterministic inequality for all $\boldsymbol{u},\widetilde{\boldsymbol{u}},\boldsymbol{a},\widetilde{\boldsymbol{a}},\boldsymbol{b},\widetilde{\boldsymbol{b}}\in\mathbb{R}^K$ with $\|\boldsymbol{u}\|\vee\|\widetilde{\boldsymbol{u}}\|\le 1$:

$$\begin{aligned}\big||K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top\widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top\widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top\widetilde{\boldsymbol{b}}|\big| &\le (\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2) + \sqrt{K}(\|\boldsymbol{a}-\boldsymbol{b}\| + \|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|)\\ &\quad + 2^{-1}(|K\|\boldsymbol{u}\|^2 - \|\boldsymbol{b}\|^2| + |K\|\widetilde{\boldsymbol{u}}\|^2 - \|\widetilde{\boldsymbol{b}}\|^2|)\end{aligned} \tag{46}$$

We prove this inequality later. Applying this inequality with $\boldsymbol{u} = \boldsymbol{\rho}/\|\boldsymbol{f}\|$ and $(\boldsymbol{a}, \boldsymbol{b})$ defined above, using $\|\boldsymbol{a}-\boldsymbol{b}\|^2 \le 2\Xi$, we get

$$\begin{aligned}|K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top \widetilde{\boldsymbol{a}}| &\le 2(\Xi + \widetilde{\Xi}) + \sqrt{2K}(\Xi^{1/2} + \widetilde{\Xi}^{1/2}) \\ &\quad + 2^{-1}(|K\|\boldsymbol{u}\|^2 - \|\boldsymbol{b}\|^2| + |K\|\widetilde{\boldsymbol{u}}\|^2 - \|\widetilde{\boldsymbol{b}}\|^2|) \\ &\quad + |K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top \widetilde{\boldsymbol{b}}|\end{aligned}$$

Taking the expectation, using the moment bound (45), we are left with

$$\begin{aligned}\mathbb{E}|K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top \widetilde{\boldsymbol{a}}| &\lesssim \mathbb{E}[\Xi + \widetilde{\Xi}] + \sqrt{K}\mathbb{E}[(\Xi^{1/2} + \widetilde{\Xi}^{1/2})] \\ &\quad + \{1 + \mathbb{E}[\Xi]\}^{1/2} + \mathbb{E}[\Xi] + \{1 + \mathbb{E}[\widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\widetilde{\Xi}] \\ &\quad + \{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}]\end{aligned}$$

Using Jensen's inequality $\mathbb{E}[X^{1/2}] \le \sqrt{\mathbb{E}[X]}$ for any non-negative random variable $X$ and $\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$ for any non-negative scalars $a, b$, the RHS is bounded from above by $\sqrt{K}\{1 + \mathbb{E}[\Xi + \widetilde{\Xi}]\}^{1/2} + \mathbb{E}[\Xi + \widetilde{\Xi}]$ up to some universal constant. This finishes the proof.

Below we prove the deterministic inequality (46). By multiple applications of the triangle inequality and Cauchy–Schwarz inequality,

$$\begin{aligned}\big||K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top \widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top \widetilde{\boldsymbol{b}}|\big| &\le |\boldsymbol{a}^\top \widetilde{\boldsymbol{a}} - \boldsymbol{b}^\top \widetilde{\boldsymbol{b}}| \\ &\le |(\boldsymbol{a}-\boldsymbol{b})^\top(\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}) + (\boldsymbol{a}-\boldsymbol{b})^\top \widetilde{\boldsymbol{b}} + \boldsymbol{b}^\top(\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}})| \\ &\le \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\| + \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| + \|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\|\|\boldsymbol{b}\| \\ &\le \frac{\|\boldsymbol{a}-\boldsymbol{b}\|^2 + \|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\|^2}{2} + \|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| + \|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\|\|\boldsymbol{b}\|\end{aligned}$$

Using $\|\boldsymbol{b}\| \le \sqrt{|\|\boldsymbol{b}\|^2 - K\|\boldsymbol{u}\|^2|} + \sqrt{K\|\boldsymbol{u}\|^2}$ and $\|\boldsymbol{u}\| \le 1$, $\|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\|$ can be bounded from above as

$$\|\boldsymbol{a}-\boldsymbol{b}\|\|\widetilde{\boldsymbol{b}}\| \le \|\boldsymbol{a}-\boldsymbol{b}\|\sqrt{|\|\widetilde{\boldsymbol{b}}\|^2 - K\|\widetilde{\boldsymbol{u}}\|^2|} + \sqrt{K}\|\boldsymbol{a}-\boldsymbol{b}\| \le \frac{\|\boldsymbol{a}-\boldsymbol{b}\|^2}{2} + \frac{|\|\widetilde{\boldsymbol{b}}\|^2 - K\|\widetilde{\boldsymbol{u}}\|^2|}{2} + \sqrt{K}\|\boldsymbol{a}-\boldsymbol{b}\|.$$

The same argument gives $\|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\|\boldsymbol{b}\| \le \frac{\|\widetilde{\boldsymbol{a}}-\widetilde{\boldsymbol{b}}\|^2}{2} + \frac{\|\boldsymbol{b}\|^2 - K\|\boldsymbol{u}\|^2|}{2} + \sqrt{K}\|\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{b}}\|$. Putting them all together, we obtained the desired upper bound of $\big||K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{a}^\top \widetilde{\boldsymbol{a}}| - |K\boldsymbol{u}^\top \widetilde{\boldsymbol{u}} - \boldsymbol{b}^\top \widetilde{\boldsymbol{b}}|\big|$.

**B.5. Proof of trace convergence.** We assume without loss of generality that $I = [n]$. Throughout this section, we denote $\boldsymbol{h} = \widehat{\boldsymbol{h}}_{\mu,K}$ and $\boldsymbol{\psi} = \boldsymbol{\psi}_{\mu,K}$ for simplicity.

LEMMA 22. *For any $\mu \in (0,1]$ with $\mu^{-1} = o(n^{1/4})$,*

$$\mathrm{tr}[\boldsymbol{V}] = \|\boldsymbol{h}\|^{-2}\big(\|\boldsymbol{\psi}\|^2 \mathrm{tr}[\boldsymbol{A}] - \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h}\big) + \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})$$

$$\mathrm{tr}[\boldsymbol{A}]^2 = \frac{(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top \boldsymbol{\psi}\|^2)}{\|\boldsymbol{\psi}\|^4} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-2})$$

*and $\mathbb{P}(\mathrm{tr}[\boldsymbol{A}]^2 \le C) \to 1$ for a constant $C > 0$ that only depend on $(\alpha, \beta, \delta)$.*

PROOF. First we show the stochastic representation of $\mathrm{tr}[\boldsymbol{V}]$ by $(\boldsymbol{h}, \boldsymbol{\psi}, \mathrm{tr}[\boldsymbol{A}])$.

LEMMA 23 ([17]). *Let* $\boldsymbol{\rho} : \mathbb{R}^{K\times Q} \to \mathbb{R}^Q$ *and* $\boldsymbol{\zeta} : \mathbb{R}^{K\times Q} \to \mathbb{R}^K$ *be two locally Lipschitz functions with differentiable components. If* $\boldsymbol{Z} \in \mathbb{R}^{K\times Q}$ *has i.i.d.* $\mathcal{N}(0,1)$ *entries, we have*

$$\mathbb{E}\Big[\Big(\frac{\boldsymbol{\zeta}^\top \boldsymbol{Z}\boldsymbol{\rho} - \sum_{kq} \frac{\partial(\zeta_k \rho_q)}{\partial g_{kq}}}{\|\boldsymbol{\zeta}\|^2 + \|\boldsymbol{\rho}\|^2}\Big)^2\Big] \le C_{11}(1 + \mathbb{E}[\Xi])$$

*where* $\Xi := \frac{1}{\|\boldsymbol{\rho}\|^2+\|\boldsymbol{\zeta}\|^2} \sum_{k\in[K]} \sum_{q\in[Q]} \Big(\big\|\frac{\partial \boldsymbol{\rho}}{\partial g_{kq}}\big\|^2 + \big\|\frac{\partial \boldsymbol{\zeta}}{\partial g_{kq}}\big\|^2\Big)$.

By Lemma 23 with $(\boldsymbol{\zeta}, \boldsymbol{\rho}) = (p^{-1/2}\boldsymbol{\psi}, \boldsymbol{h})$ and $\boldsymbol{Z} = \boldsymbol{G}$, we have

$$\mathbb{E}\Big[\Big(\frac{\frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \frac{1}{\sqrt{p}}\sum_{ij} \frac{\partial(\psi_i h_j)}{\partial g_{ij}}}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2}\Big)^2\Big] \le C_{12}(1 + \mathbb{E}[\Xi])$$

where $\Xi$ is the same estimate as in (42). Using the upper estimate $\Xi \le C_{13}\|\boldsymbol{A}\|_{\mathrm{op}}^2 \|\boldsymbol{G}\|_{\mathrm{op}}^4$ in (43), $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\mathbb{E}[\|\boldsymbol{G}\|^4] \le C(\delta)n^2$, we find that the RHS in the above display is bounded from away by $C(\delta)(1+\mu^{-2})$. Since the denominator $\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2$ is $\mathcal{O}_{\mathbb{P}}(1)$, we get

$$\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \sum_{ij} \frac{\partial \psi_i h_j}{\partial g_{ij}} = \sqrt{p} \cdot \mathcal{O}_{\mathbb{P}}(1) \cdot \mathcal{O}_{\mathbb{P}}(\mu^{-1}) = \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1}).$$

For the sum of derivative $\sum_{ij} \frac{\partial(\psi_i h_j)}{\partial g_{ij}}$, using $\|\boldsymbol{A}\|_{\mathrm{op}} \le (p\mu)^{-1}$ and $\|\boldsymbol{G}\|_{\mathrm{op}} = \mathcal{O}_{\mathbb{P}}(\sqrt{n})$, we have

$$\begin{aligned}\sum_{ij} \frac{\partial(\psi_i h_j)}{\partial g_{ij}} &= \|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] - \boldsymbol{h}^\top \boldsymbol{G}^\top \boldsymbol{D}\boldsymbol{\psi} - \boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{h} - \|\boldsymbol{h}\|^2 \operatorname{tr}[\boldsymbol{V}] \\ &= \|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] + \mathcal{O}_{\mathbb{P}}(n^{1/2}) + \mathcal{O}_{\mathbb{P}}(\mu^{-1}) - \|\boldsymbol{h}\|^2 \operatorname{tr}[\boldsymbol{V}]\end{aligned}$$

so that we are left with

$$\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} - \|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] + \|\boldsymbol{h}\|^2 \operatorname{tr}[\boldsymbol{V}] = \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1}).$$

Dividing by $\|\boldsymbol{h}\|^2$, noting $\|\boldsymbol{h}\|^{-1} = \mathcal{O}_{\mathbb{P}}(1)$ from $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, we obtain the representation of $\operatorname{tr}[\boldsymbol{V}]$.

Next, we show the stochastic representation of $\operatorname{tr}[\boldsymbol{A}^2]$ by $(\boldsymbol{G}, \boldsymbol{h}, \boldsymbol{\psi})$. By the stochastic representation of $\operatorname{tr}[\boldsymbol{V}]$,

$$\begin{aligned}\big|\|\boldsymbol{\psi}\|^4 \operatorname{tr}[\boldsymbol{A}]^2 - (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \operatorname{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2)^2\big| &\le \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1})\big|\|\boldsymbol{\psi}\|^2 \operatorname{tr}[\boldsymbol{A}] + \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} + \operatorname{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2\big| \\ &\le \mathcal{O}_{\mathbb{P}}(\sqrt{n}\mu^{-1})\mathcal{O}_{\mathbb{P}}(n\mu^{-1} + n + n) \\ &= \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}).\end{aligned}$$

By Lemma 21 with $\boldsymbol{\rho} = \widetilde{\boldsymbol{\rho}} = \boldsymbol{\psi}/\sqrt{p}$ and $\boldsymbol{\zeta} = \widetilde{\boldsymbol{\zeta}} = \boldsymbol{h}$, we have

$$\mathbb{E}\Big[\frac{\big|p\|\frac{\boldsymbol{\psi}}{\sqrt{p}}\|^2 - \sum_{j\in[p]}(\frac{\boldsymbol{\psi}^\top}{\sqrt{p}}\boldsymbol{G}\boldsymbol{e}_j - \frac{1}{\sqrt{p}}\sum_{i\in[n]}\frac{\partial \boldsymbol{\psi}}{\partial g_{ij}})^2\big|}{\|\boldsymbol{h}\|^2 + p^{-1}\|\boldsymbol{\psi}\|^2}\Big] \le C_{14}(\sqrt{n}(1 + \mathbb{E}[\Xi]^{1/2}) + \mathbb{E}[\Xi]).$$

where $\Xi$ is the same estimate as in (42). Using $\mathbb{E}[\Xi] \le C_{15}\mu^{-2}$ again, we get

$$\|\boldsymbol{\psi}\|^2 - \frac{1}{p}\sum_{j\in[p]}(\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j - \sum_{i\in[n]}\frac{\partial \psi_i}{\partial g_{ij}})^2 = \mathcal{O}_{\mathbb{P}}(1) \cdot \mathcal{O}_{\mathbb{P}}(\sqrt{n}(1 + \mu^{-1}) + \mu^{-2}) = (n^{1/2}\mu^{-1})$$

Using $\sum_{i\in[n]} \frac{\partial \psi_i}{\partial g_{ij}} = -\boldsymbol{\psi}^\top \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j - \operatorname{tr}[\boldsymbol{V}]h_j$ by the derivative formula (34), it holds that

$$\begin{aligned}
&\Big\| \sum_{j\in[p]} (\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{e}_j - \sum_{i\in[n]} \frac{\partial \psi_i}{\partial g_{ij}})^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 \Big\| \\
&= \Big| \|\boldsymbol{G}^\top\boldsymbol{\psi} + \boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 \Big| \\
&\le \|\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi}\| \Big( \|\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{\psi}\| + 2\|\boldsymbol{G}^\top\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\| \Big) \\
&= \mathcal{O}_{\mathbb{P}}(\mu^{-1})(\mathcal{O}_{\mathbb{P}}(\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(n)) = \mathcal{O}_{\mathbb{P}}(\mu^{-1}n).
\end{aligned}$$

Combining the above displays, we are left with

$$\|\boldsymbol{\psi}\|^2 - p^{-1}\|\boldsymbol{G}^\top\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 = \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(\mu^{-1}n^{-1}) = \mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1}).$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{\psi}\|^4 \operatorname{tr}[\boldsymbol{A}]^2 &= (\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} + \operatorname{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2)^2 + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}) \\
&= (\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(2\operatorname{tr}[\boldsymbol{V}]\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} + \operatorname{tr}[\boldsymbol{V}]^2\|\boldsymbol{h}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2})) \\
&= (\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(\|\boldsymbol{G}^\top\boldsymbol{\psi} + \operatorname{tr}[\boldsymbol{V}]\boldsymbol{h}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2})) \\
&= (\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1}) + \mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-2}).
\end{aligned}$$

Multiplying by $\|\boldsymbol{\psi}\|^{-4}$, which is $\mathcal{O}_{\mathbb{P}}(n^2)$ since $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\text{p}} \beta^2 > 0$, we obtain this representation of $\operatorname{tr}[\boldsymbol{A}]^2$. The upper bound $\mathbb{P}(\operatorname{tr}[\boldsymbol{A}]^2 \le C_{16})$ follows from the stochastic representation and the convergences: $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\text{p}} \beta^2 > 0$, $\|\boldsymbol{h}\|^2 \xrightarrow{\text{p}} \alpha^2 > 0$, $\|\boldsymbol{G}\|_{\text{op}}/\sqrt{n} \xrightarrow{\text{p}} 1 + \sqrt{\delta} > 0$. □

LEMMA 24. *Letting $\overline{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z},\boldsymbol{\theta}]$ be the conditional expectation with respect to the design matrix $\boldsymbol{G}$, for any $\mu \in (0,1]$ with $\mu^{-2} = o(n)$, we have*

$$\frac{(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2 + \|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)}{\|\boldsymbol{\psi}\|^4} = \frac{(\overline{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}])^2 + \overline{\mathbb{E}}[\|\boldsymbol{h}\|]^2(p\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2 - \|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])}{\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2} + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).$$

PROOF. By the Gaussian Poincaré inequality, we claim the following:

$$\begin{aligned}
\overline{\mathbb{E}}[(\|\boldsymbol{\psi}\|^2 - \overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])^2] &\le C_{17}n\mu^{-2}, \\
\overline{\mathbb{E}}[(\|\boldsymbol{h}\|^2 - \overline{\mathbb{E}}[\|\boldsymbol{h}\|^2])^2] &\le C_{18}n^{-1}\mu^{-2}, \\
\overline{\mathbb{E}}[(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \overline{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}])^2] &\le C_{19}n\mu^{-2}, \\
\overline{\mathbb{E}}[(\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2 - \overline{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])^2] &\le C_{20}n^3\mu^{-2}.
\end{aligned}$$

First and second moment inequalities immediately follow from Lemma 17 with $\widetilde{I} = I = [n]$. For the third moment inequality, the Gaussian Poincaré inequality gives the upper bound $\mathbb{E}[(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \overline{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}])^2] \le \sum_{ij}\mathbb{E}\Big(\frac{\partial \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}}{\partial g_{ij}}\Big)^2$, where

$$\begin{aligned}
\text{RHS}_{ij} := \frac{\partial \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}}{\partial g_{ij}} &= \Big(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i - \boldsymbol{V}\boldsymbol{e}_i h_j\Big)^\top \boldsymbol{G}\boldsymbol{h} + \psi_i h_j + \boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{A}\Big(\boldsymbol{e}_j\psi_i - \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{e}_i h_j\Big) \\
&= \boldsymbol{e}_j^\top\boldsymbol{A}^\top\Big(-\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{h} + \boldsymbol{G}^\top\boldsymbol{\psi}\Big)\psi_i - \boldsymbol{e}_i^\top\Big(\boldsymbol{V}^\top\boldsymbol{G}\boldsymbol{h} + \boldsymbol{D}\boldsymbol{G}\boldsymbol{A}^\top\boldsymbol{G}^\top\boldsymbol{\psi}\Big)h_j + \psi_i h_j
\end{aligned}$$

By $\boldsymbol{V}=\boldsymbol{D}-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top\boldsymbol{D}$ and $\|\boldsymbol{A}\|_{\rm op}\le(p\mu)^{-1}$ in (35), we have $\sum_{i,j}\mathbb{E}[\mathrm{RHS}_{ij}^2]\le C_{21}n\mu^{-2}$. Next,

$$\mathbb{E}\Big[\Big(\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2-\bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2]\Big)^2\Big]\le\sum_{i,j}\mathbb{E}\Big(\frac{\partial\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2}{\partial g_{ij}}\Big)^2=4\sum_{i,j}\mathbb{E}(\boldsymbol{\psi}^\top\boldsymbol{G}\frac{\partial\boldsymbol{G}^\top\boldsymbol{\psi}}{\partial g_{ij}})^2$$

where

$$\begin{aligned}\boldsymbol{\psi}^\top\boldsymbol{G}\frac{\partial\boldsymbol{G}^\top\boldsymbol{\psi}}{\partial g_{ij}}&=\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{e}_j\psi_i+\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{G}^\top(-\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\psi_i-\boldsymbol{V}\boldsymbol{e}_ih_j)\\&=\boldsymbol{\psi}^\top\boldsymbol{G}(\boldsymbol{I}_p-\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G}\boldsymbol{A})\boldsymbol{e}_j\psi_i-\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{V}\boldsymbol{e}_ih_j\end{aligned}$$

so $\sum_{ij}\mathbb{E}[\mathrm{RHS}_{ij}^2]\le C_{22}n^3\mu^{-2}$. Thus, we get

$$\begin{aligned}\|\boldsymbol{\psi}\|^2&=\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]+\mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\\\|\boldsymbol{h}\|^2&=\bar{\mathbb{E}}[\|\boldsymbol{h}\|^2]+\mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1})\\\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}&=\bar{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}]+\mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\\\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2&=\bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2]+\mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})\end{aligned}$$

Since $\mu^{-2}=o(n)$, the concentration of $\|\boldsymbol{\psi}\|^2$ on the conditional expectation $\mathbb{E}[\|\boldsymbol{\psi}\|^2]$ and the convergence $\|\boldsymbol{\psi}\|^2/p\xrightarrow{\mathrm{p}}\beta^2$ yield $\mathbb{E}[\|\boldsymbol{\psi}\|^2]/p\xrightarrow{\mathrm{p}}\beta^2>0$. This implies that $1/\|\boldsymbol{\psi}\|^2$ and $1/\mathbb{E}[\|\boldsymbol{\psi}\|^2]$ are $\mathcal{O}_{\mathbb{P}}(n^{-1})$. Then, the error from replacing the denominator $\|\boldsymbol{\psi}\|^4$ by $\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2$ is estimated as

$$\begin{aligned}&(\|\boldsymbol{\psi}\|^{-4}-\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2)\big((\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2+\|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2-\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)\big)\\&=\frac{(\|\boldsymbol{\psi}\|^2-\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])(\|\boldsymbol{\psi}\|^2+\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2])}{\|\boldsymbol{\psi}\|^4\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2}\cdot\mathcal{O}_{\mathbb{P}}(n^2)\\&=\mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\mathcal{O}_{\mathbb{P}}(n\cdot n^{-4})\mathcal{O}_{\mathbb{P}}(n^2)=\mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).\end{aligned}$$

For the error from replacing the numerator with the conditional one, the error of replacing each term with the conditional one is given by

$$\begin{aligned}(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2&=(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2+\mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})\\\|\boldsymbol{h}\|^2\|\boldsymbol{\psi}\|^2&=\bar{\mathbb{E}}[\|\boldsymbol{h}\|^2]\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]+\mathcal{O}_{\mathbb{P}}(n^{1/2}\mu^{-1})\\\|\boldsymbol{h}\|^2\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2&=\bar{\mathbb{E}}[\|\boldsymbol{h}\|^2]\bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2]+\mathcal{O}_{\mathbb{P}}(n^{3/2}\mu^{-1})\end{aligned}$$

Thus, noting that $\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^{-1}=\mathcal{O}_{\mathbb{P}}(n^{-1})$, we get

$$\frac{(\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h})^2+\|\boldsymbol{h}\|^2(p\|\boldsymbol{\psi}\|^2-\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2)-(\bar{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}])^2-\bar{\mathbb{E}}[\|\boldsymbol{h}\|^2](p\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]-\bar{\mathbb{E}}[\|\boldsymbol{G}^\top\boldsymbol{\psi}\|^2])}{\bar{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]^2}$$
$$=\mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}).$$

This finishes the proof. □

LEMMA 25. *Suppose $\mu\in(0,1]$ and $\mu^{-1}=O(n^{1/8})$. Then, there exist a non-negative random variables $\widehat{\kappa}\ge0$, which is independent of $\boldsymbol{G}$, such that*

$$\mathbb{P}\Big(|\operatorname{tr}[\boldsymbol{A}]-\widehat{\kappa}|\le2n^{-1/16}\quad\textit{and}\quad|\widehat{\kappa}|\le C\Big)\to1$$

*where $C$ is a constant depending on $(\alpha,\beta,\delta)$ only.*

PROOF. By Lemma 22 and Lemma 24, there exists a random variable $A_n$, which is independent of $\boldsymbol{G}$, such that

$$\operatorname{tr}[\boldsymbol{A}]^2 = A_n + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-2}) + \mathcal{O}_{\mathbb{P}}(n^{-1/2}\mu^{-1}) = A_n + \mathcal{O}_{\mathbb{P}}(n^{-1/4}), \quad \text{and} \quad \mathbb{P}(|A_n| \le C) \to 1$$

Noting $\operatorname{tr}[\boldsymbol{A}]^2 \ge 0$ and $\mathcal{O}_{\mathbb{P}}(n^{-1/4}) = o_{\mathbb{P}}(n^{-1/8})$, this implies that the event

$$\Omega := \{|\operatorname{tr}[\boldsymbol{A}]^2 - A_n| \le n^{-1/8}\} \cap \{-n^{-1/8} \le A_n \le C\}$$

holds with high probability. Let us take $\widehat{\kappa} := \sqrt{(A_n + 2n^{-1/8})_+}$. Note in passing that under the event $\Omega$, we have $\widehat{\kappa} = \sqrt{A_n + 2n^{-1/8}}$ and $\widehat{\kappa} \le \sqrt{C+1}$. By the non-negativeness $\operatorname{tr}[\boldsymbol{A}] \ge 0$ and the $(1/2)$-Hölder continuity of the square root $\mathbb{R}_{\ge 0} \ni x \mapsto \sqrt{x}$, under the event $\Omega$, it holds that

$$|\operatorname{tr}[\boldsymbol{A}] - \widehat{\kappa}| = |\sqrt{\operatorname{tr}[\boldsymbol{A}]^2} - \sqrt{A_n + 2n^{-1/8}}| \le |\operatorname{tr}[\boldsymbol{A}]^2 - A_n - 2n^{-1/8}|^{1/2}$$

and the RHS is less than $|n^{-1/8} + 2n^{-1/8}|^{1/2} \le 2n^{-1/16}$ by the triangle inequality. □

B.5.1. *Proof of Lemma 16.* Applying Lemma 18 with $M = 1$ and $(\boldsymbol{z}, \boldsymbol{F}(\boldsymbol{z})) = (\boldsymbol{g}_i, \boldsymbol{h})$ for each $i \in [n]$, using $\sum_{ij} \overline{\mathbb{E}}[\|\partial_{ij}\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(\mu^{-2}) = o_{\mathbb{P}}(n)$ we get

$$\sum_{i\in[n]} (\boldsymbol{g}_i^\top \boldsymbol{h} - \operatorname{tr}[\boldsymbol{A}]\psi_i - \|\boldsymbol{h}\|\widehat{u}_i)^2 = o_{\mathbb{P}}(n), \quad \widehat{u}_i|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{G}_{-i} \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$$

By Lemma 25, there exists a non-negative random variable $\widehat{\kappa}$ independent of $\boldsymbol{G}$ such that

$$\operatorname{tr}[\boldsymbol{A}] = \widehat{\kappa} + o_{\mathbb{P}}(1), \quad \mathbb{P}(\widehat{\kappa} \le C) \to 1$$

for a positive constant $C$. Then, combined with $\|\boldsymbol{h}\| = \alpha + o_{\mathbb{P}}(1)$, we get

$$\frac{1}{2n}\sum_{i\in[n]} (\boldsymbol{g}_i^\top \boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i)^2 = o_{\mathbb{P}}(1) + (\widehat{\kappa} - \operatorname{tr}[\boldsymbol{A}])^2 \frac{\|\boldsymbol{\psi}\|^2}{n} + (\|\boldsymbol{h}\| - \alpha)^2 \frac{\sum_{i\in[n]} \widehat{\boldsymbol{u}}_i^2}{n} = o_{\mathbb{P}}(1).$$

Furthermore, using the independence of $(\widehat{\kappa}, \boldsymbol{G})$ and the upper bound $\widehat{\kappa} \le C$, by the same argument in the proof of Theorem 2, we can easily show that the conditional expectation $\overline{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\boldsymbol{z}, \boldsymbol{\theta}]$ of the square of LHS is bounded by $C$ with high probability for some constant $C$. This implies that the conditional expectation $\overline{\mathbb{E}}$ of LHS is also $o(1)$, i.e.,

$$\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}[(\boldsymbol{g}_i^\top \boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i)^2] = o_{\mathbb{P}}(1).$$

Let us define $\Xi_i := \boldsymbol{g}_i^\top \boldsymbol{h} - \widehat{\kappa}\psi_i - \alpha\widehat{u}_i$ so that the above display reads $n^{-1}\sum_i \overline{\mathbb{E}}[\Xi_i^2] = o_{\mathbb{P}}(1)$. Noting $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ for all $i \in [n]$, the residual can be written as

$$z_i - \boldsymbol{g}_i^\top \boldsymbol{h} = \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i - \Xi_i; \widehat{\kappa})$$

for all $i \in [n]$. Since $\mathsf{prox}_f(\cdot)$ is 1-Lipschitz for any convex function, we have

$$\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\Big(z_i - \boldsymbol{g}_i^\top \boldsymbol{h} - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})\Big)^2\Big] \le \frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}[\Xi_i^2] = o_{\mathbb{P}}(1).$$

Since $\mathsf{loss}'$ is Lipschitz, the above display lets us approximate $\psi_i = \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \boldsymbol{h})$ by $\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})$:

$$\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\Big(\psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})\Big)^2\Big] = o_{\mathbb{P}}(1).$$

Applying this approximation to the concentration $\beta^2 = \|\boldsymbol{\psi}\|^2/p + o_{\mathbb{P}}(1) = \overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2]/p + o_{\mathbb{P}}(1)$ by Lemma 17 with $I = \widetilde{I} = [n]$, we get

$$\beta^2 = \frac{n}{p}\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\psi_i^2\Big] + o_{\mathbb{P}}(1) = \delta\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i; \widehat{\kappa})^2\Big] + o_{\mathbb{P}}(1)$$

$$= \delta\frac{1}{n}\sum_{i\in[n]} \int_{-\infty}^{\infty} \varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x; \widehat{\kappa})^2 \,\mathrm{d}x + o_{\mathbb{P}}(1). \tag{47}$$

where $\varphi(x)$ is the pdf of $\mathcal{N}(0,1)$. Let us define the functions $F, \widehat{F} : [0,\infty) \to \mathbb{R}$ by

$$F(\tau) := \beta^2 - \delta\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(Z + \alpha G; \tau)^2], \quad \widehat{F}(\tau) := \beta^2 - \delta\frac{1}{n}\sum_{i\in[n]}\int_{-\infty}^{\infty} \varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x; \tau)^2 \,\mathrm{d}x$$

so that $F(\kappa) = 0$ by (4b) in System 1a (with $c = 1$), while (47) reads $\widehat{F}(\widehat{\kappa}) = o_{\mathbb{P}}(1)$. Note in passing that the weak law of large numbers implies $\widehat{F}(\tau) \xrightarrow{\mathrm{p}} F(\tau)$ pointwise, and $F$ and $\widehat{F}$ are strictly increasing functions in $\tau$ since $-2^{-1}\mathsf{env}'_{\mathsf{loss}}(x;\tau)^2$ is the derivative of the convex function $\tau \mapsto \mathsf{env}_{\mathsf{loss}}(x;\tau)$, which is strictly convex under Assumption D-(4) (see [95, Lemma 4.4]). Then, for any $\epsilon > 0$, we have

$$F(\kappa + \epsilon) > 0 = F(\kappa) > F(\kappa - \epsilon).$$

By the pointwise convergence $\widehat{F}(\tau) \xrightarrow{\mathrm{p}} F(\tau)$, it holds that

$$\widehat{F}(\kappa + \epsilon) > 2^{-1}F(\kappa + \epsilon) > 0 > 2^{-1}F(\kappa - \epsilon) > \widehat{F}(\kappa - \epsilon).$$

with high probability. Then, combined with $\widehat{F}(\widehat{\kappa}) = o_{\mathbb{P}}(1)$, we have

$$\widehat{F}(\kappa + \epsilon) > 2^{-1}F(\kappa + \epsilon) > \widehat{F}(\widehat{\kappa}) > 2^{-1}F(\kappa - \epsilon) > \widehat{F}(\kappa - \epsilon).$$

with high probability. Since $\widehat{F}$ is non-decreasing with probability 1, this gives $\mathbb{P}(|\widehat{\kappa} - \kappa| \le \epsilon) \to 1$. Since we took $\epsilon > 0$ arbitrarily, we obtain $\widehat{\kappa} \xrightarrow{\mathrm{p}} \kappa$.

Going back to the place where we replaced $\mathrm{tr}[\boldsymbol{A}]$ by $\widehat{\kappa}$, now replacing $\mathrm{tr}[\boldsymbol{A}]$ by $\kappa$ instead, we obtain

$$\frac{1}{n}\sum_{i\in[n]} (\boldsymbol{g}_i^\top \boldsymbol{h} - \kappa\psi_i - \alpha\widehat{u}_i)^2 = o_{\mathbb{P}}(1),$$

and this approximation also holds in $\overline{\mathbb{E}}$. Thus, we get the approximation of residual and $\psi_i$

$$\frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\Big(z_i - \boldsymbol{g}_i^\top\boldsymbol{h} - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\Big)^2\Big] = o_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\Big(\psi_i - \mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)\Big)^2\Big] = o_{\mathbb{P}}(1)$$

Let us show $\mathrm{tr}[\boldsymbol{V}]/p \xrightarrow{\mathrm{p}} \nu$. Using the concentration $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} - \mathrm{tr}[\boldsymbol{A}]\|\boldsymbol{\psi}\|^2 + \mathrm{tr}[\boldsymbol{V}]\|\boldsymbol{h}\|^2 = o_{\mathbb{P}}(n)$ from Lemma 22 and $\|\boldsymbol{h}\|^2 \xrightarrow{\mathrm{p}} \alpha^2 > 0$, $\|\boldsymbol{\psi}\|^2/p \xrightarrow{\mathrm{p}} \beta^2$, $\mathrm{tr}[\boldsymbol{A}] \xrightarrow{\mathrm{p}} \kappa$, we have

$$p^{-1}\,\mathrm{tr}[\boldsymbol{V}] = \alpha^{-2}\kappa\beta^2 - \alpha^{-2}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}/p + o_{\mathbb{P}}(1).$$

Recall that we have shown the concentration $\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} = \overline{\mathbb{E}}[\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}] + o_{\mathbb{P}}(n)$ in the proof of Lemma 24. Applying the proximal approximation of the residual $z_i - \boldsymbol{g}_i^\top\boldsymbol{h}$ and $\psi_i$ to this, noting $\overline{\mathbb{E}}[\|\boldsymbol{\psi}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$ and $\overline{\mathbb{E}}[\|\boldsymbol{G}\boldsymbol{h}\|^2] = \mathcal{O}_{\mathbb{P}}(n)$, we are left with

$$\frac{1}{n}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h} = \frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}[\psi_i \cdot \boldsymbol{g}_i^\top\boldsymbol{h}] + o_{\mathbb{P}}(1)$$

$$= \frac{1}{n}\sum_{i\in[n]} \overline{\mathbb{E}}\Big[\mathsf{env}'_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa)(z_i - \mathsf{prox}_{\mathsf{loss}}(z_i - \alpha\widehat{u}_i;\kappa))\Big] + o_{\mathbb{P}}(1)$$

$$= \frac{1}{n}\sum_{i\in[n]} \int_{-\infty}^{\infty} \varphi(x)\mathsf{env}'_{\mathsf{loss}}(z_i + \alpha x;\kappa)(z_i - \mathsf{prox}_{\mathsf{loss}}(z_i + \alpha x;\kappa))\,\mathrm{d}x + o_{\mathbb{P}}(1)$$

so that the weak law of large numbers yields

$$\begin{aligned}\frac{1}{n}\boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{h} &= \mathbb{E}[\mathsf{env}_{\mathsf{loss}'}(Z+\alpha G;\kappa)(Z - \mathsf{prox}_{\mathsf{loss}}(Z+\alpha G;\kappa))] + o_{\mathbb{P}}(1)\\ &= \mathbb{E}[\mathsf{env}_{\mathsf{loss}'}(Z+\alpha G;\kappa)(\alpha G + Z - \mathsf{prox}_{\mathsf{loss}}(Z+\alpha G;\kappa) - \alpha G)] + o_{\mathbb{P}}(1)\\ &= \kappa\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)^2] - \alpha\mathbb{E}[G\cdot \mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)]\\ &= \kappa\cdot\beta^2/\delta - \alpha\cdot\nu\alpha/\delta.\end{aligned}$$

where we have used (4b) and (4d) in System 1a (with $c=1$) for the last equation. Combined with $p^{-1}\operatorname{tr}[\boldsymbol{V}] = \alpha^{-2}\kappa\beta^2 - \alpha^{-2}\boldsymbol{\psi}^\top\boldsymbol{G}\boldsymbol{h}/p + o_{\mathbb{P}}(1)$, this gives $p^{-1}\operatorname{tr}[\boldsymbol{V}] = \nu + o_{\mathbb{P}}(1)$ and finishes the proof.

**B.6. Convergence of error vector norm squared under Assumption D.** In this section we verify that Assumption C and Assumption D-(1)-(3) are sufficient for [95, Theorem 4.1] to hold. Comparing our assumptions and the condition assumed in the theorem, it suffices to show that the conditions (10) and (12) in [95] can be omitted.

Indeed, the authors used the condition (10) to show that any $\widehat{\boldsymbol{u}} \in \partial\mathsf{loss}(\boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}}^B)/\sqrt{p}$ belongs to a compact set with high probability, where $\widehat{\boldsymbol{h}}^B$ is a "bounded" estimator. More precisely, $\widehat{\boldsymbol{h}}^B$ is a solution to the constrained optimization problem $\min_{\|\boldsymbol{h}\|\le K_\alpha}\mathsf{obj}(\boldsymbol{h})$ for a positive constant $K_\alpha > 0$ where $\mathsf{obj}(\boldsymbol{h}) = \sum_{i\in[n]}\mathsf{loss}(z_i - \boldsymbol{g}_i^\top\boldsymbol{h}) + \sum_{j\in[p]}\mathsf{reg}(\sqrt{p}h_j + \theta_j)$ is the objective function for the original unconstrained M-estimator. Since loss is differentiable with Lipschitz derivative $\mathsf{loss}'$ by Assumption C, using the same argument in Lemma 14, the norm of $\widehat{\boldsymbol{u}}$ is bounded as $\|\widehat{\boldsymbol{u}}\| \le (\|\mathsf{loss}'(\boldsymbol{z})\| + \|\mathsf{loss}'\|_{\mathrm{Lip}}\|\boldsymbol{G}\|_{\mathrm{op}}K_\alpha)/\sqrt{p}$. Since $\boldsymbol{G}$ has i.i.d. $\mathcal{N}(0,1)$ entries while $\mathsf{loss}'(z_i)$ has a finite second moment by Assumption D-(1), this gives $\|\widehat{\boldsymbol{u}}\|^2 \le C$ with probability approaching to 1 for a positive constant $C$.

Next, we claim that the condition (12) in [95] is not necessary. Noting that the condition (12) is used for [95, Assumption 2-(b)], it suffices to show that the assumption 2-(b) is satisfied given our assumptions. Here we restate the assumption 2-(b) for convenience:

$$\forall \tau > 0, \quad \lim_{c\to+\infty} c^2/(2\tau) - \mathbb{E}\big[\mathsf{env}_{\mathsf{reg}}(cH + \Theta) - \mathsf{reg}(\Theta)\big] = +\infty.$$

By the condition $\mathbb{P}(\Theta \ne 0) > 0$ in Assumption D-(3), either $\mathbb{P}(\Theta > 0) > 0$ or $\mathbb{P}(\Theta < 0) > 0$ holds. Let us consider the case $\mathbb{P}(\Theta > 0) > 0$. Define a measurable function $u(H,\Theta)$ of $(H,\Theta)$ as follows:

$$u(H,\Theta) := -\min(\Theta, 1)I\{\Theta > 0 \text{ and } H < 0\}.$$

Note that $u(H,\Theta)$ is always bounded as $|u(H,\Theta)| \le 1$. By the definition of Moreau envelope, i.e., $\mathsf{env}_{\mathsf{reg}}(cH+\Theta) := \operatorname{argmin}_{p\in\mathbb{R}}(cH + \Theta - p)^2/(2\tau) + \mathsf{reg}(p)$, taking the point $p = \Theta + u(H,\Theta)$, we get the lower estimate as follows:

$$\begin{aligned}\frac{c^2}{2\tau} - \mathbb{E}\big[\mathsf{env}_{\mathsf{reg}}(cH+\Theta) - \mathsf{reg}(\Theta)\big] &\ge \frac{c^2}{2\tau} - \mathbb{E}\Big[\frac{(cH - u(H,\Theta))^2}{2\tau} + \mathsf{reg}(\Theta + u(H,\Theta)) - \mathsf{reg}(\Theta)\Big]\\ &= \frac{\mathbb{E}[uH]}{\tau}\cdot c - \frac{\mathbb{E}[u^2]}{2\tau} - \mathbb{E}\Big[\mathsf{reg}(\Theta + u) - \mathsf{reg}(\Theta)\Big].\end{aligned}$$

Thus it suffices to show that $\mathbb{E}[uH] > 0$ and $\mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)]$ is finite for the RHS to diverge as $c \to +\infty$. By the definition of $u = u(\Theta, H)$, the expectation $\mathbb{E}[uH]$ can be written as

$$\mathbb{E}[uH] = \mathbb{E}[-H \min(\Theta,1) I\{\Theta > 0, H < 0\}] = \mathbb{E}[|H| \min(\Theta,1) I\{\Theta > 0, H < 0\}].$$

Here $|H|\min(\Theta,1)$ is always strictly positive under the event $\{\Theta > 0, H > 0\}$, and this event has positive probability $\mathbb{P}(\Theta > 0, H > 0) = \mathbb{P}(\Theta > 0)\mathbb{P}(H > 0) = \mathbb{P}(\Theta > 0) \cdot 2^{-1}$ by the assumption $\mathbb{P}(\Theta > 0) > 0$ and the independence of $\Theta$ and $H \sim \mathcal{N}(0,1)$. This means that $\mathbb{E}[uH]$ is strictly positive. Regarding $\mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)]$, we have

$$\mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)] = \mathbb{E}\Big[\big\{\mathsf{reg}(\Theta - \min(\Theta,1)) - \mathsf{reg}(\Theta)\big\} I\{\Theta > 0, H < 0\}\Big].$$

Note that $0 < \Theta - \min(\Theta,1) < \Theta$ for all $\Theta > 0$. Then, by the convexity of reg and the condition $\mathsf{reg}(0) = \min_x \mathsf{reg}(x)$ in Assumption C, under the event $I\{\Theta > 0, H < 0\}$ it holds that

$$0 > \mathsf{reg}(\Theta - \min(\Theta,1)) - \mathsf{reg}(\Theta) > -d_\Theta \min(\Theta,1) > -d_\Theta \quad \text{for all } d_\Theta \in \partial \mathsf{reg}(\Theta)$$

By Assumption D-(1), $d_\Theta$ has a finite second moment for any choice of sub-derivative $d_\Theta$. Therefore, we have $0 > \mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)] > -\mathbb{E}[d_\Theta I\{\Theta > 0, H < 0\}] > -\infty$ so that $\mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)]$ is finite.

In the other case $\mathbb{P}(\Theta < 0) > 0$, we may take $u(H,\Theta) := \min(-\Theta, 1) I\{\Theta < 0, H > 0\}$. Then the same argument leads to $\mathbb{E}[uH] > 0$ and $|\mathbb{E}[\mathsf{reg}(\Theta+u) - \mathsf{reg}(\Theta)]| < +\infty$.

**B.7. Convergence of loss gradient norm squared under Assumption D.** Let $\boldsymbol{\psi} = \mathsf{loss}'(\boldsymbol{z} - \boldsymbol{G}\boldsymbol{h})$ and $\mathsf{loss}^*$ be the conjugate of loss. By the same argument in [95], restricting the range of $\boldsymbol{\psi}$ to a compact set so that the strong duality holds, we observe that $\boldsymbol{\psi}$ is a solution to the following min-max problem with probability approaching to 1:

$$\max_{\boldsymbol{\psi} \in \mathbb{R}^n} \min_{\boldsymbol{h} \in \mathbb{R}^p} \boldsymbol{\psi}^\top (\boldsymbol{z} - \boldsymbol{G}\boldsymbol{h}) - \mathsf{loss}^*(\boldsymbol{\psi}) + \mathsf{reg}(\sqrt{p}\boldsymbol{h} + \boldsymbol{\theta}) = \max_{\boldsymbol{\psi}} \boldsymbol{\psi}^\top \boldsymbol{z} - \mathsf{loss}^*(\boldsymbol{\psi}) + \boldsymbol{\psi}^\top \boldsymbol{G}\boldsymbol{\theta}/\sqrt{p} - \mathsf{reg}^*(\boldsymbol{G}^\top \boldsymbol{\psi}/\sqrt{p})$$

If we write $\widehat{\boldsymbol{u}} = \boldsymbol{\psi}/\sqrt{p} \in \mathbb{R}^n$ then $\widehat{\boldsymbol{u}}$ is the M-estimator of the form:

$$\widehat{\boldsymbol{u}} \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \mathsf{F}(-\boldsymbol{G}^\top \boldsymbol{u}) + \mathsf{L}(\sqrt{n}\boldsymbol{u}), \quad \text{where} \quad \begin{aligned} &\mathsf{F} : \mathbb{R}^p \mapsto \mathbb{R}, \quad \boldsymbol{v} \mapsto \mathsf{reg}^*(-\boldsymbol{v}) + \boldsymbol{\theta}^\top \boldsymbol{v} \\ &\mathsf{L} : \mathbb{R}^n \mapsto \mathbb{R}, \quad \boldsymbol{u} \mapsto \mathsf{loss}^*(\sqrt{\frac{p}{n}}\boldsymbol{w}) - \sqrt{\frac{p}{n}}\boldsymbol{z}^\top \boldsymbol{w} \end{aligned}$$

For any $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}_n, \boldsymbol{I}_n)$, for all $c \in \mathbb{R}$ and $\tau > 0$ it holds that

$$\frac{1}{p}\big(\mathsf{env}_{\mathsf{F}}(c\boldsymbol{h}; \tau) - \mathcal{F}(\boldsymbol{0})\big) \xrightarrow{\mathrm{p}} \frac{c^2}{2\tau} - \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(\frac{c}{\tau}H + \Theta_0; \frac{1}{\tau}) - \mathsf{reg}(0)]$$

$$\frac{1}{n}\big(\mathsf{env}_{\mathsf{L}}(c\boldsymbol{g}; \tau) - \mathcal{L}(\boldsymbol{0})\big) \xrightarrow{\mathrm{p}} \frac{c^2}{2\tau} - \mathbb{E}[\mathsf{env}_{\mathsf{loss}}(\sqrt{\delta}\frac{c}{\tau}G + Z; \frac{\delta}{\tau}) - \mathsf{loss}(0)]$$

Thus, letting $F(c,\tau) := \mathbb{E}[\mathsf{env}_{\mathsf{reg}}(cH + \Theta; \tau)]$ and $L(c,\tau) := \mathbb{E}[\mathsf{env}_{\mathsf{loss}}(cG + \Theta; \tau)]$, [95] implies that $\|\widehat{\boldsymbol{u}}\|^2 \xrightarrow{\mathrm{p}} \widetilde{\alpha}^2$ where $\widetilde{\alpha}$ is the minimizer of the min-max optimization problem:

$$\inf_{\widetilde{\alpha}, \widetilde{\tau}_g} \sup_{\widetilde{\beta}, \widetilde{\tau}_h} \frac{\widetilde{\beta}\widetilde{\tau}_g}{2} + \delta^{-1}\Big(\frac{\widetilde{\alpha}^2\widetilde{\beta}}{2\widetilde{\tau}_g} - F(\frac{\widetilde{\alpha}\widetilde{\beta}}{\widetilde{\tau}_g}, \frac{\widetilde{\beta}}{\widetilde{\tau}_g})\Big) - \frac{\widetilde{\alpha}\widetilde{\tau}_h}{2} - \frac{\widetilde{\alpha}\widetilde{\beta}^2}{2\widetilde{\tau}_h} + \Big(\frac{\widetilde{\alpha}\widetilde{\beta}^2}{2\widetilde{\tau}_h} - L(\sqrt{\delta}\widetilde{\beta}, \delta\frac{\widetilde{\tau}_h}{\widetilde{\alpha}})\Big)$$

Thus, by the change of variables $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\tau}_g, \widetilde{\tau}_h) \mapsto (\beta, \alpha/\sqrt{\delta}, \tau_h/\sqrt{\delta}, \tau_g/\delta)$ and multiplying the potential by $-\delta$, we are left with $\|\widehat{\boldsymbol{u}}\|^2 \xrightarrow{\mathrm{p}} \beta_*^2$ where $\beta_*$ is the minimizer of

$$\sup_{\beta\tau_h} \inf_{\alpha,\tau_g} -\frac{\alpha\tau_h}{2} - \frac{\beta^2\alpha}{2\tau_g} + F(\frac{\beta\alpha}{\tau_h}, \frac{\alpha}{\tau_h}) + \frac{\beta\tau_g}{2} + \delta L(\alpha, \frac{\tau_g}{\beta}),$$

which is the same potential in [95].

**B.8. Explicit expressions for degrees of freedom.** See Table 6.

TABLE 6

*Explicit formulae for the degrees of freedom and residual degrees of freedom of the estimator $\widehat{\boldsymbol{\theta}}_I$ defined using (2). Here the loss functions $\mathsf{loss}(r)$ are: 1) squared loss: $r^2/2$, 2) Huber loss: $r^2/2$ for $|r| \le 1$ and $|r| - 1/2$ for $|r| > 1$; and the regularization functions $\mathsf{reg}(b)$ are: 1) ridge penalty: $\frac{\lambda_1}{2} b^2$, 2) lasso penalty: $\lambda_2 |b|$, and 3) elastic net penalty: $\frac{\lambda_1}{2} b^2 + \lambda_2 |b|$. And other quantities are: 1) $\widehat{S}_I = \{j \in [p] : \widehat{\boldsymbol{\theta}}_I(j) \neq 0\}$ is the set of active variables of $\widehat{\boldsymbol{\theta}}_I$ and $\boldsymbol{X}_{\widehat{S}_I}$ is the submatrix of $\boldsymbol{X}_I$ made of columns indexed in $\widehat{S}_I$, 2) $\widehat{T}_I = \{i \in I : \mathsf{loss}''(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_I) > 0\}$ is the set of detected inliers (active observations), and 3) the matrix $\boldsymbol{D}_I = \mathrm{diag}(\mathsf{loss}''(\boldsymbol{y}_I - \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I))$.*

| **Loss** (loss) | **Regularizer** (reg) | **Degrees of freedom** (df) | **Residual degrees of freedom** ($\mathrm{tr}[\boldsymbol{V}_I]$) |
|---|---|---|---|
| Square | Ridge | $\mathrm{tr}\left[\left(\boldsymbol{X}_I^\top \boldsymbol{X}_I + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_I^\top \boldsymbol{X}_I\right]$ | $\lvert I\rvert - \mathsf{df}_I$ |
| Square | Lasso | $\lvert\widehat{S}_I\rvert$ | $\lvert I\rvert - \mathsf{df}_I$ |
| Square | Elastic net | $\mathrm{tr}\left[\left(\boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{X}_{\widehat{S}_I} + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{X}_{\widehat{S}_I}\right]$ | $\lvert I\rvert - \mathsf{df}_I$ |
| Huber | Ridge | $\mathrm{tr}\left[\left(\boldsymbol{X}_I^\top \boldsymbol{D}_I \boldsymbol{X}_I + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_I^\top \boldsymbol{D}_I \boldsymbol{X}_I\right]$ | $\lvert\widehat{T}_I\rvert - \mathsf{df}_I$ |
| Huber | Lasso | $\lvert\widehat{S}_I\rvert$ | $\lvert\widehat{T}_I\rvert - \mathsf{df}_I$ |
| Huber | Elastic net | $\mathrm{tr}\left[\left(\boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{D}_I \boldsymbol{X}_{\widehat{S}_I} + \lambda_1 \boldsymbol{I}\right)^{-1} \boldsymbol{X}_{\widehat{S}_I}^\top \boldsymbol{D}_I \boldsymbol{X}_{\widehat{S}_I}\right]$ | $\lvert\widehat{T}_I\rvert - \mathsf{df}_I$ |
| Convex | Convex | $\mathrm{tr}[(\partial/\partial \boldsymbol{y}_I) \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I]$ | $\mathrm{tr}[(\partial/\partial \boldsymbol{y}_I) \mathsf{loss}'(\boldsymbol{y}_I - \boldsymbol{X}_I \widehat{\boldsymbol{\theta}}_I)]$ |

## APPENDIX C: PROOFS FOR RESULTS IN SECTION 4

**C.1. Proof of Proposition 7.** Since the risk limit $\mathcal{R}_M$ is given by $\mathcal{R}_M = M^{-1}\alpha^2 + (1 - M^{-1})\alpha^2 \eta_{\scriptscriptstyle G}$ and $\eta_{\scriptscriptstyle G}$ is non-negative in the homogeneous case (see the discussion in Appendix A), it suffices to show $\eta_{\scriptscriptstyle G} < 1$. We proceed by contradiction. Let $\eta_{\scriptscriptstyle H}$ be the associated scalar such that $\eta_{\scriptscriptstyle G} = F_{\mathsf{reg}}(\eta_{\scriptscriptstyle H})$ and $\eta_{\scriptscriptstyle H} = F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G})$. Note that $|\eta_{\scriptscriptstyle H}| = |F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G})| \le c < 1$ by $c < 1$. If $|\eta_{\scriptscriptstyle G}| = 1$ then the equality case $F_{\mathsf{reg}}(\eta_{\scriptscriptstyle H}) = 1$ for the Cauchy–Schwarz inequality implies that with probability 1,

$$\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu} H + \Theta; \frac{1}{\nu}\Big) - \frac{\beta}{\nu} H = \frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu} \widetilde{H} + \Theta; \frac{1}{\nu}\Big) - \frac{\beta}{\nu} \widetilde{H}$$

with $\widetilde{H} := \eta_{\scriptscriptstyle H} H + \sqrt{1 - \eta_{\scriptscriptstyle H}^2} H_0$ for any independent standard normals $(H, H_0)$. Multiplying the above display by $H_0$ and taking the expectation, the LHS becomes 0 by the independence of $(H, H_0)$ and $\mathbb{E}[H_0] = 0$. As for the RHS, using (4c) in System 1a and Stein's lemma, we have

$$\mathbb{E}\Big[\Big(\frac{1}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\Big(\frac{\beta}{\nu} \widetilde{H} + \Theta; \frac{1}{\nu}\Big) - \frac{\beta}{\nu} \widetilde{H}\Big) H_0\Big] = \sqrt{1 - \eta_{\scriptscriptstyle H}^2}(-\kappa\beta).$$

Thus, we get $0 = \sqrt{1 - \eta_{\scriptscriptstyle H}^2}(-\kappa\beta)$, which is a contradiction since $\kappa\beta > 0$ and $|\eta_{\scriptscriptstyle H}| < 1$.

**C.2. Proof of Proposition 8.** Let us fix $\psi = (c\delta)^{-1}$ and take the derivative of $\mathcal{R}_M = M^{-1}\alpha^2 + (1 - M^{-1})\alpha^2 \eta_{\scriptscriptstyle G}$ with respect to $\phi = \delta^{-1}$. Note $c = \psi/\phi < 1$. With this parameterization, $(\alpha, \beta, \kappa, \nu)$ are all fixed since they depend on $\psi$ only. Below, we derive the partial derivative of $\eta_{\scriptscriptstyle G}$ with respect to $\phi$. Using $\psi$ and $\phi$, we observe that $\eta_{\scriptscriptstyle G}$ is the unique solution to the fixed point equation

$$\eta_{\scriptscriptstyle G} = F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G}; \phi).$$

Here, $F_{\mathsf{reg}}$ does not depend on $\phi$ but $F_{\mathsf{loss}}$ depends on $\phi$ as:

$$F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G};\phi) = \frac{\phi}{\psi^2\beta^2}\cdot\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z;\kappa)\cdot\mathsf{env}'_{\widetilde{\mathsf{loss}}}(\alpha\widetilde{G} + Z;\kappa)],$$

Since the map $\eta_{\scriptscriptstyle G} \mapsto F_{\mathsf{reg}} \circ F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G};\phi)$ is differentiable and $c$-Lipschitz with $c = \psi/\phi < 1$ (see Theorem 1), the implicit function theorem implies that $\eta_{\scriptscriptstyle G}$ is differentiable with respect to $\phi$ and the derivative satisfies:

$$\frac{\partial\eta_{\scriptscriptstyle G}}{\partial\phi} = (F_{\mathsf{reg}}\circ F_{\mathsf{loss}})'(\eta_{\scriptscriptstyle G})\frac{\partial\eta_{\scriptscriptstyle G}}{\partial\phi} + F'_{\mathsf{reg}}(\eta_{\scriptscriptstyle H})\frac{F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G})}{\phi}.$$

Rearranging the above display, we get

$$\frac{\partial\eta_{\scriptscriptstyle G}}{\partial\phi} = \frac{F'_{\mathsf{reg}}(\eta_{\scriptscriptstyle H})}{\phi}\frac{F_{\mathsf{loss}}(\eta_{\scriptscriptstyle G})}{1-(F_{\mathsf{reg}}\circ F_{\mathsf{loss}})'(\eta_{\scriptscriptstyle G})} = \frac{F'_{\mathsf{reg}}(\eta_{\scriptscriptstyle H})}{\phi}\frac{\eta_{\scriptscriptstyle H}}{1-(F_{\mathsf{reg}}\circ F_{\mathsf{loss}})'(\eta_{\scriptscriptstyle G})} \geq 0.$$

where the last inequality follows from the fact that $F_{\mathsf{reg}}$ is non-decreasing (Theorem 1) and $\eta_{\scriptscriptstyle H}$ is non-negative (see the discussion on homogeneous cases in Appendix A). Combined with $\partial_\phi\mathcal{R}_M = (1-M^{-1})\alpha^2\frac{\partial\eta_G}{\partial\phi}$, we observe that $\mathcal{R}_M$ is non-decreasing in $\phi$. Therefore, for any two $\phi_1 \leq \phi_2$, we have that

$$\mathcal{R}_M(\phi_2,\psi) \geq \mathcal{R}_M(\phi_1,\psi) \geq \inf_{\psi>\phi_2}\mathcal{R}_M(\phi_1,\psi) \geq \inf_{\psi>\phi_1}\mathcal{R}_M(\phi_1,\psi) \quad \text{for all } \psi \geq \phi_2.$$

This gives $\inf_{\psi>\phi_2}\mathcal{R}_M(\phi_2,\psi) \geq \inf_{\psi>\phi_1}\mathcal{R}_M(\phi_1,\psi)$ for any $\phi_1 \leq \phi_2$, which means the map $\phi \mapsto \inf_{\psi>\phi}\mathcal{R}_M(\phi,\psi)$ is non-decreasing. Reverting to the original parametrization $(\phi,\psi) \mapsto (\delta, c) = (\phi^{-1},\psi/\phi)$, we conclude that $\delta \mapsto \inf_{c\in(0,1)}\mathcal{R}_M(\delta,c)$ is non-increasing.

**C.3. Derivation of System 2.** We will first reformulate Systems 1a and 1b in a slightly different set of parameters. Since the purpose of this reparameterization is to match with existing work, we will also consider regularizer $\mathsf{reg}$ with an explicit regularization level $\lambda$. The mapping with respect to the parameters in Systems 1a and 1b is as follows: $a = \frac{\lambda}{\beta}$, $\tau = \sqrt{c\delta}\frac{\beta}{\nu}$. Under this parameterization, note that $\frac{\beta}{\nu} = \frac{\tau}{\sqrt{c\delta}}$ and $\frac{\lambda}{\nu} = \frac{a\tau}{\sqrt{c\delta}}$.

REMARK 9 (Scaling differences in design). It is worth remarking that in the literature on risk characterization of regularized M-estimator under proportional asymptotics, the scaling of $\lambda$ can be slightly different, up to a factor of $\sqrt{c\delta}$. One of the reasons for the differences is how the design matrix $\boldsymbol{X}$ is scaled. We assume that the entries of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ each have variance $1/p$ and thus each row has a unit average norm squared. It is also common to assume that the entries of $\boldsymbol{X}$ each have variance $1/n$ and thus each column has a unit average norm squared. This brings in a factor of $\sqrt{\delta}$. For the subsampled design $\boldsymbol{X}_I \in \mathbb{R}^{k\times p}$, we get an additional factor of $\sqrt{c}$. Consequently, some expressions may appear different up to this scaling.

DERIVATION OF SYSTEM 2. The derivation is straightforward. We will use some simple relationships between proximal operators and Moreau envelopes. Recall from (1) that $\mathsf{env}'_f(x;\tau) = \frac{1}{\tau}(x - \mathsf{prox}_f(x;\tau))$ so that the derivative identity $\mathsf{env}''_f(x;\tau) = \frac{1}{\tau}(1 - \mathsf{prox}'_f(x;\tau))$ holds for almost every $x$ by the non-expansiveness of the proximal operator.

*(1) Case of $m = \ell$.* The original system of equations is given by System 1a. For regularizers of the form $\lambda$reg, from (4a), we have

$$\begin{aligned}\alpha^2 &= \mathbb{E}\Big[\big(\tfrac{\lambda}{\nu}\cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \tfrac{\beta}{\nu}H\big)^2\Big]\\ &= \mathbb{E}\Big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\beta}{\nu}H + \Theta; \tfrac{\lambda}{\nu}\big) - \Theta\big)^2\Big]\\ &= \mathbb{E}\Big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\big)^2\Big] \end{aligned} \tag{48}$$

where the variables $\tau$ and $a$ are defined as:

$$\tau = \sqrt{c\delta}\frac{\beta}{\nu} \overset{(4b),(4d)}{=\!=\!=} \frac{\alpha\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]}}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)\cdot G]}, \tag{49}$$

$$a = \frac{\lambda}{\beta} \overset{(4b)}{=\!=} \frac{\lambda}{\sqrt{c\delta}\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]}}. \tag{50}$$

For squared loss $\mathsf{loss}(x) = x^2/2$, since $\mathsf{env}'_{\mathsf{loss}}(x;\tau) = x/(1+\tau)$ from Table 7, squaring the final expression in (49) yields

$$\tau^2 = \alpha^2 + \sigma^2. \tag{51}$$

Thus, $\tau^2$ and $\alpha^2$ are the limiting total and excess risks, respectively. Combining (48) and (51) gives the first desired equation (16a).

Similarly, (4c) after diving by $\tau$ yields

$$\frac{\kappa\beta}{\tau} = \frac{\beta}{\nu\tau} - \frac{\lambda}{\nu\tau}\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\cdot H\big].$$

Multiplying both sides by $\sqrt{c\delta}a\tau$ and noting that $a\beta = \lambda$, $\frac{\beta}{\nu} = \frac{\tau}{\sqrt{c\delta}}$, $\frac{\lambda}{\nu} = \frac{a\tau}{\sqrt{c\delta}}$ then gives

$$\sqrt{c\delta}\kappa\lambda = a\tau - \frac{(a\tau)^2}{\tau}\mathbb{E}\big[\mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\cdot H\big]. \tag{52}$$

From Stein's lemma, we have

$$\mathbb{E}[\mathsf{env}'_{\mathsf{reg}}(\tau H + \Theta; \kappa)\cdot H] = \tau\mathbb{E}[\mathsf{env}''_{\mathsf{reg}}(\tau H + \Theta; \kappa)], \tag{53}$$

and so (52) reduces to

$$\begin{aligned}\sqrt{c\delta}\kappa\lambda &\overset{(53)}{=\!=} a\tau - \frac{(a\tau)^2}{\sqrt{c\delta}}\mathbb{E}\big[\mathsf{env}''_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]\\ &= a\tau - a\tau\mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]\\ &= a\tau\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]. \end{aligned} \tag{54}$$

Using similar manipulations as above, (49) reduces to

$$\tau = \frac{\kappa\sqrt{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G + Z; \kappa)^2]}}{\mathbb{E}[1 - \mathsf{prox}'_{\mathsf{loss}}(\alpha G + Z; \kappa)]}. \tag{55}$$

Combining (54) and (55), we get

$$0 = \kappa\lambda\Big(1 - \frac{a\tau}{\sqrt{c\delta}\kappa\lambda}\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H + \Theta; \tfrac{a\tau}{\sqrt{c\delta}}\big)\big]\Big). \tag{56}$$

From (50), under squared loss, we also have

$$1 + \kappa = \frac{\sqrt{c\delta}a\tau}{\lambda}. \tag{57}$$

Combining (56) with (57) then leads to the second desired equation (16b).

*(2) Case of $m \neq \ell$.* Step 2. $(\eta_G, \eta_H)$ is the solution to the following 2-scalar fix-point equations:

$$\eta_G = \frac{\mathbb{E}\Big[\big(\frac{\lambda}{\nu}\cdot \mathsf{env}'_{\mathsf{reg}}\big(\frac{\beta}{\nu}H+\Theta;\frac{\lambda}{\nu}\big)-\frac{\beta}{\nu}H\big)\cdot\big(\frac{\lambda}{\nu}\cdot \mathsf{env}'_{\mathsf{reg}}\big(\frac{\beta}{\nu}\widetilde{H}+\Theta;\frac{\lambda}{\nu}\big)-\frac{\beta}{\nu}\widetilde{H}\big)\Big]}{\mathbb{E}\Big[\big(\frac{\lambda}{\nu}\cdot \mathsf{env}'_{\mathsf{reg}}\big(\frac{\beta}{\nu}H+\Theta;\frac{\lambda}{\nu}\big)-\frac{\beta}{\nu}H\big)^2\Big]}$$

$$\eta_H = c\cdot\frac{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G+Z;\kappa)\cdot \mathsf{env}'_{\mathsf{loss}}(\alpha\widetilde{G}+Z;\kappa)]}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G+Z;\kappa)^2]},$$

where $\binom{G}{\widetilde{G}} \sim N\Big(\binom{0}{0}, \begin{pmatrix}1 & \eta_G\\ \eta_G & 1\end{pmatrix}\Big)$ and $\binom{H}{\widetilde{H}} \sim N\Big(\binom{0}{0}, \begin{pmatrix}1 & \eta_H\\ \eta_H & 1\end{pmatrix}\Big)$. Similarly, by variable substitution, we can rewrite the equations as:

(58a) $$\eta_G\alpha^2 = \mathbb{E}\Big[\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}H+\Theta;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\big)\big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}+\Theta;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\big)\Big]$$

(58b) $$\eta_H = c\cdot\frac{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G+Z;\kappa)\cdot \mathsf{env}'_{\mathsf{loss}}(\alpha\widetilde{G}+Z;\kappa)]}{\mathbb{E}[\mathsf{env}'_{\mathsf{loss}}(\alpha G+Z;\kappa)^2]}.$$

Since $\mathsf{env}'_{\mathsf{loss}}(x;\tau) = x/(1+\tau)$ for squared loss, we obtain the third desired equation (17a). □

**C.4. Proximal operators and Moreau envelopes.** See Table 7.

TABLE 7

*Proximal operators, Moreau envelopes, and their derivatives for ridge (first row) and lasso (second row) regularizers and Huber (third row) loss considered in Sections 4.2.1 and 4.2.2.*

| $f(x)$ | $\mathsf{prox}_f(x;\tau)$ | $\mathsf{prox}'_f(x;\tau)$ | $\mathsf{env}_f(x;\tau)$ | $\mathsf{env}'_f(x;\tau)$ |
|---|---|---|---|---|
| $\frac{1}{2}x^2$ | $\frac{x}{1+\tau}$ | $\frac{1}{1+\tau}$ | $\frac{1}{2}\frac{x^2}{1+\tau}$ | $\frac{x}{1+\tau}$ |
| $\lvert x\rvert$ | $(\lvert x\rvert-\tau)_+\operatorname{sign}(x)$ | $\mathbb{1}\{\lvert x\rvert\geq\tau\}$ | $\begin{cases}\frac{1}{2\tau}x^2 & \lvert x\rvert<\tau\\ \lvert x\rvert-\frac{1}{2}\tau & \lvert x\rvert\geq\tau\end{cases}$ | $\min\left\{\frac{\lvert x\rvert}{\tau},1\right\}\operatorname{sign}(x)$ |
| $\begin{cases}\frac{x^2}{2} & \lvert x\rvert\leq 1\\ \lvert x\rvert-\frac{1}{2} & \lvert x\rvert>1\end{cases}$ | $\begin{cases}\frac{x}{1+\tau} & \lvert x\rvert\leq 1+\tau\\ x-\tau\operatorname{sign}(x) & \lvert x\rvert>1+\tau\end{cases}$ | $\begin{cases}\frac{1}{1+\tau} & \lvert x\rvert\leq 1+\tau\\ 1 & \lvert x\rvert>1+\tau\end{cases}$ | $\begin{cases}\frac{1}{2}\frac{x^2}{1+\tau} & \lvert x\rvert\leq 1+\tau\\ \lvert x\rvert-\tau-\frac{1}{2} & \lvert x\rvert>1+\tau\end{cases}$ | $\left(\frac{\lvert x\rvert}{1+\tau}-1\right)_+\operatorname{sign}(x)$ |

## APPENDIX D: PROOFS FOR RESULTS IN SECTION 5

**D.1. Proof of (24).** Applying Stein's lemma to (21b) and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
1 &= \frac{1}{c\delta}\mathbb{E}\Big[\mathsf{prox}'_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)\Big]\\
&= \frac{1}{\sqrt{c\delta}}\frac{1}{\tau}\mathbb{E}\Big[H\cdot\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)\Big]\\
&\leq \frac{1}{\sqrt{c\delta}\tau}\mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta+\tfrac{\tau}{\sqrt{c\delta}}H;\tfrac{a\tau}{\sqrt{c\delta}}\big)-\Theta\Big)^2\Big]^{1/2}\\
&= \frac{1}{\sqrt{c\delta}\tau}\sqrt{\tau^2-\sigma^2} \qquad \text{(by (21a)).}
\end{aligned}$$

Taking the square of both sides and rearranging the resulting $\tau^2$ term, we get the lower estimate $\tau^2 \geq \sigma^2/(1-c\delta)$ as desired.

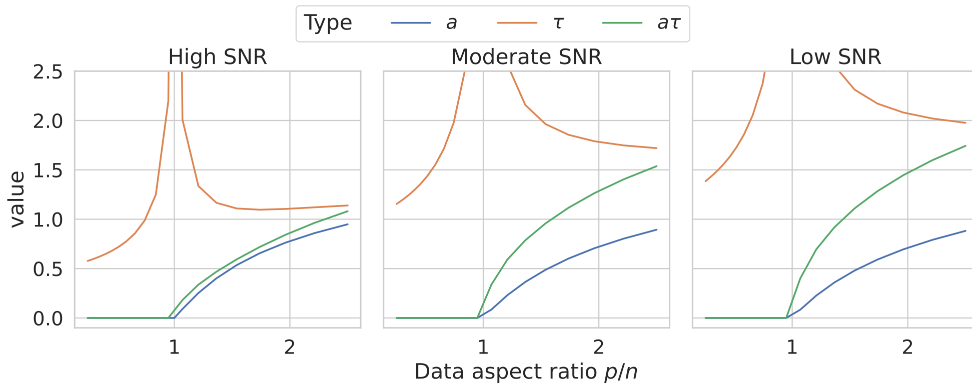


Fig 12: Fixed-point quantities for lassoless at different data aspect ratios $p/n$. The data model is given by (61) with signal strength $\rho = 0.5$ and sparsity levels $s = 0.2$ at different noise levels $\sigma$. *Left*: High SNR $\sigma = 0.5$. *Middle*: Moderate SNR $\sigma = 1$. *Right*: Low SNR $\sigma = 1.2$.

**D.2. Derivation of System 4 as a limit of System 2 as $\lambda \to 0^+$.** Fix $\mathrm{reg}(x) = |x|^q$ for $q \in \{1, 2\}$. Let $(a_\lambda, \tau_\lambda, \xi_\lambda)$ be the solution to System 2 for any $\lambda > 0$ and let $(a_*, \tau_*, \xi_*)$ be the solution to System 4. Note that previous papers showed that $(a_\lambda, \tau_\lambda)$ converges to the solution $(a_*, \tau_*)$ as $\lambda \to (0)^+$; $q = 1$ case is given by the proof of Lemma A.1 in [70] while the $q = 2$ case immediately follows from the explicit formulae of $(a_\lambda, \tau_\lambda)$ and $(a_*, \tau_*)$. Below we claim the continuity of $\xi$, i.e., $\xi_\lambda \to \xi_*$. Denoting $\eta_\lambda = c\xi_\lambda^2/\tau_\lambda^2$ and $\eta_* = c\xi_*^2/\tau_*^2$, what we want to show is $\eta_\lambda \to \eta_*$.

Observe that the systems for $\xi_\lambda$ and $\xi_*$ read to $\eta_\lambda = F(\eta_\lambda; \tau_\lambda, a_\lambda)$ and $\eta_* = F(\eta_*; \tau_*, a_*)$ where

$$F(\eta; \tau, a) := \frac{c}{\tau^2}\Big\{\mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\Big) \cdot \Big(\mathsf{prox}_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta\Big)\Big] + \sigma^2\Big\}$$

with $(H, \widetilde{H})$ being the mean zero jointly normals such that $\mathbb{E}[H^2] = \mathbb{E}[\widetilde{H}^2] = 1$ and $\mathbb{E}[H\widetilde{H}] = \eta$. Note that the map $\eta \mapsto F(\eta; \tau_\lambda, a_\lambda)$ is $c$-Lipschitz over $[-1, 1]$ by the same argument in Appendix B.1, while the map $(\tau, a) \mapsto F(\eta; \tau, a)$ is continuous over $(0, \infty)^2$ for any $\eta \in [-1, 1]$ by the moment assumption in Assumption D-1 and the dominated convergence theorem. Then, $|\eta_\lambda - \eta_*|$ is bounded from above as

$$\begin{aligned}
|\eta_\lambda - \eta_*| &= |F(\eta_\lambda; \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)| \\
&\le |F(\eta_\lambda; \tau_\lambda, a_\lambda) - F(\eta_*, \tau_\lambda, a_\lambda)| + |F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)| \\
&\le c|\eta_\lambda - \eta_*| + |F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)|
\end{aligned}$$

so that $|\eta_\lambda - \eta_*| \le (1-c)^{-1}|F(\eta_*, \tau_\lambda, a_\lambda) - F(\eta_*, \tau_*, a_*)|$ holds. This upper bound converges to 0 as $\lambda \to 0$ by the continuity of $F(\eta_*, \tau, a)$ in $(\tau, a)$ and the convergence $(\tau_\lambda, a_\lambda) \to (\tau_*, a_*)$.

**D.3. Additional details for Remark 7.** Expanding the product term in the fixed point equation (22a) by

$$\mathsf{prox}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) - \Theta = \tfrac{\tau}{\sqrt{c\delta}}H - \tfrac{a\tau}{\sqrt{c\delta}}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big),$$

we have

$$\begin{aligned}
\xi^2 - \sigma^2 = {}& \frac{\tau^2}{c\delta}\mathbb{E}[H\widetilde{H}] - 2\frac{a\tau^2}{c\delta}\mathbb{E}\Big[\widetilde{H}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
&+ \frac{(a\tau)^2}{c\delta}\mathbb{E}\Big[\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big) \cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big],
\end{aligned}$$

where $\frac{\tau^2}{c\delta}\mathbb{E}[H\widetilde{H}] = \frac{\tau^2}{c\delta}\eta_{H} = \frac{\xi^2}{\delta}$ by the definition $\eta_{H} = c\xi^2/\tau^2$. For the second term, realizing $\widetilde{H} = \eta_{H} H + \sqrt{1-\eta_{H}^2}\,\overline{H}$ for a standard normal $\overline{H} \sim \mathcal{N}(0,1)$ independent of $(H, \widetilde{H}, \Theta)$, noting that $\mathsf{env}'_{\mathsf{reg}}(\Theta + \frac{\tau}{\sqrt{c\delta}}H)$ is bounded in second moment by Assumption D-(1), we have

$$\begin{aligned}
\mathbb{E}\Big[\widetilde{H}\cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] &= \eta_{H}\mathbb{E}\Big[H\cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
&= \eta_{H}\tfrac{\tau}{\sqrt{c\delta}}\mathbb{E}\Big[\mathsf{env}''_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] && \text{(by Stein's lemma)} \\
&= \eta_{H}\tfrac{\tau}{\sqrt{c\delta}}\frac{\sqrt{c\delta}}{a\tau}\mathbb{E}\Big[1 - \mathsf{prox}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big] \\
&= \frac{\eta_{H}}{a}(1 - c\delta) && \text{(by (21b))}.
\end{aligned}$$

Combined with $\eta_{H} = c\xi^2/\tau^2$, we have

$$\frac{a\tau^2}{c\delta}\mathbb{E}\Big[\widetilde{H}\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H\big)\Big] = \frac{\xi^2}{\delta}(1 - c\delta).$$

Therefore, we get

$$\xi^2\Big(1 - \frac{1}{\delta} + 2\frac{(1-c\delta)}{\delta}\Big) = \sigma^2 + \frac{(a\tau)^2}{c\delta}\mathbb{E}\Big[\mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}H; \tfrac{a\tau}{\sqrt{c\delta}}\big)\cdot \mathsf{env}'_{\mathsf{reg}}\big(\Theta + \tfrac{\tau}{\sqrt{c\delta}}\widetilde{H}; \tfrac{a\tau}{\sqrt{c\delta}}\big)\Big].$$

This means that $\xi^2 \to \frac{\delta}{\delta-1}\sigma^2$ holds if and only if the rightmost term converges to $0$ as $c \to (\delta^{-1})^-$. This is true if reg is Lipschitz and $\lim_{c\to(\delta^{-1})^-} a\tau = 0$, as $|\mathsf{env}'_{\mathsf{reg}}(x;\tau)| \le \|\mathsf{reg}\|_{\mathrm{Lip}}$ for all $x \in \mathbb{R}$ and $\tau > 0$. However, we are not able to provably establish that $a\tau \to 0$ as $c \to \delta^{-1}$. For the lasso regularizer, we observe in Figure 12 that $a\tau \to 0$ appears to hold as $c \to \delta^{-1}$.

## APPENDIX E: PROOF FOR RESULTS IN SECTION 6

**E.1. Proof of Theorem 9.** Here the proof is similar to Appendix B.1. Our goal is to show

1. $|F_{\mathsf{loss}}(\eta_{G})| \le \sqrt{c\widetilde{c}}$ and $|F_{\mathsf{Reg}}(\eta_{H})| \le 1$ for all $\eta_{G} \in [-1,1]$ and $\eta_{H} \in [-1,1]$.
2. $F_{\mathsf{loss}}$ and $F_{\mathsf{Reg}}$ are non-decreasing, differentiable, and the compositions $F_{\mathsf{loss}} \circ F_{\mathsf{Reg}}$ and $F_{\mathsf{Reg}} \circ F_{\mathsf{loss}}$ are $\min\{c, \widetilde{c}\}$-Lipschitz.
3. System 5b admits a unique solution $(\eta_{G}^\star, \eta_{H}^\star) \in [-1,1] \times [-\sqrt{c\widetilde{c}}, \sqrt{c\widetilde{c}}]$.

The first claim immediately follows from the definition of $(F_{\mathsf{loss}}, F_{\mathsf{Reg}})$, the Cauchy–Schwarz inequality, and (27a)-(27b). The third claim follows from the second claim by the argument in Appendix B.1 using Brouwer's fixed-point theorem. Thus, it suffices to show the second claim.

For $F_{\mathsf{loss}}$, we have already shown in Appendix B.1 that $F_{\mathsf{loss}}$ is differentiable with

$$0 \le F'_{\mathsf{loss}}(\eta_{G}) \le \frac{\alpha\widetilde{\alpha}}{\beta\widetilde{\beta}}\Big(\frac{c\widetilde{\nu}}{\kappa} \wedge \frac{\widetilde{c}\nu}{\widetilde{\kappa}}\Big).$$

For the derivative of $F_{\mathsf{Reg}}$, we will use the following lemma that generalizes Lemma 12.

LEMMA 26. *Given two Lipschitz functions $\boldsymbol{f}, \widetilde{\boldsymbol{f}} : \mathbb{R}^p \to \mathbb{R}^p$, define $\phi : [-1,1] \to \mathbb{R}$ as*

$$\phi(t) = \mathbb{E}[\boldsymbol{f}(\boldsymbol{h})^\top \widetilde{\boldsymbol{f}}(t\boldsymbol{h} + \sqrt{1-t^2}\boldsymbol{h}')]$$

*where the expectation is taken with respect to independent standard normal vectors* $\boldsymbol{h}$, $\boldsymbol{h}' \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)$. *Then the map* $\phi$ *is differentiable with its derivative given by*

$$\phi'(t) = \mathbb{E}\Big[\operatorname{tr}\Big(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{h})^\top \frac{\partial \widetilde{\boldsymbol{f}}}{\partial \boldsymbol{x}}(t\boldsymbol{h} + \sqrt{1-t^2}\boldsymbol{h}')\Big)\Big]$$

*where* $\partial \boldsymbol{f}/(\partial \boldsymbol{x})$ *and* $\partial \widetilde{\boldsymbol{f}}/(\partial \boldsymbol{x})$ *are the weak derivatives of* $\boldsymbol{f}$ *and* $\widetilde{\boldsymbol{f}}$.

We omit the proof of this lemma, as it follows directly from the proof of Lemma 12. Now we apply this lemma with Lemma 27 below. Here we recall the generalized Moreau envelope:

$$\forall \boldsymbol{v} \in \mathbb{R}^p,\ \forall \boldsymbol{\Lambda} \succ \mathbf{0}_{p\times p} \quad \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda}) := \min_{\boldsymbol{x}\in\mathbb{R}^p} \frac{1}{2}(\boldsymbol{v}-\boldsymbol{x})^\top \boldsymbol{\Lambda}^{-1}(\boldsymbol{v}-\boldsymbol{x}) + \mathsf{Reg}(\boldsymbol{x})$$

and denote the unique minimizer by $\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda})$. Recall $\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^{-1}(\boldsymbol{v} - \mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda}))$, where $\nabla$ is the gradient with respect to $\boldsymbol{v}$.

LEMMA 27. *Fix* $\boldsymbol{\Lambda} \succ \mathbf{0}_{p\times p}$. *Then, for any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, *we have*

$$\|\boldsymbol{\Lambda}^{1/2}(\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) - \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda}))\|_2 \le \|\boldsymbol{\Lambda}^{-1/2}(\boldsymbol{x}-\boldsymbol{y})\|_2. \tag{59}$$

*Therefore,* $\boldsymbol{x} \mapsto \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$ *is* $\lambda_{\min}(\boldsymbol{\Lambda})^{-1}$*-Lipschitz. Furthermore, there exists a weak derivative* $\frac{\partial}{\partial \boldsymbol{x}} \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$, *denoted by* $\nabla^2 \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$, *which is positive semi-definite and satisfies*

$$\mathbf{0}_{p\times p} \preceq \boldsymbol{\Lambda}^{1/2} \nabla^2 \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) \boldsymbol{\Lambda}^{1/2} \preceq \boldsymbol{I}_p.$$

PROOF. Note that (59) follows easily from the firm nonexpansiveness of the proximal operator $\mathbf{prox}_{\mathsf{Reg}}(\cdot; \boldsymbol{\Lambda})$ with respect to the inner product $\langle \boldsymbol{v}, \boldsymbol{u}\rangle_{\boldsymbol{\Lambda}^{-1/2}} := \boldsymbol{v}^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{u}$. For completeness, however, we prove (59) directly.

By the KKT conditions, $\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$ is also a minimizer of the following convex function $f_{\boldsymbol{x}} : \mathbb{R}^p \mapsto \mathbb{R}$:

$$f_{\boldsymbol{x}}(\boldsymbol{p}) := \frac{1}{2}(\boldsymbol{x}-\boldsymbol{p})^\top \boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{p}) - \frac{1}{2}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) - \boldsymbol{p})^\top \boldsymbol{\Lambda}^{-1}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) - \boldsymbol{p}) + \mathsf{Reg}(\boldsymbol{p}),$$

so that $f_{\boldsymbol{x}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})) \le f_{\boldsymbol{x}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda}))$. By symmetry, we also have $f_{\boldsymbol{y}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda})) \le f_{\boldsymbol{y}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}))$. Putting them together, we get

$$f_{\boldsymbol{x}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})) + f_{\boldsymbol{y}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda})) \le f_{\boldsymbol{x}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda})) + f_{\boldsymbol{y}}(\mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})).$$

Substituting the definitions of $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{y}}$ into the above display and rearranging it with $\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^{-1}(\boldsymbol{v} - \mathbf{prox}_{\mathsf{Reg}}(\boldsymbol{v}; \boldsymbol{\Lambda}))$, we obtain

$$\|\boldsymbol{\Lambda}^{1/2}(\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) - \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda}))\|_2^2 \le (\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) - \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{y}; \boldsymbol{\Lambda}))^\top (\boldsymbol{x}-\boldsymbol{y}).$$

Applying the Cauchy–Schwarz inequality to the RHS, we obtain (59).

By Rademacher's theorem, $\boldsymbol{x} \mapsto \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$ is differentiable almost everywhere, and for any differentiable points $\boldsymbol{x}$, the Jacobian is $\partial \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}) = \nabla^2 \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$, which is positive semi-definite since the mapping $\boldsymbol{x} \mapsto \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda})$ is convex. Let us fix $\boldsymbol{x}$ at which $\nabla \mathsf{env}$ is differentiable. For any $\boldsymbol{u} \in \mathbb{R}^p$ and any $t > 0$, using the established Lipschitz continuity (59), we have

$$\|\boldsymbol{\Lambda}^{1/2}(\nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x} + t\boldsymbol{\Lambda}^{1/2}\boldsymbol{u}; \boldsymbol{\Lambda}) - \nabla \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x}; \boldsymbol{\Lambda}))\|_2 \le \|\boldsymbol{\Lambda}^{-1/2} \cdot t\boldsymbol{\Lambda}^{1/2}\boldsymbol{u}\|_2 = t\|\boldsymbol{u}\|_2.$$

Dividing both sides by $t$, taking the limit $t \to 0$, and using the chain rule, we have

$$\|\mathbf{\Lambda}^{1/2}\nabla^2 \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x};\mathbf{\Lambda})\mathbf{\Lambda}^{1/2}\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_2$$

for all $\boldsymbol{u}\in\mathbb{R}^p$. This means $\mathbf{\Lambda}^{1/2}\nabla^2 \mathsf{env}_{\mathsf{Reg}}(\boldsymbol{x};\mathbf{\Lambda})\mathbf{\Lambda}^{1/2} \preceq \boldsymbol{I}_p$, completing the proof. $\square$

Let us apply Lemma 26 and Lemma 27 to $F_{\mathsf{Reg}}$. Note that $F_{\mathsf{Reg}}$ can be written as

$$F_{\mathsf{Reg}}(\eta_{{}_H}) = \frac{1}{\alpha\widetilde{\alpha}}\frac{1}{p}\mathbb{E}[\boldsymbol{f}(\boldsymbol{h})^\top \widetilde{\boldsymbol{f}}(\widetilde{\boldsymbol{h}})], \quad \widetilde{\boldsymbol{h}} = \eta_{{}_H}\boldsymbol{h} + \sqrt{1-\eta_{{}_H}^2}\boldsymbol{h}',$$

where $\boldsymbol{f},\widetilde{\boldsymbol{f}}:\mathbb{R}^p\to\mathbb{R}^p$ are defined as

$$\begin{aligned}
\boldsymbol{f}(\boldsymbol{x}) &:= \tfrac{\beta}{\nu}\boldsymbol{x} - \nu^{-1}\mathbf{\Sigma}^{-1/2}\nabla\mathsf{env}_{\mathsf{Reg}}(\boldsymbol{\theta} + \tfrac{\beta}{\nu}\mathbf{\Sigma}^{-1/2}\boldsymbol{x};(\nu\mathbf{\Sigma})^{-1})\\
\widetilde{\boldsymbol{f}}(\boldsymbol{x}) &:= \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\boldsymbol{x} - \widetilde{\nu}^{-1}\mathbf{\Sigma}^{-1/2}\nabla\mathsf{env}_{\widetilde{\mathsf{Reg}}}(\boldsymbol{\theta} + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\mathbf{\Sigma}^{-1/2}\boldsymbol{x};(\widetilde{\nu}\mathbf{\Sigma})^{-1}).
\end{aligned}$$

By Lemma 27, $\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ are Lipschitz and their weak derivatives satisfy

$$\begin{aligned}
\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}} &= \frac{\beta}{\nu}\Big(\boldsymbol{I}_p - (\nu\mathbf{\Sigma})^{-1/2}\nabla^2\mathsf{env}_{\mathsf{Reg}}(\boldsymbol{\theta}+\tfrac{\beta}{\nu}\mathbf{\Sigma}^{-1/2}\boldsymbol{x};(\nu\mathbf{\Sigma})^{-1})(\nu\mathbf{\Sigma})^{-1/2}\Big), \quad \mathbf{0}_{p\times p}\preceq\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}\preceq\frac{\beta}{\nu}\boldsymbol{I}_p\\
\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}} &= \frac{\widetilde{\beta}}{\widetilde{\nu}}\Big(\boldsymbol{I}_p - (\widetilde{\nu}\mathbf{\Sigma})^{-1/2}\nabla^2\mathsf{env}_{\widetilde{\mathsf{Reg}}}(\boldsymbol{\theta}+\tfrac{\widetilde{\beta}}{\widetilde{\nu}}\mathbf{\Sigma}^{-1/2}\boldsymbol{x};(\widetilde{\nu}\mathbf{\Sigma})^{-1})(\widetilde{\nu}\mathbf{\Sigma})^{-1/2}\Big), \quad \mathbf{0}_{p\times p}\preceq\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}\preceq\frac{\widetilde{\beta}}{\widetilde{\nu}}\boldsymbol{I}_p.
\end{aligned}$$

Therefore, using Lemma 26, noting that $\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}$ and $\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}$ are now symmetric, we know that $F_{\mathsf{Reg}}$ is differentiable with its derivative given by

$$F'_{\mathsf{Reg}}(\eta_{{}_H}) = \frac{1}{\alpha\widetilde{\alpha}p}\mathbb{E}\Big[\mathrm{tr}\Big\{\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}(\boldsymbol{h})\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}(\widetilde{\boldsymbol{h}})\Big\}\Big].$$

Now we use the following fact: $0\le\mathrm{tr}[\boldsymbol{A}\boldsymbol{B}]\le\mathrm{tr}[\boldsymbol{A}]$ for any positive semi-definite matrix $\boldsymbol{A},\boldsymbol{B}$ with $\mathbf{0}_{p\times p}\preceq\boldsymbol{A},\boldsymbol{B}\preceq\boldsymbol{I}_p$. This inequality immediately follows from the identity $\mathrm{tr}[\boldsymbol{A}\boldsymbol{B}] = \mathrm{tr}[\boldsymbol{A}\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}] = \mathrm{tr}[\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2}]$ and $\mathbf{0}_{p\times p}\preceq\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2}\preceq\boldsymbol{A}$. Combine this with $\mathbf{0}_{p\times p}\preceq\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}\preceq\frac{\widetilde{\beta}}{\widetilde{\nu}}\boldsymbol{I}_p$, $\mathbf{0}_{p\times p}\preceq\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}\preceq\frac{\beta}{\nu}\boldsymbol{I}_p$, we have

$$0\le F'_{\mathsf{Reg}}(\eta_{{}_H})\le\frac{1}{p\alpha\widetilde{\alpha}}\bigg(\frac{\widetilde{\beta}}{\widetilde{\nu}}\mathbb{E}\Big[\mathrm{tr}\Big\{\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}(\boldsymbol{h})\Big\}\Big]\wedge\frac{\beta}{\nu}\mathbb{E}\Big[\mathrm{tr}\Big\{\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}(\widetilde{\boldsymbol{h}})\Big\}\Big]\bigg).$$

Using Stein's lemma and the fact that $(\alpha,\beta,\nu,\kappa)$ satisfies (27c), recalling the definition of $\boldsymbol{f}(\boldsymbol{x})$, we have

$$\mathbb{E}\Big[\mathrm{tr}\Big\{\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{x}}(\boldsymbol{h})\Big\}\Big] = \mathbb{E}[\boldsymbol{h}^\top\boldsymbol{f}(\boldsymbol{h})] = \mathbb{E}\Big[\boldsymbol{h}^\top\Big\{\frac{\beta}{\nu}\boldsymbol{h} - \frac{\mathbf{\Sigma}^{-1/2}}{\nu}\nabla\mathsf{env}_{\mathsf{Reg}}\Big(\boldsymbol{\theta}+\frac{\beta}{\nu}\mathbf{\Sigma}^{-1/2}\boldsymbol{h};\frac{\mathbf{\Sigma}^{-1}}{\nu}\Big)\Big\}\Big] = p\kappa\beta.$$

By the same argument, we have $\mathbb{E}\Big[\mathrm{tr}\Big\{\frac{\partial\widetilde{\boldsymbol{f}}}{\partial\boldsymbol{x}}(\widetilde{\boldsymbol{h}})\Big\}\Big] = p\widetilde{\kappa}\widetilde{\beta}$. Substituting these equations to the previous display, we get $F'_{\mathsf{Reg}}(\eta_{{}_H})\le\frac{\beta\widetilde{\beta}}{\alpha\widetilde{\alpha}}(\frac{\kappa}{\widetilde{\nu}}\wedge\frac{\widetilde{\kappa}}{\nu})$.

Putting the above display together,

$$0\le F'_{\mathsf{loss}}(\eta_{{}_G})\le\frac{\alpha\widetilde{\alpha}}{\beta\widetilde{\beta}}\Big(\frac{c\widetilde{\nu}}{\kappa}\wedge\frac{\widetilde{c}\nu}{\widetilde{\kappa}}\Big), \quad 0\le F'_{\mathsf{Reg}}(\eta_{{}_H})\le\frac{\beta\widetilde{\beta}}{\alpha\widetilde{\alpha}}\Big(\frac{\kappa}{\widetilde{\nu}}\wedge\frac{\widetilde{\kappa}}{\nu}\Big).$$

Therefore, by the chain rule, we know that $\eta_{{}_H}\mapsto F_{\mathsf{loss}}\circ F_{\mathsf{Reg}}(\eta_{{}_H})$ and $\eta_{{}_G}\mapsto F_{\mathsf{Reg}}\circ F_{\mathsf{loss}}(\eta_{{}_G})$ are both $(c\wedge\widetilde{c})$-Lipschitz. This completes the proof.

**E.2. Heuristic proof of Conjecture 10.** We outline a heuristic proof strategy for Conjecture 10, leaving a complete rigorous proof to future work.

Consider the change of variables $\boldsymbol{b} \mapsto \boldsymbol{h} = p^{-1/2}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b}-\boldsymbol{\theta})$, and define $\boldsymbol{G} = \sqrt{p}\,\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n\times p}$, whose entries are i.i.d. $\mathcal{N}(0,1)$ (note that this $\boldsymbol{h}$ is different from the Gaussian vectors $\boldsymbol{h}, \widetilde{\boldsymbol{h}}$ appearing in System 5b). Then the regularized M-estimator $\widehat{\boldsymbol{\theta}}$ and its residual vector can be expressed as

$$\widehat{\boldsymbol{\theta}} = \sqrt{p}\,\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{h}} + \boldsymbol{\theta}, \quad \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{z} - \boldsymbol{G}\widehat{\boldsymbol{h}},$$

where $\widehat{\boldsymbol{h}}$ solves the M-estimation problem with a non-separable regularizer $\mathsf{R}(\cdot)$:

$$\widehat{\boldsymbol{h}} \in \operatorname*{argmin}_{\boldsymbol{h}\in\mathbb{R}^p} \sum_{i\in I} \mathsf{loss}(z_i - \boldsymbol{g}_i^\top \boldsymbol{h}) + \mathsf{R}(\boldsymbol{h}), \quad \mathsf{R}(\boldsymbol{h}) := \mathsf{Reg}(\sqrt{p}\,\boldsymbol{\Sigma}^{-1/2}\boldsymbol{h} + \boldsymbol{\theta}).$$

Assume for simplicity that Reg is $\tau$-strongly convex (otherwise apply the smoothing argument in Appendix B.2.1), so that $\mathsf{R}(\cdot)$ is $(\tau p)/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$-strongly convex. By Lemma 15, the mappings $\boldsymbol{G} \mapsto \widehat{\boldsymbol{h}}$ and $\boldsymbol{G} \mapsto \boldsymbol{\psi} := \sum_{i\in I} \mathsf{loss}'(z_i - \boldsymbol{g}_i^\top \widehat{\boldsymbol{h}})\boldsymbol{e}_i$ are differentiable, with derivatives

$$\forall i \in [n],\, j \in [p], \quad \frac{\partial \widehat{\boldsymbol{h}}}{\partial g_{ij}} = \boldsymbol{A}\Big(\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi} - \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{e}_i\boldsymbol{e}_j^\top\widehat{\boldsymbol{h}}\Big), \quad \frac{\partial \boldsymbol{\psi}}{\partial g_{ij}} = -\boldsymbol{D}\boldsymbol{G}\boldsymbol{A}\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{\psi} - \boldsymbol{V}\boldsymbol{e}_i\boldsymbol{e}_j^\top\widehat{\boldsymbol{h}}.$$

Following the argument of Appendix B.2.4, the Gaussian Poincaré inequality yields

$$\frac{\widehat{\boldsymbol{h}}^\top\widetilde{\boldsymbol{h}}}{\|\widehat{\boldsymbol{h}}\|\|\widetilde{\boldsymbol{h}}\|} = \widehat{\eta}_{G} + o_{\mathbb{P}}(1), \quad \frac{\widehat{\boldsymbol{\psi}}^\top\widetilde{\boldsymbol{\psi}}}{\|\boldsymbol{\psi}\|\|\widetilde{\boldsymbol{\psi}}\|} = \widehat{\eta}_{H} + o_{\mathbb{P}}(1),$$

for random variables $\widehat{\eta}_{G}, \widehat{\eta}_{H} \in [-1,1]$ independent of $\boldsymbol{G}$.

Using the contraction property in Theorem 9, it remains to show

$$\widehat{\eta}_{H} = F_{\mathsf{loss}}(\widehat{\eta}_{G}) + o_{\mathbb{P}}(1), \quad \widehat{\eta}_{G} = F_{\mathsf{Reg}}(\widehat{\eta}_{H}) + o_{\mathbb{P}}(1).$$

The first relation $\widehat{\eta}_{H} = F_{\mathsf{loss}}(\widehat{\eta}_{G}) + o_{\mathbb{P}}(1)$ follows from the same reasoning as in Appendix B.2.4, as $F_{\mathsf{loss}}$ is not altered from the original isotropic and random signal setting.

The main challenge lies in establishing the second one, $\widehat{\eta}_{G} = F_{\mathsf{Reg}}(\widehat{\eta}_{H}) + o_{\mathbb{P}}(1)$. By carefully tracing the argument in Appendix B.2.4, applying Lemma 18 with $\boldsymbol{F} = \Big[\frac{\boldsymbol{\psi}}{\sqrt{p\beta}}\ \frac{\widetilde{\boldsymbol{\psi}}}{\sqrt{p\widetilde{\beta}}}\Big] \in \mathbb{R}^{n\times 2}$ and $\boldsymbol{z} = \boldsymbol{G}\boldsymbol{e}_j \in \mathbb{R}^n$, combined with the concentration results in (28), one can construct a random vector $\boldsymbol{w}_j \in \mathbb{R}^2$ for each $j \in [p]$ such that

$$\boldsymbol{w}_j \mid (\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{G}^{-j}, I, \widetilde{I}) \overset{\mathrm{d}}{=} \mathcal{N}(\boldsymbol{0}_2, \boldsymbol{I}_2), \text{ here } \boldsymbol{G}^{-j} \text{ is } \boldsymbol{G} \text{ with } j\text{th column removed} \tag{60}$$

and obtain

$$\widehat{\eta}_{G} = \mathcal{F}_{\mathsf{Reg}}(\widehat{\eta}_{H}) + o_{\mathbb{P}}(1),$$

where the map $\mathcal{F}_{\mathsf{Reg}} : [-1,1] \to \mathbb{R}$ is given by

$$\mathcal{F}_{\mathsf{Reg}}(\eta) := \frac{1}{p\alpha\widetilde{\alpha}}\mathbb{E}\Bigg[\Big(\frac{\boldsymbol{\Sigma}^{-1/2}}{\nu}\nabla\mathsf{env}_{\mathsf{Reg}}\big(\boldsymbol{\theta} + \tfrac{\beta}{\nu}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}_\eta; \tfrac{\boldsymbol{\Sigma}^{-1}}{\nu}\big) - \tfrac{\beta}{\nu}\boldsymbol{w}_\eta\Big)^\top \Big(\frac{\boldsymbol{\Sigma}^{-1/2}}{\widetilde{\nu}}\nabla\mathsf{env}_{\widetilde{\mathsf{Reg}}}\big(\boldsymbol{\theta} + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{w}}_\eta; \tfrac{\boldsymbol{\Sigma}^{-1}}{\widetilde{\nu}}\big) - \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\widetilde{\boldsymbol{w}}_\eta\Big)\Bigg],$$

with $\boldsymbol{w}_\eta = (w_{\eta,j})_{j\in[p]}$ and $\widetilde{\boldsymbol{w}}_\eta = (\widetilde{w}_{\eta,j})_{j\in[p]}$ defined by

$$\forall j \in [p], \quad \begin{pmatrix} w_{\eta,j} \\ \widetilde{w}_{\eta,j} \end{pmatrix} = \begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix}^{1/2} \boldsymbol{w}_j.$$

Therefore, it suffices to prove $\mathcal{F}_{\mathsf{Reg}} = F_{\mathsf{Reg}}$, but at this point, this cannot be concluded directly. Indeed, by (60), we have $(w_{\eta,j}, \widetilde{w}_{\eta,j}) \sim \mathcal{N}\left(\mathbf{0}_2, \begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix}\right)$ marginally, but the vectors $(w_{\eta,j}, \widetilde{w}_{\eta,j})_{j\in[p]}$ are not necessarily independent across $j$ since (60) is not strong enough to claim the independence of $(\boldsymbol{w}_j)_{j\in[p]}$. Now recalling the definition of $F_{\mathsf{Reg}}$ in System 5b, the bivariate Gaussian vectors $(\boldsymbol{h}, \widetilde{\boldsymbol{h}}) = (h_j, \widetilde{h}_j)_{j\in[p]}$ in $F_{\mathsf{Reg}}$ are explicitly assumed to have the **i.i.d.** marginals $\mathcal{N}\left(\mathbf{0}_2, \begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix}\right)$. This distinction prevents us from directly concluding $\mathcal{F}_{\mathsf{Reg}} = F_{\mathsf{Reg}}$.

Nonetheless, we can conclude that $\mathcal{F}_{\mathsf{Reg}} = F_{\mathsf{Reg}}$ when $\boldsymbol{\Sigma}$ is diagonal (not necessarily the identity) and $(\mathsf{Reg}, \widetilde{\mathsf{Reg}})$ are separable such that $\mathsf{Reg}(\boldsymbol{x}) = \sum_{j\in[p]} \mathsf{reg}_j(x_j)$, $\widetilde{\mathsf{Reg}}(\boldsymbol{x}) = \sum_{j\in[p]} \widetilde{\mathsf{reg}}_j(x_j)$. Indeed, since the Moreau envelope can be expressed as a separable form in this case, we can write $\mathcal{F}_{\mathsf{Reg}}$ and $F_{\mathsf{Reg}}$ as

$$\mathcal{F}_{\mathsf{Reg}}(\eta) = \frac{1}{p\alpha\widetilde{\alpha}} \sum_{j=1}^{p} \mathbb{E}\Big[\big(\frac{\Sigma_{jj}^{-1/2}}{\nu} \mathsf{env}'_{\mathsf{reg}_j}(\theta_j + \tfrac{\beta}{\nu}\Sigma_{jj}^{-1/2} w_{\eta,j}; \tfrac{\Sigma_{jj}^{-1}}{\nu}) - \tfrac{\beta}{\nu} w_{\eta,j}\big)\big(\frac{\Sigma_{jj}^{-1/2}}{\widetilde{\nu}} \mathsf{env}'_{\widetilde{\mathsf{reg}}_j}(\theta_j + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\Sigma_{jj}^{-1/2} \widetilde{w}_{\eta,j}; \tfrac{\Sigma_{jj}^{-1}}{\widetilde{\nu}}) - \tfrac{\widetilde{\beta}}{\widetilde{\nu}} \widetilde{w}_{\eta,j}\big)\Big]$$

$$F_{\mathsf{Reg}}(\eta) = \frac{1}{p\alpha\widetilde{\alpha}} \sum_{j=1}^{p} \mathbb{E}\Big[\big(\frac{\Sigma_{jj}^{-1/2}}{\nu} \mathsf{env}'_{\mathsf{reg}_j}(\theta_j + \tfrac{\beta}{\nu}\Sigma_{jj}^{-1/2} h_j; \tfrac{\Sigma_{jj}^{-1}}{\nu}) - \tfrac{\beta}{\nu} h_j\big)\big(\frac{\Sigma_{jj}^{-1/2}}{\widetilde{\nu}} \mathsf{env}'_{\widetilde{\mathsf{reg}}_j}(\theta_j + \tfrac{\widetilde{\beta}}{\widetilde{\nu}}\Sigma_{jj}^{-1/2} \widetilde{h}_j; \tfrac{\Sigma_{jj}^{-1}}{\widetilde{\nu}}) - \tfrac{\widetilde{\beta}}{\widetilde{\nu}} \widetilde{h}_j\big)\Big],$$

for all $\eta \in [-1, 1]$, which implies $\mathcal{F}_{\mathsf{Reg}} = F_{\mathsf{Reg}}$ thanks to $(w_{\eta,j}, \widetilde{w}_{\eta,j}) =^d (h_j, \widetilde{h}_j)$ marginally for $j \in [p]$. In this case, the independence of $(w_{\eta,j}, \widetilde{w}_{\eta,j})$ across $j$ is not required since the expectation $\mathbb{E}$ is taken inside $\sum_j$.

In summary, the key technical obstacle is to establish independence of $(\boldsymbol{w}_j)_{j\in[p]}$ in (60), which lies beyond the reach of Lemma 18. Overcoming this requires new tools, and we leave this gap for future research.

## APPENDIX F: ADDITIONAL NUMERICAL SIMULATIONS

**F.1. Minimum $\ell_1$-norm interpolator.** See Figures 13 and 14.

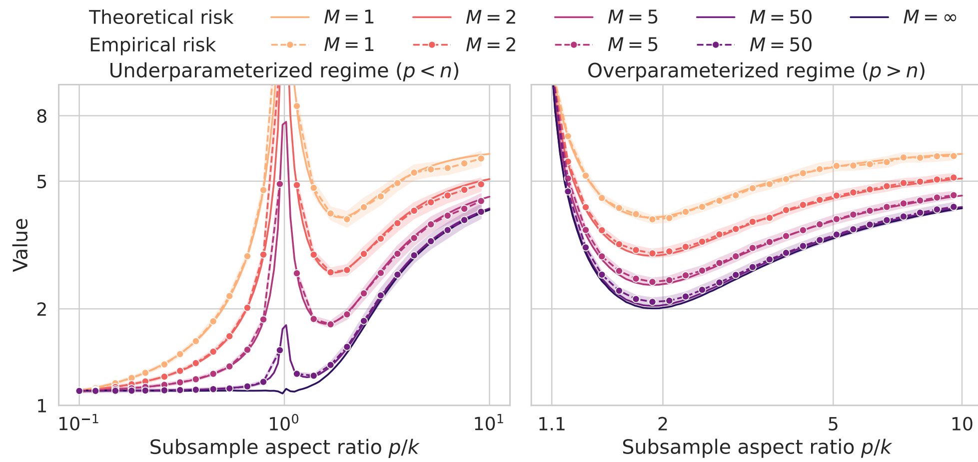


Fig 13: The prediction risk for lasso ensemble at different subsample aspect ratios $p/k$ with regularization parameter $\lambda = 0.001$ and varying ensemble size $M$. The solid lines represent the theoretical risks, the dashed lines represent the empirical risks averaged over 50 simulations, and the shaded regions represent the empirical standard errors. The data model is given by (61) with signal strength $\rho = 2$, noise level $\sigma = 1$, and support proportion $s = 0.1$. *Left*: underparameterized regime when $p/n = 0.1$ and $n = 2000$. *Right*: overparameterized regime when $p/n = 1.1$ and $n = 500$.

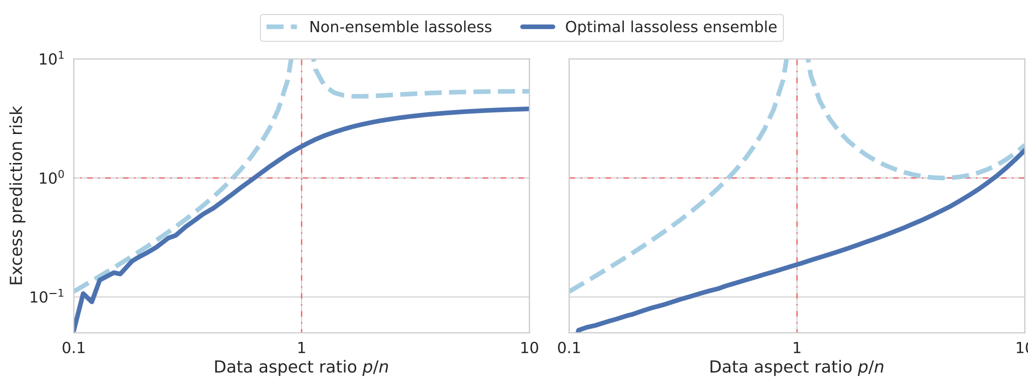


Fig 14: **Optimal subsample risk of the lassoless ensemble is monotonic in the data aspect ratio.** Excess risk of the lasso and optimal lasso ensemble, at different data aspect ratios $p/n$ ranging from 0.1 to 10. The data model is given by (61) with signal strength $\rho = 2$, noise level $\sigma = 1$, data aspect ratio $p/n = 0.1$, feature size $p = 500$, and varying support proportion $s$. *Left*: dense regime with $s = 0.9$. *Right*: sparse regime with $s = 0.01$.

**F.2. Lasso under varying covariate distributions.** See Figure 15.

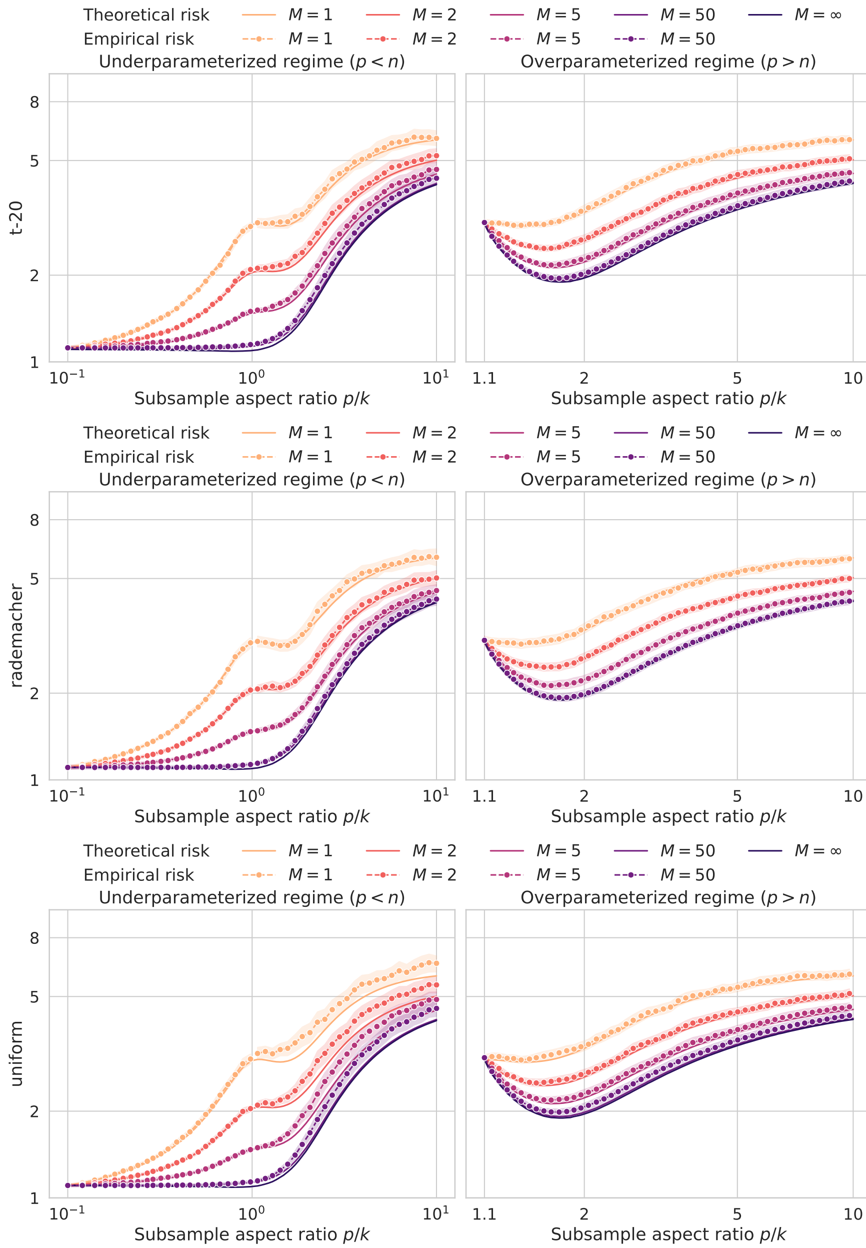


Fig 15: The prediction risk for lasso ensemble at different subsample aspect ratios $p/k$ with regularization parameter $\lambda = 0.1$ and varying ensemble size $M$. The covariates $X_{ij}$'s are i.i.d. and follow (a) $t_{20}$ (b) Rademacher distribution, and (c) uniform distribution in $[-1/1]$, with a proper normalization such that the mean is zero and variance is one. The solid lines represent the theoretical risks, the dashed lines represent the empirical risks under Gaussian covariate averaged over $50$ simulations, and the shaded regions represent the empirical standard errors. The data model is given by (61) with covariate distribution replaced by the corresponding distribution, and with signal strength $\rho = 2$, noise level $\sigma = 1$, and support proportion $s = 0.1$. *Left*: underparameterized regime when $p/n = 0.1$ and $n = 2000$. *Right*: overparameterized regime when $p/n = 1.1$ and $n = 500$.

### F.3. Optimal subagging versus optimal regularization.

F.3.1. *Squared loss, lasso regularizer.* See Figure 16.

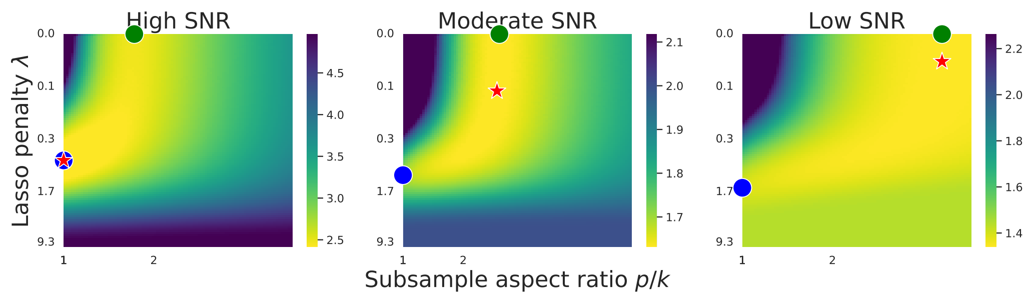


Fig 16: Heatmaps of theoretical prediction risk in $\lambda$ and subsample aspect ratio $p/k = (p/n) \cdot c^{-1}$ of full lasso ensemble in the overparameterized regime ($p/n = 1.1$) and sparse regime ($s = 0.2$). The data model is given by (61) with support proportion $s = 0.2$ and noise level $\sigma = 1$ at different signal levels $\rho$. Blue dots (●): the optimal lasso. Green dots (●): the optimally subsampled lassoless regression. Red stars ($\star$): the optimal lasso ensemble. *Left*: high SNR $\rho = 2$. The optimal lasso (●) is better than the optimally subsampled lassoless (●). *Middle*: moderate SNR $\rho = 1$. The optimal lasso ensemble ($\star$) is better than both the optimal lasso (●) and the optimally subsampled lassoless(●). *Right*: low SNR $\rho = 0.67$. The optimally subsampled lassoless (●) is better than the optimal lasso (●).

F.3.2. *Huber loss, unregularized.* See Figure 17.

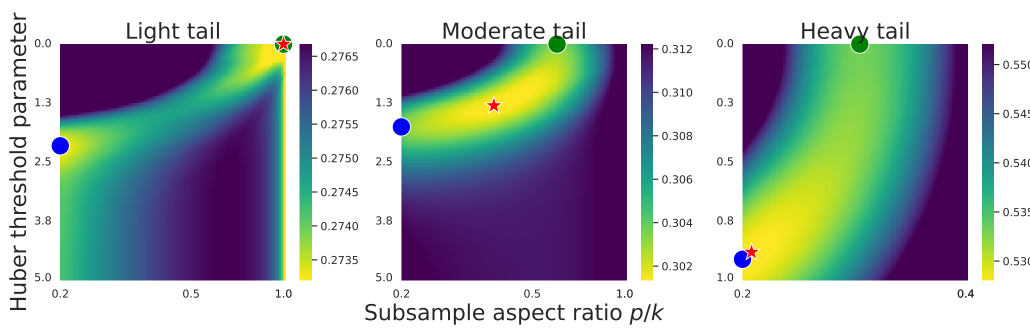


Fig 17: Heatmaps of theoretical prediction risk in Huber threshold parameter and subsample aspect ratio $p/k \in [p/n, 1)$ of full Huber ensemble in the underparameterized regime ($p/n = 0.2$). Blue dots (●): the optimal Huber regression. Green dots (●): the optimally subsampled LAD regression (Huber regression with threshold parameter $\to 0^+$). Red stars ($\star$): the optimal Huber ensemble. *Left*: noise follows Student's $t$ distribution $t_{20}$. The optimally subsampled LAD (●) is better than the optimal Huber (●). *Middle*: noise follows Student's $t$ distribution $t_{10}$ The optimal subsample Huber ($\star$) is better than the both optimal subsample LAD (●) and the optimal Huber (●). *Right*: noise follows Student's $t$ distribution $t_2$. The optimal Huber (●) is better than the optimally subsampled LAD (●).

F.3.3. *$\ell_1$-regularized Huber regression: varying threshold parameter.* See Figure 18.

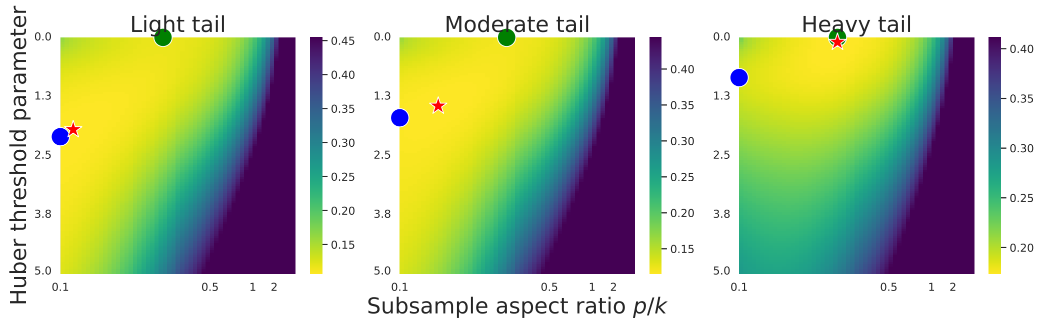


Fig 18: Heatmaps of theoretical prediction risk in Huber threshold parameter and subsample aspect ratio $p/k = (p/n) \cdot c^{-1}$ of full $\ell_1$-regularized Huber ensemble with lasso regularization level $0.5$ in the underparameterized regime $(p/n = 0.1)$. Blue dots (•): the optimal $\ell_1$-regularized Huber regression. Green dots (•): the optimally subsampled $\ell_1$-regularized LAD regression ($\ell_1$-regularized Huber regression with huber threshold parameter $\to 0^+$). Red stars ($\star$): the optimal $\ell_1$-regularized Huber ensemble. *Left*: noise follows Student's $t$ distribution $t_{20}$. The optimal $\ell_1$-regularized Huber (•) is better than the optimally subsampled $\ell_1$-regularized LAD (•). *Middle*: noise follows Student's $t$ distribution $t_{10}$. The optimal $\ell_1$-regularized Huber ensemble ($\star$) attains nearly the same risk as both the optimal $\ell_1$-regularized Huber (•) and the optimally subsampled $\ell_1$-regularized LAD (•). *Right*: noise follows Student's $t$ distribution $t_2$. The optimally subsampled $\ell_1$-regularized LAD (•) is better than the optimal $\ell_1$-regularized Huber (•).

F.3.4. *$\ell_1$-regularized Huber regression: varying regularization level.* See Figure 19.

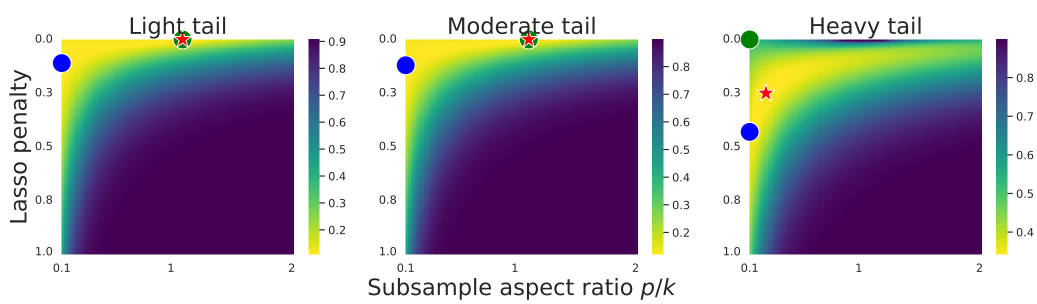


Fig 19: Heatmaps of theoretical prediction risk in lasso regularization level and subsample aspect ratio $p/k = (p/n) \cdot c^{-1}$ of full $\ell_1$-regularized Huber ensemble with Huber threshold parameter $10$ in the underparameterized regime $(p/n = 0.1)$. Blue dots (•): the optimal $\ell_1$-regularized Huber regression. Green dots (•): the optimally subsampled lassoless Huber regression ($\ell_1$-norm interpolator when $p/k \geq 1$ and unregularized Huber regression when $p/k < 1$). Red stars ($\star$): the optimal $\ell_1$-regularized Huber ensemble. *Left*: noise follows Student's $t$ distribution $t_{20}$. The optimally subsampled lassoless Huber (•) is better than the optimal $\ell_1$-regularized Huber (•). *Middle*: noise follows Student's $t$ distribution $t_{10}$. The optimally subsampled lassoless Huber (•) is better than the optimal $\ell_1$-regularized Huber (•). *Right*: noise follows Student's $t$ distribution $t_2$ The optimal $\ell_1$-regularized Huber ensemble ($\star$) is better than both the optimally subsampled lassoless Huber (•) and the optimal $\ell_1$-regularized Huber (•)

**F.4. Details of numerical experiments.**

F.4.1. *Data model for lasso experiments.* We consider a linear model whose signal variables are generated from two-point distribution:

(61)
$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + z, \qquad \boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, p^{-1}\boldsymbol{I}_p), \quad z \sim \mathcal{N}(0, \sigma^2), \quad \boldsymbol{\theta} \overset{i.i.d.}{=\sim} s\mathcal{P}_{\rho/\sqrt{s}} + (1-s)\mathcal{P}_0,$$

where $\mathcal{P}_c$ denotes the Dirac measure at point $c \in \mathbb{R}$, and $\rho > 0$ is some given quantity that determines the signal magnitude. Here, $s \in (0,1)$ is the support proportion. The signal-to-noise-ratio (SNR) under the above model obeys

$$\mathrm{SNR} = \frac{\mathbb{E}[(\boldsymbol{x}^\top \boldsymbol{\theta})^2]}{\sigma^2} = \frac{s \cdot (\rho^2/s)}{\sigma^2} = \frac{\rho^2}{\sigma^2}.$$

F.4.2. *Data model for $\ell_1$-regularized Huber experiments.* We consider the $\ell_1$-regularized Huber regression

$$\widehat{\boldsymbol{\theta}}_I \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i \in I} \mathsf{Huber}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}; \rho) + \lambda \|\boldsymbol{\theta}\|_1$$

where $\lambda \geq 0$ is a regularization parameter and Huber$(\cdot; \mu)$ is the Huber loss with threshold parameter $\mu > 0$, which is defined as follows [62]:

(62) For all $x \in \mathbb{R}$,
$$\mathsf{Huber}(x;\mu) := \mathsf{env}_{|\cdot|}(x;\mu) = \begin{cases} \frac{1}{2\mu}x^2 & \text{if } |x| \leq \mu \\ |x| - \frac{1}{2}\mu & \text{if } |x| > \mu. \end{cases}$$

Observe that $\mathsf{Huber}(\cdot;\mu)$ behaves like the squared loss for large $\mu$ and like the absolute loss for small $\mu$. More precisely, for all $x \in \mathbb{R}$ it holds that

$$\lim_{\mu \to 0^+} \mathsf{Huber}(x;\mu) = |x| \quad \text{and} \quad \lim_{\mu \to +\infty} \mu \cdot \mathsf{Huber}(x;\mu) = x^2/2.$$

We consider the linear model $y = \boldsymbol{x}^\top \boldsymbol{\theta} + z$ where the marginal distribution of the signal is set to the mixture $\epsilon \mathcal{N}(0,1) + (1-\epsilon)\mathcal{P}_0$ (recall that $\mathcal{P}_0$ denotes the Dirac measure at 0) with support proportion $\epsilon = 0.1$, while the noise $z$ follows Student's t-distribution $t_d$ for some degree of freedom $d \geq 2$, which will be specified for each numerical simulation.

## REFERENCES


[1] ADLAM, B. and PENNINGTON, J. (2020). Understanding Double Descent Requires a Fine-Grained Bias-Variance Decomposition. In *Advances in Neural Information Processing Systems*.

[2] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* 242–252.

[3] ANDO, R. and KOMAKI, F. (2023). On High-Dimensional Asymptotic Properties of Model Averaging Estimators. *arXiv preprint arXiv:2308.09476*.

[4] BARTLETT, P. L., MONTANARI, A. and RAKHLIN, A. (2021). Deep learning: a statistical viewpoint. *Acta Numerica* **30** 87–201.

[5] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory* **57** 764–785.

[6] BAYATI, M. and MONTANARI, A. (2011). The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory* **58** 1997–2017.

[7] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal $M$-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences* **110** 14563–14568.

[8] BELKIN, M. (2021). Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica* **30** 203–248.

[9] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*.

[10] BELLEC, P. C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46** 30–59.
[11] BELLEC, P. C. (2023). Out-of-sample error estimation for M-estimators with convex penalty. *Information and Inference: A Journal of the IMA* **12** 2782–2817.
[12] BELLEC, P. C. (2025). Observable adjustments in single-index models for regularized M-estimators with bounded p/n. *The Annals of Statistics* **53** 531–560.
[13] BELLEC, P. C., DU, J.-H., KORIYAMA, T., PATIL, P. and TAN, K. (2024). Corrected generalized cross-validation for finite ensembles of penalized estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.
[14] BELLEC, P. C. and KORIYAMA, T. (2023). Existence of solutions to the nonlinear equations characterizing the precise error of M-estimators. *arXiv preprint arXiv:2312.13254*.
[15] BELLEC, P. C. and KORIYAMA, T. (2024). Asymptotics of resampling without replacement in robust and logistic regression. *arXiv preprint arXiv:2404.02070*.
[16] BELLEC, P. C. and KORIYAMA, T. (2025). Error estimation and adaptive tuning for unregularized robust M-estimator. *Journal of Machine Learning Research* **26** 1–40.
[17] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*.
[18] BERTSEKAS, D. (2016). *Nonlinear Programming*, Third edition ed. Athena Scientific.
[19] BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.
[20] BREIMAN, L. (2001). Using iterated bagging to debias regressions. *Machine Learning* **45** 261–277.
[21] BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *The Annals of Statistics* **30** 927–961.
[22] BUJA, A. and STUETZLE, W. (2006). Observations on bagging. *Statistica Sinica* 323–351.
[23] CELENTANO, M. and MONTANARI, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics* **50** 170–196.
[24] CELENTANO, M. and MONTANARI, A. (2024). Correlation adjusted debiased Lasso: debiasing the Lasso with inaccurate covariate model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.
[25] CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics* **51** 2194–2220.
[26] CHEN, X., ZENG, Y., YANG, S. and SUN, Q. (2023). Sketched Ridgeless Linear Regression: The Role of Downsampling. In *International Conference on Machine Learning*.
[27] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for overparameterized models using optimal transport. *Advances in Neural Information Processing Systems*.
[28] CLARTÉ, L., VANDENBROUCQUE, A., DALLE, G., LOUREIRO, B., KRZAKALA, F. and ZDEBOROVÁ, L. (2024). Analysis of Bootstrap and Subsampling in High-dimensional Regularized Regression. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence* **244** 787–819. PMLR.
[29] DAI, D., RIGOLLET, P., L., X. and T., Z. (2014). Aggregation of Affine Estimators. *Electon. J. Stat.* **8** 302-327.
[30] DAI, D., RIGOLLET, P. and ZHANG, T. (2012). Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics* **40** 1878–1905.
[31] DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning* **72** 39–61.
[32] DALALYAN, A. S., GRAPPIN, E. and PARIS, Q. (2018). On the exponentially weighted aggregate with the Laplace prior. *Ann. Statist.*
[33] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics* **40** 2327–2355.
[34] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces* **1**.
[35] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA* **11** 435–495.
[36] DOBRIBAN, E. and SHENG, Y. (2020). WONDER: Weighted One-shot Distributed Ridge Regression in High Dimensions. *Journal of Machine Learning Research* **21** 1–52.
[37] DOBRIBAN, E. and SHENG, Y. (2021). Distributed linear regression by averaging. *The Annals of Statistics* **49** 918–943.
[38] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46** 247–279.
[39] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **106** 18914–18919.
[40] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory* **57** 6920–6941.

[41] DONOHO, D. L. and MONTANARI, A. (2016). High dimensional robust $M$-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields* **166** 935–969.
[42] DU, J.-H. and PATIL, P. (2024). Implicit Regularization Paths of Weighted Neural Representations. In *Advances in Neural Information Processing Systems* **37** 30261–30299.
[43] DU, J.-H. and PATIL, P. (2024). Python package sklearn_ensemble_cv v0.2.3. PyPI.
[44] DU, J.-H., PATIL, P. and KUCHIBHOTLA, A. K. (2023). Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation. In *International Conference on Machine Learning*.
[45] DU, J.-H., PATIL, P., ROEDER, K. and KUCHIBHOTLA, A. K. (2024). Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*.
[46] DU, S., ZHAI, X., POCZOS, B. and SINGH, A. (2018). Gradient descent provably optimizes overparameterized neural networks. *arXiv preprint arXiv:1810.02054*.
[47] D'ASCOLI, S., REFINETTI, M., BIROLI, G. and KRZAKALA, F. (2020). Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*.
[48] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.
[49] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* **110** 14557–14562.
[50] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.
[51] FRIEDMAN, J. H. and HALL, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference* **137** 669–683.
[52] FU, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7** 397–416.
[53] GERBELOT, C. and BERTHIER, R. (2023). Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA* **12** 2562–2628.
[54] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*.
[55] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*.
[56] HALL, P. and SAMWORTH, R. J. (2005). Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 363–379.
[57] HAN, Q. and SHEN, Y. (2023). Universality of regularized regression estimators in high dimensions. *The Annals of Statistics* **51** 1799–1823.
[58] HAN, Q. and XU, X. (2023). The distribution of Ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*.
[59] HANSEN, L. and SALAMON, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** 993–1001.
[60] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics* **50** 949–986.
[61] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Second edition.
[62] HUBER, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35** 73–101.
[63] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA* **2** 115–144.
[64] KAROUI, N. E. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields* **170** 95–175.
[65] KROGH, A. and SOLLICH, P. (1997). Statistical mechanics of ensemble learning. *Physical Review E* **55** 811.
[66] LECUÉ, G. and RIGOLLET, P. (2014). Optimal learning with Q-aggregation. *Ann. Statist.* **42** 211-224.
[67] LEE, J., XIAO, L., SCHOENHOLZ, S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*.
[68] LEJEUNE, D., JAVADI, H. and BARANIUK, R. (2020). The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*.
[69] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on* **52** 3396–3410.

[70] LI, Y. and WEI, Y. (2021). Minimum $\ell_1$-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

[71] LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-$\ell_1$-norm interpolated classifiers. *The Annals of Statistics* **50** 1669 – 1695.

[72] LOUREIRO, B., GERBELOT, C., CUI, H., GOLDT, S., KRZAKALA, F., MEZARD, M. and ZDEBOROVÁ, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems* **34** 18137–18151.

[73] LOUREIRO, B., GERBELOT, C., REFINETTI, M., SICURO, G. and KRZAKALA, F. (2022). Fluctuations, Bias, Variance and Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension. In *International Conference on Machine Learning*.

[74] MAI, X., LIAO, Z. and COUILLET, R. (2019). A large scale analysis of logistic regression: Asymptotic performance and new insights. In *International Conference on Acoustics, Speech and Signal Processing*.

[75] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics* **49** 2313–2335.

[76] MONTANARI, A., RUAN, F., SOHN, Y. and YAN, J. (2025). The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *The Annals of Statistics* **53** 822 – 853.

[77] MÜCKE, N., REISS, E., RUNGENHAGEN, J. and KLEIN, M. (2022). Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*.

[78] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Foundations of Computational Mathematics* **16** 965–1029.

[79] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In *Allerton Conference on Communication, Control, and Computing*.

[80] PARIKH, N., BOYD, S. et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization* **1** 127–239.

[81] PATIL, P. and DU, J.-H. (2023). Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems*.

[82] PATIL, P., DU, J.-H. and KUCHIBHOTLA, A. K. (2023). Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research* **24** 1–113.

[83] PATIL, P., DU, J.-H. and TIBSHIRANI, R. J. (2024). Revisiting optimism and model complexity in the wake of overparameterized machine learning. *arXiv preprint arXiv:2410.01259*.

[84] PATIL, P., KUCHIBHOTLA, A. K., WEI, Y. and RINALDO, A. (2022). Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*.

[85] PATIL, P. and LEJEUNE, D. (2024). Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*.

[86] PATIL, P., WEI, Y., RINALDO, A. and TIBSHIRANI, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*.

[87] PERRONE, M. (1993). Putting it all together: Methods for combining neural networks. In *Advances in Neural Information Processing Systems*.

[88] PESCE, L., KRZAKALA, F., LOUREIRO, B. and STEPHAN, L. (2023). Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*.

[89] RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639-665.

[90] SALEHI, F., ABBASI, E. and HASSIBI, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*.

[91] SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics* **40** 2733–2763.

[92] SOLLICH, P. and KROGH, A. (1995). Learning with ensembles: How overfitting can be useful. In *Advances in Neural Information Processing Systems*.

[93] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151.

[94] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116** 14516–14525.

[95] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory* **64** 5592–5628.

[96] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*.

[97] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7** 1456–1490.

[98] VAN DER VAART, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press.

[99] WANG, S., WENG, H. and MALEKI, A. (2020). Which bridge estimator is the best for variable selection? *The Annals of Statistics* **48** 2791 – 2823.

[100] WENG, H., MALEKI, A. and ZHENG, L. (2018). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics* **46** 3099 – 3129.