# REGULARIZED ADAPTIVE HUBER MATRIX REGRESSION AND DISTRIBUTED LEARNING

Yue Wang[1], Wenqi Lu[2], Lei Wang[2], Zhongyi Zhu[3], Hongmei Lin[4] and Heng Lian[1,5]

*[1]City University of Hong Kong, [2]Nankai University, [3]Fudan University*

*[4]Shanghai University of International Business and Economics*

*and [5]City University of Hong Kong Shenzhen Research Institute*

**Supplementary Material**

## S1. Proofs of main results

Let $\mathbb{S}^{p-1}$ denote the $p$-dimensional unit sphere, $P$ and $P_n$ denote the expectation under the population measure and the empirical counterpart, respectively. Let $Pf = \mathbb{E}f(x)$ and $P_n f = (1/n)\sum_{i=1}^{n} f(x_i)$.

### S1.1 Derivation of Algorithm 1

**Update for r:** Let $\mathbf{h} = \mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta}^k - \mathbf{u}^k/\rho$. According to the Huber loss function, we split into two cases.

Case 1: $|r_i| \leq \tau$. The original optimization transforms to

$$\min_{r_i} \frac{r_i^2}{2n} + \frac{\rho}{2}(r_i - h_i)^2.$$

Therefore, we obtain $\widehat{r}_i = n\rho h_i/(1+n\rho)$. Substituting this into $|r_i| \leq \tau$ yields $|h_i| \leq \tau(1+n\rho)/(n\rho)$.

Case 2: $|r_i| > \tau$. The original optimization transforms to

$$\min_{r_i} \frac{1}{2}(r_i - h_i)^2 + \frac{\tau}{n\rho}|r_i|.$$

Therefore, we have $\widehat{r}_i = \mathrm{Soft}[h_i, \tau/(n\rho)]$ with $\mathrm{Soft}[a,\kappa] = (a-\kappa)_+ - (-a-\kappa)_+$ representing the soft thresholding operator. That is, $\widehat{r}_i = h_i - \tau/(n\rho)$ if $h_i > \tau(1+n\rho)/(n\rho)$ and $\widehat{r}_i = h_i + \tau/(n\rho)$ if $h_i < -\tau(1+n\rho)/(n\rho)$.

**Update for B/b:** Let $\sum_{j=1}^{\min\{p,q\}} \omega_j \mathbf{a}_j \mathbf{c}_j^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\mathbf{\Theta}^k - \mathbf{V}^k/\rho$. Then this problem can be solved by singular value thresholding given by

$$\mathbf{B}^{k+1} = \sum_{j=1}^{\min\{p,q\}} \max\left(\omega_j - \lambda/\rho, 0\right) \mathbf{a}_j \mathbf{c}_j^{\mathrm{T}}.$$

**Update for $\boldsymbol{\theta}$:** Setting the derivative equals to zero implies

$$\boldsymbol{\theta}^{k+1} = (\bar{\mathbf{X}}^\top \bar{\mathbf{X}} + \mathbf{I}_{pq})^{-1}[\bar{\mathbf{X}}^\top(-\mathbf{r}^{k+1} + \mathbf{y} - \mathbf{u}^k/\rho)].$$

## S1.2 Proof of Theorem 1

By the first order optimality condition of (2.1), there exists $\bar{\mathbf{\Theta}} \in \partial \|\widehat{\mathbf{\Theta}}\|_*$ such that $\nabla L(\widehat{\mathbf{\Theta}}) + \lambda \bar{\mathbf{\Theta}} = \mathbf{0}$. Then we have

$$\langle \nabla L(\widehat{\mathbf{\Theta}}) - \nabla L(\mathbf{\Theta}_0), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \rangle = \langle \lambda \bar{\mathbf{\Theta}} + \nabla L(\mathbf{\Theta}_0), \mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}} \rangle.$$

Using Hölder's inequality, we have

$$\langle \lambda \bar{\mathbf{\Theta}}, \mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}} \rangle \leq \lambda \|\bar{\mathbf{\Theta}}\|_{op} \|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}\|_* \leq 8\lambda \|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r\|_*,$$

where the last inequality uses the fact that $\|\bar{\mathbf{\Theta}}\|_{op} \leq 2$ and Lemma 1 which states that $\|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}\|_* \leq \|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_{r^c}\|_* + \|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r\|_* \leq 4\|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r\|_*$. Similarly, we have

$$\langle \nabla L(\mathbf{\Theta}_0), \mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}} \rangle \quad \leq \quad \|\nabla L(\mathbf{\Theta}_0)\|_{op} \|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}\|_* \leq 2\lambda \|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r\|_*,$$

where the last inequality uses $2\|\nabla L(\mathbf{\Theta}_0)\|_{op} \leq \lambda$ by Lemma 6. It follows from Lemma 1 of Negahban and Wainwright (2011a) which states that $\mathrm{rank}((\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r) \leq 2r$ and Cauchy-Schwarz inequality that

$$\|(\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}})_r\|_* \leq \sqrt{2r} \|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}\|_F.$$

Therefore, combining the above inequalities yields

$$\langle \nabla L(\widehat{\mathbf{\Theta}}) - \nabla L(\mathbf{\Theta}_0), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \rangle \leq C\lambda \sqrt{r} \|\mathbf{\Theta}_0 - \widehat{\mathbf{\Theta}}\|_F. \tag{S1.1}$$

Let $\gamma = C\tau\sqrt{r(p+q)\log n/n}$. Next we will construct $\widehat{\mathbf{\Theta}}_\eta = \mathbf{\Theta}_0 + \eta(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$ for some $\eta \in (0, 1]$ to satisfy $\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq \gamma$. If $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F < \gamma$, we set $\eta = 1$ and thus $\widehat{\mathbf{\Theta}}_\eta = \widehat{\mathbf{\Theta}}$; otherwise, we pick $\eta = \gamma / \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \in (0, 1)$ such that $\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F = \gamma$. Since $\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 = \eta(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)$ and $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \in \mathbb{C}$, it is easy to see $\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 \in \mathbb{C}$. Therefore, we have $\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 \in \mathbb{C} \cap \{\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq \gamma\}$. Since $\tau = C(\sigma_\delta n/((p+q)\log n))^{1/(1+\delta)}$ and $n \geq Cr(p+q)\log n$ for sufficiently large $C$, we have $\tau \geq C\gamma$, $\tau \geq C\sigma_\delta^{1/(1+\delta)}$ and $\gamma^2 \geq C\tau^2 r(p+q)\log n/n$. Therefore, the conditions of Lemma 5 have been verified. A direct application of Lemma 5 yields

$$\langle \nabla L(\widehat{\mathbf{\Theta}}_\eta) - \nabla L(\mathbf{\Theta}_0), \widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 \rangle \geq C\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F^2, \tag{S1.2}$$

with probability at least $1 - n^{-C}$. Moreover, by invoking Lemma C.1 of Sun et al. (2020), we have

$$\begin{aligned} \langle \nabla L(\widehat{\mathbf{\Theta}}_\eta) - \nabla L(\mathbf{\Theta}_0), \widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 \rangle &\leq \eta \langle \nabla L(\widehat{\mathbf{\Theta}}) - \nabla L(\mathbf{\Theta}_0), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 \rangle \\ &\leq C\eta\lambda\sqrt{r}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F. \end{aligned} \tag{S1.3}$$

Therefore, it follows from (S1.2) and (S1.3) that

$$C\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F^2 \leq C\eta\lambda\sqrt{r}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r}\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F,$$

where the last inequality uses $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0 = \eta^{-1}(\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0)$. Rearranging the above inequality yields

$$\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r} < \gamma,$$

where we use $((p+q)\log n/n)^{\delta/(1+\delta)} < ((p+q)\log n/n)^{1/2-1/(1+\delta)}$ for any $\delta$.

By the construction of $\widehat{\mathbf{\Theta}}_\eta$ and $\|\widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F < \gamma$, we get $\widehat{\mathbf{\Theta}}_\eta = \widehat{\mathbf{\Theta}}$, implying

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r}.$$

This completes the proof for the first part.

Let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0$. By Lemma 1 which states that $\|\widehat{\mathbf{\Delta}}_{r^c}\|_* \leq 3\|\widehat{\mathbf{\Delta}}_r\|_*$, we have $\|\widehat{\mathbf{\Delta}}\|_* \leq \|\widehat{\mathbf{\Delta}}_{r^c}\|_* + \|\widehat{\mathbf{\Delta}}_r\|_* \leq 4\|\widehat{\mathbf{\Delta}}_r\|_* \leq C\sqrt{r}\|\widehat{\mathbf{\Delta}}_r\|_F \leq C\sqrt{r}\|\widehat{\mathbf{\Delta}}\|_F$. Therefore, we have

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_* \leq C\sqrt{r}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \leq C\lambda r,$$

which completes the proof of Theorem 1. □

## S1.3 Proof of Theorem 2

By the first order optimality condition of (2.1), there exists $\bar{\mathbf{\Theta}} \in \partial\|\widehat{\mathbf{\Theta}}\|_*$ such that

$$-\frac{1}{n}\sum_{i=1}^{n} \ell'_\tau(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle)\mathbf{X}_i + \lambda\bar{\mathbf{\Theta}} = \mathbf{0}. \tag{S1.4}$$

Let $\widehat{r} = \text{rank}(\widehat{\mathbf{\Theta}})$. By (2.1) of Koltchinskii et al. (2011), if $\widehat{\mathbf{\Theta}}$ has singular value decomposition $\widehat{\mathbf{\Theta}} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{p\times\widehat{r}}$ and $\mathbf{V} \in \mathbb{R}^{q\times\widehat{r}}$, we have

$$\partial\|\widehat{\mathbf{\Theta}}\|_* = \{\mathbf{U}\mathbf{V}^\top + \mathbf{U}^\perp\mathbf{W}(\mathbf{V}^\perp)^\top : \|\mathbf{W}\|_{op} \leq 1\},$$

where $\mathbf{U}^\perp \in \mathbb{R}^{p\times(p-\widehat{r})}$ and $\mathbf{V}^\perp \in \mathbb{R}^{q\times(q-\widehat{r})}$ are orthogonal matrices with columns orthogonal to those of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $\mathbf{U}_j$ and $\mathbf{V}_j$

denote the column of $\mathbf{U}$ and $\mathbf{V}$, respectively. Premultiplying $\mathbf{U}_j$ and postmultiplying $\mathbf{V}_j$ to (S1.4) yield

$$\frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i-\langle\mathbf{X}_i,\widehat{\boldsymbol{\Theta}}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j=\lambda,\quad j=1,\ldots,\widehat{r}. \tag{S1.5}$$

On the other hand, by the first order condition of $\boldsymbol{\Theta}_{0,\tau}:=\arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^{p\times q}}\mathbb{E}L(\boldsymbol{\Theta})$, we have $\mathbb{E}[\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_{0,\tau}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j]=0$. Then we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i-\langle\mathbf{X}_i,\widehat{\boldsymbol{\Theta}}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j\\
=\;&(P_n-P)[(\ell_\tau'(Y_i-\langle\mathbf{X}_i,\widehat{\boldsymbol{\Theta}}\rangle)-\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_0\rangle))\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j]\\
&+\mathbb{E}[(\ell_\tau'(Y_i-\langle\mathbf{X}_i,\widehat{\boldsymbol{\Theta}}\rangle)-\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_0\rangle))\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j]\\
&+(P_n-P)(\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_0\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j)\\
&+\mathbb{E}[\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_0\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j]-\mathbb{E}[\ell_\tau'(Y_i-\langle\mathbf{X}_i,\boldsymbol{\Theta}_{0,\tau}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j]\\
:=\;&T_{1j}+T_{2j}+T_{3j}+T_{4j}.
\end{aligned}$$

Let $\mathbf{T}_k=(T_{k1},\ldots,T_{k\widehat{r}})^\top$ for $k=1,2,3,4$. Then (S1.5) implies

$$\lambda\sqrt{\widehat{r}}\;\leq\;\|\mathbf{T}_1\|_2+\|\mathbf{T}_2\|_2+\|\mathbf{T}_3\|_2+\|\mathbf{T}_4\|_2.$$

We proceed to bound $\|\mathbf{T}_k\|_2$. Let $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{\widehat{r}})^\top$ and $\boldsymbol{\beta}=(\mathbf{V}_1\otimes\mathbf{U}_1,\ldots,\mathbf{V}_{\widehat{r}}\otimes\mathbf{U}_{\widehat{r}})\boldsymbol{\alpha}$. Then for any $\boldsymbol{\alpha}\in\mathbb{S}^{\widehat{r}-1}$, we have $\|\boldsymbol{\beta}\|_2^2=\sum_{j=1}^{\widehat{r}}\alpha_j^2\|\mathbf{V}_j\otimes\mathbf{U}_j\|_2^2=1$. Define $\bar{\Omega}=\{\boldsymbol{\Theta}\in\mathbb{R}^{p\times q}:\|\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\|_F\leq a_n\}$. Since $\ell_\tau'$ is Lipschitz continuous, we have

$$\begin{aligned}\|\mathbf{T}_2\|_2 = \sup_{\boldsymbol{\alpha}\in\mathbb{S}^{\widehat{r}-1}} \boldsymbol{\alpha}^\top \mathbf{T}_2 &\leq \sup_{\boldsymbol{\beta}\in\mathbb{S}^{pq-1},\boldsymbol{\Theta}\in\bar{\Omega}} C\mathbb{E}\left[|\langle \mathbf{X}_i, \boldsymbol{\Theta}-\boldsymbol{\Theta}_0\rangle||\boldsymbol{\beta}^\top \mathrm{vec}(\mathbf{X}_i)|\right] \\ &\leq \sup_{\boldsymbol{\beta}\in\mathbb{S}^{pq-1},\boldsymbol{\Theta}\in\bar{\Omega}} C\sqrt{\mathbb{E}|\langle \mathbf{X}_i, \boldsymbol{\Theta}-\boldsymbol{\Theta}_0\rangle|^2}\sqrt{\mathbb{E}|\boldsymbol{\beta}^\top \mathrm{vec}(\mathbf{X}_i)|^2} \\ &\leq \sup_{\boldsymbol{\Theta}\in\bar{\Omega}} C\|\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\|_F \leq Ca_n,\end{aligned}$$

where we use the sub-Gaussian property of $\mathrm{vec}(\mathbf{X}_i)$ in the last line. Similarly, by Lemma 9 which states that $\|\boldsymbol{\Theta}_{0,\tau}-\boldsymbol{\Theta}_0\|_F \leq C\sigma_\delta \tau^{-\delta} \leq a_n$, we have

$$\|\mathbf{T}_4\|_2 = \sup_{\boldsymbol{\alpha}\in\mathbb{S}^{\widehat{r}-1}} \boldsymbol{\alpha}^\top \mathbf{T}_4 \leq \sup_{\boldsymbol{\beta}\in\mathbb{S}^{pq-1},\boldsymbol{\Theta}\in\bar{\Omega}} C\mathbb{E}\left[|\langle \mathbf{X}_i, \boldsymbol{\Theta}-\boldsymbol{\Theta}_0\rangle||\boldsymbol{\beta}^\top \mathrm{vec}(\mathbf{X}_i)|\right] \leq Ca_n.$$

By similar arguments as that in Lemma 7, with high probability, we have

$$\|\mathbf{T}_1\|_2 \leq Ca_n\sqrt{\frac{\widehat{r}(p+q)\log n}{n}} + Ca_n\frac{\widehat{r}^{3/2}(p+q)^2(\log n)^2}{n}.$$

By similar arguments as Lemma 6 and $\tau = Cv_\delta\{n/((p+q)\log n)\}^{1/(1+\delta)}$, we have

$$\|\mathbf{T}_3\|_2 \leq Cv_\delta\left(\frac{(p+q)\log n}{n}\right)^{\delta/(1+\delta)},$$

with high probability. Therefore, we have

$$\begin{aligned}\lambda\sqrt{\widehat{r}} &\leq \|\mathbf{T}_1\|_2 + \|\mathbf{T}_2\|_2 + \|\mathbf{T}_3\|_2 + \|\mathbf{T}_4\|_2 \\ &\leq Ca_n\sqrt{\frac{\widehat{r}(p+q)\log n}{n}} + Ca_n\frac{\widehat{r}^{3/2}(p+q)^2(\log n)^2}{n} + Ca_n + Cv_\delta\left(\frac{(p+q)\log n}{n}\right)^{\delta/(1+\delta)}.\end{aligned}$$

By assumption $\lambda = Cv_\delta\left((p+q)\log n/n\right)^{\delta/(1+\delta)}$, the above inequality implies that $\widehat{r} \leq Cr$. This completes the proof. □

## S1.4 Proof of Theorem 3

This follows similar arguments as that in the proof of Theorem 1. By the first order condition of (3.5), there exists $\bar{\mathbf{\Theta}} \in \partial\|\widetilde{\mathbf{\Theta}}\|_*$ such that $\nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}) + \lambda\bar{\mathbf{\Theta}} = \mathbf{0}$. Then we have

$$\langle \nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}) - \nabla\widetilde{L}(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\rangle = \langle \lambda\bar{\mathbf{\Theta}} + \nabla\widetilde{L}(\mathbf{\Theta}_0), \mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\rangle.$$

Using Hölder's inequality yields

$$\langle \lambda\bar{\mathbf{\Theta}}, \mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\rangle \leq \lambda\|\bar{\mathbf{\Theta}}\|_{op}\|\mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\|_* \leq 8\lambda\|(\mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}})_r\|_*,$$

where the last inequality uses $\|\bar{\mathbf{\Theta}}\|_{op} \leq 2$ and $\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0 \in \mathbb{C}$ (proved similarly as Lemma 1). Similarly, we get

$$\langle \nabla\widetilde{L}(\mathbf{\Theta}_0), \mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\rangle \quad \leq \quad \|\nabla\widetilde{L}(\mathbf{\Theta}_0)\|_{op}\|\mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\|_* \leq 2\lambda\|(\mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}})_r\|_*,$$

where the last inequality uses Lemma 8. Therefore, we have

$$\langle \nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}) - \nabla\widetilde{L}(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\rangle \leq C\lambda\sqrt{r}\|\mathbf{\Theta}_0 - \widetilde{\mathbf{\Theta}}\|_F. \tag{S1.6}$$

On the other hand, by the definition of $\widetilde{L}$, it is easy to see

$$\langle \nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}) - \nabla\widetilde{L}(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\rangle = \langle \nabla L_1(\widetilde{\mathbf{\Theta}}) - \nabla L_1(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\rangle. \tag{S1.7}$$

Let $\gamma = C\tau\sqrt{r(p+q)\log n/n}$. Next we will construct $\widetilde{\mathbf{\Theta}}_\eta = \mathbf{\Theta}_0 + \eta(\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0)$ for some $\eta \in (0,1]$ to satisfy $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq \gamma$. If $\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F < \gamma$, we set $\eta = 1$ and thus $\widetilde{\mathbf{\Theta}}_\eta = \widetilde{\mathbf{\Theta}}$; otherwise, we pick $\eta = \gamma/\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F$ such that $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F = \gamma$. Since $\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0 \in \mathbb{C}$, we have $\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0 \in \mathbb{C} \cap \{\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq \gamma\}$. By assumptions $\tau = C(\sigma_\delta N/((p+q)\log N))^{1/(1+\delta)}$ and $n \geq Cr^{4/3}((p+q)\log n)^{5/3}$, we have $\tau \geq C\gamma$, $\tau \geq C\sigma_\delta^{1/(1+\delta)}$ and $\gamma^2 \geq C\tau^2 r(p+q)\log n/n$. Therefore, the conditions of Lemma 5 have been verified. Combining (S1.7) along with Lemma 5 yields

$$\langle \nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}_\eta) - \nabla\widetilde{L}(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\rangle \geq C\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F^2. \tag{S1.8}$$

Moreover, by invoking Lemma C.1 of Sun et al. (2020), we have

$$\langle \nabla\widetilde{L}(\widetilde{\mathbf{\Theta}}_\eta) - \nabla\widetilde{L}(\mathbf{\Theta}_0), \widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\rangle \;\;\leq\;\; C\eta\lambda\sqrt{r}\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F. \tag{S1.9}$$

Therefore, it follows from (S1.8) and (S1.9) that

$$C\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F^2 \leq C\eta\lambda\sqrt{r}\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r}\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F.$$

If $\gamma > C\lambda\sqrt{r}$, rearranging the above inequality yields $\|\widetilde{\mathbf{\Theta}}_\eta - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r} < \gamma$. By the construction of $\widetilde{\mathbf{\Theta}}_\eta$, we have $\widetilde{\mathbf{\Theta}}_\eta = \widetilde{\mathbf{\Theta}}$, which implies

$$\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F \leq C\lambda\sqrt{r}.$$

Therefore, it remains to prove $\gamma > C\lambda\sqrt{r}$. By our assumptions, we have

$$\begin{aligned}\lambda &= Ca_n\sqrt{\frac{r(p+q)\log n}{n}} + Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n} + Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\frac{\delta}{1+\delta}}\\ &= Cv_\delta r\left(\frac{(p+q)\log n}{n}\right)^{\frac{\delta}{1+\delta}+\frac{1}{2}} + Cv_\delta r^2\left(\frac{(p+q)\log n}{n}\right)^{\frac{\delta}{1+\delta}}\frac{((p+q)\log n)^2}{n}\\ &\quad + Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\frac{\delta}{1+\delta}}.\end{aligned}$$

Since $n \geq Cr^{4/3}((p+q)\log n)^{5/3}$, $\delta/(1+\delta) > 1/2 - 1/(1+\delta)$ and $\tau = C(\sigma_\delta N/((p+q)\log N))^{1/(1+\delta)}$, a direct calculation yields

$$C\sqrt{r}\lambda \;\leq\; Cv_\delta\sqrt{r}\left(\frac{(p+q)\log n}{n}\right)^{\frac{1}{2}-\frac{1}{1+\delta}} < C\tau\sqrt{r(p+q)\log n/n} = \gamma,$$

and thus we complete the proof of the first part.

Since $\widetilde{\boldsymbol{\Delta}} = \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0 \in \mathbb{C}$, we have $\|\widetilde{\boldsymbol{\Delta}}\|_* \leq C\sqrt{r}\|\widetilde{\boldsymbol{\Delta}}\|_F$, implying

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_* \leq C\sqrt{r}\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \leq C\lambda r,$$

which completes the proof of Theorem 3. □

## S1.5 Proof of Corollary 2

Let $\widetilde{\boldsymbol{\Theta}}^{t+1}$ denote the distributed estimator obtained at $(t+1)$-th round, $b_n = Cr\sqrt{\frac{(p+q)\log n}{n}} + C\frac{r^2(p+q)^2(\log n)^2}{n}$ and $S_N = Cv_\delta\sqrt{r}\left(\frac{(p+q)\log N}{N}\right)^{\frac{\delta}{1+\delta}}$. Applying

Theorem 3 recursively yields

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Theta}}^{t+1} - \boldsymbol{\Theta}_0\|_F &\leq b_n\|\widetilde{\boldsymbol{\Theta}}^{t} - \boldsymbol{\Theta}_0\|_F + S_N \leq b_n(b_n\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}_0\|_F + S_N) + S_N \\
&\leq \cdots \\
&\leq (b_n)^{t+1}\|\widetilde{\boldsymbol{\Theta}}^{0} - \boldsymbol{\Theta}_0\|_F + S_N\sum_{l=0}^{t}(b_n)^l \\
&= a_n(b_n)^{t+1} + \frac{S_N(1-(b_n)^{t+1})}{1-b_n}.
\end{aligned}$$

It is easy to see that when the number of iterations $T \geq \frac{\log(S_N/a_n)}{\log b_n}$, we have $a_n b_n^T \leq CS_N$, therefore implying

$$\|\widetilde{\boldsymbol{\Theta}}^{T} - \boldsymbol{\Theta}_0\|_F \leq CS_N = Cv_\delta\sqrt{r}\left(\frac{(p+q)\log N}{N}\right)^{\frac{\delta}{1+\delta}}.$$

Essentially the same arguments also apply to upper bound $\|\widetilde{\boldsymbol{\Theta}}^{T} - \boldsymbol{\Theta}_0\|_*$, which completes the proof. □

### S1.6 Proof of Theorem 4

For notational simplicity, we let $\{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ denote the full sample and $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ denote the subsample that is stored in $\mathcal{M}_1$. By the first order optimality condition of (3.5), there exists $\bar{\boldsymbol{\Theta}} \in \partial\|\widetilde{\boldsymbol{\Theta}}\|_*$ such that

$$\begin{aligned}
\lambda\bar{\boldsymbol{\Theta}} &= \frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i - \langle\mathbf{X}_i, \widetilde{\boldsymbol{\Theta}}\rangle)\mathbf{X}_i - \frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i - \langle\mathbf{X}_i, \widehat{\boldsymbol{\Theta}}\rangle)\mathbf{X}_i \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\ell_\tau'(Y_i - \langle\mathbf{X}_i, \widehat{\boldsymbol{\Theta}}\rangle)\mathbf{X}_i. \qquad \text{(S1.10)}
\end{aligned}$$

Let $\widetilde{r} = \text{rank}(\widetilde{\mathbf{\Theta}})$. By (2.1) of Koltchinskii et al. (2011), if $\widetilde{\mathbf{\Theta}}$ has singular value decomposition $\widetilde{\mathbf{\Theta}} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{p\times\widetilde{r}}$ and $\mathbf{V} \in \mathbb{R}^{q\times\widetilde{r}}$, we have

$$\partial\|\widetilde{\mathbf{\Theta}}\|_* = \{\mathbf{U}\mathbf{V}^\top + \mathbf{U}^\perp\mathbf{W}(\mathbf{V}^\perp)^\top : \|\mathbf{W}\|_{op} \leq 1\},$$

where $\mathbf{U}^\perp \in \mathbb{R}^{p\times(p-\widetilde{r})}$ and $\mathbf{V}^\perp \in \mathbb{R}^{q\times(q-\widetilde{r})}$ are orthogonal matrices with columns orthogonal to those of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $\mathbf{U}_j$ and $\mathbf{V}_j$ denote the column of $\mathbf{U}$ and $\mathbf{V}$, respectively. Premultiplying $\mathbf{U}_j$ and postmultiplying $\mathbf{V}_j$ to (S1.10) yield

$$\begin{aligned}\lambda &= \frac{1}{n}\sum_{i=1}^n \ell_\tau'(Y_i - \langle\mathbf{X}_i, \widetilde{\mathbf{\Theta}}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j - \frac{1}{n}\sum_{i=1}^n \ell_\tau'(Y_i - \langle\mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j \\ &\quad + \frac{1}{N}\sum_{i=1}^N \ell_\tau'(Y_i - \langle\mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j \\ &:= M_j, \quad j = 1, \ldots, \widetilde{r}.\end{aligned}$$

Let $\mathbf{M} = (M_1, \ldots, M_{\widetilde{r}})^\top$, then the above implies

$$\lambda\sqrt{\widetilde{r}} \leq \|\mathbf{M}\|_2. \tag{S1.11}$$

We continue to bound $\|\mathbf{M}\|_2$. Define $\mathbf{\Theta}_{0,\tau} := \arg\min_{\mathbf{\Theta}\in\mathbb{R}^{p\times q}} \mathbb{E}L(\mathbf{\Theta})$, then we have $\mathbb{E}[\ell_\tau'(Y_i - \langle\mathbf{X}_i, \mathbf{\Theta}_{0,\tau}\rangle)\mathbf{U}_j^\top\mathbf{X}_i\mathbf{V}_j] = 0$. We note that $M_j$ can be further decomposed as

$$
\begin{aligned}
M_j &= (P_n - P)[(\ell_\tau'(Y_i - \langle \mathbf{X}_i, \widetilde{\mathbf{\Theta}}\rangle) - \ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle))\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j] \\
&\quad + \mathbb{E}[(\ell_\tau'(Y_i - \langle \mathbf{X}_i, \widetilde{\mathbf{\Theta}}\rangle) - \ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle))\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j] \\
&\quad - (P_n - P)[(\ell_\tau'(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle) - \ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle))\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j] \\
&\quad + (P_N - P)[(\ell_\tau'(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle) - \ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle))\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j] \\
&\quad + (P_N - P)(\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle)\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j) \\
&\quad + \mathbb{E}[\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle)\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j] - \mathbb{E}[\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_{0,\tau}\rangle)\mathbf{U}_j^\top \mathbf{X}_i \mathbf{V}_j].
\end{aligned}
$$

Since $\tau = Cv_\delta\{N/((p+q)\log N)\}^{1/(1+\delta)}$, following the similar arguments as that in Theorem 2, we have

$$
\begin{aligned}
\|\mathbf{M}\|_2 &\le Ca_N\sqrt{\frac{\widetilde{r}(p+q)\log n}{n}} + Ca_N\frac{\widetilde{r}^{3/2}(p+q)^2(\log n)^2}{n} + Ca_N \\
&\quad + Ca_n\sqrt{\frac{\widehat{r}(p+q)\log n}{n}} + Ca_n\frac{\widehat{r}^{3/2}(p+q)^2(\log n)^2}{n} \\
&\quad + Ca_n\sqrt{\frac{\widehat{r}(p+q)\log N}{N}} + Ca_n\frac{\widehat{r}^{3/2}(p+q)^2(\log N)^2}{N} + Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\delta/(1+\delta)}.
\end{aligned}
$$

By assumption $\lambda = Cv_\delta\left((p+q)\log N/N\right)^{\delta/(1+\delta)}$, $n \ge C(\min\{p,q\}(p+q)\log n)^2$, $N \ge C(\min\{p,q\}(p+q)\log N)^4$ and Theorem 2 that $\widehat{r} \le Cr$, substituting the above into (S1.11) yields $\widetilde{r} \le Cr$. This completes the proof. □

## S2. Lemmas and their proofs

The following lemmas will be used in the proof of the main theorem.

**Lemma 1.** *Let* $\mathbb{C} = \{\mathbf{\Delta} \in \mathbb{R}^{p\times q} : \|\mathbf{\Delta}_{r^c}\|_* \leq 3\|\mathbf{\Delta}_r\|_*\}$. *Assume* $\lambda \geq 2\|\nabla L(\mathbf{\Theta}_0)\|_{op}$, *then we have* $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0$ *belongs to the set* $\mathbb{C}$.

**Proof of Lemma 1.** According to the definition of $\widehat{\mathbf{\Theta}}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle) + \lambda\|\widehat{\mathbf{\Theta}}\|_* \leq \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle) + \lambda\|\mathbf{\Theta}_0\|_*.$$

Due to the convexity of $\ell_\tau$, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle) - \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle) \\ \geq \quad & -\frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle)\langle \mathbf{X}_i, \widehat{\mathbf{\Delta}}\rangle, \end{aligned}$$

which, combined with Hölder's inequality and assumption $\lambda \geq 2\|\nabla L(\mathbf{\Theta}_0)\|_{op}$, implies that

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \widehat{\mathbf{\Theta}}\rangle) - \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle) \\ \geq \quad & -\Big\|\frac{1}{n}\sum_{i=1}^{n}\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle)\mathbf{X}_i\Big\|_{op}\|\widehat{\mathbf{\Delta}}\|_* \geq -\frac{\lambda}{2}\|\widehat{\mathbf{\Delta}}\|_*. \end{aligned}$$

Therefore, we have

$$\lambda\|\mathbf{\Theta}_0\|_* - \lambda\|\widehat{\mathbf{\Theta}}\|_* \geq -\frac{\lambda}{2}\|\widehat{\mathbf{\Delta}}\|_*.$$

Recall that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0$, $\widehat{\mathbf{\Delta}}_{r^c} = \mathcal{P}_{\mathcal{N}}\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Delta}}_r = \widehat{\mathbf{\Delta}} - \widehat{\mathbf{\Delta}}_{r^c}$. Combining triangle inequalities $\|\widehat{\mathbf{\Delta}}\|_* \leq \|\widehat{\mathbf{\Delta}}_r\|_* + \|\widehat{\mathbf{\Delta}}_{r^c}\|_*$, $\|\widehat{\mathbf{\Theta}}\|_* \geq \|\mathbf{\Theta}_0 + \widehat{\mathbf{\Delta}}_{r^c}\|_* - \|\widehat{\mathbf{\Delta}}_r\|_*$

along with the fact that $\|\boldsymbol{\Theta}_0 + \widehat{\boldsymbol{\Delta}}_{r^c}\|_* = \|\boldsymbol{\Theta}_0\|_* + \|\widehat{\boldsymbol{\Delta}}_{r^c}\|_*$, we have

$$-\frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_r\|_* - \frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_{r^c}\|_* \;\; \leq \;\; \lambda\|\boldsymbol{\Theta}_0\|_* - \lambda\|\widehat{\boldsymbol{\Theta}}\|_* \leq -\lambda\|\widehat{\boldsymbol{\Delta}}_{r^c}\|_* + \lambda\|\widehat{\boldsymbol{\Delta}}_r\|_*.$$

Rearranging the above terms generates

$$\|\widehat{\boldsymbol{\Delta}}_{r^c}\|_* \leq 3\|\widehat{\boldsymbol{\Delta}}_r\|_*,$$

which completes the proof. □

**Lemma 2.** *Let* $\mathbf{A}$ *be a* $p \times q$ *matrix,* $\mathcal{U}$ *be a* $1/4$*-covering of* $\mathbb{S}^{p-1}$ *and* $\mathcal{V}$ *be a* $1/4$*-covering of* $\mathbb{S}^{q-1}$*, then*

$$\|\mathbf{A}\|_{op} \leq 2 \max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \mathbf{u}_j^\top \mathbf{A}\mathbf{v}_k,$$

*with* $|\mathcal{U}| \leq 9^p$ *and* $|\mathcal{V}| \leq 9^q$*.*

**Proof of Lemma 2.** By the definition of the covering, for any $\mathbf{u} \in \mathbb{S}^{p-1}$ and $\mathbf{v} \in \mathbb{S}^{q-1}$, there exist $\mathbf{u}_j \in \mathcal{U}$ such that $\|\mathbf{u}_j - \mathbf{u}\|_2 \leq 1/4$ and $\mathbf{v}_k \in \mathcal{V}$ such that $\|\mathbf{v}_k - \mathbf{v}\|_2 \leq 1/4$, where $j = 1, \ldots, |\mathcal{U}|$ and $k = 1, \ldots, |\mathcal{V}|$. Then we have

$$\begin{aligned}
\mathbf{u}^\top \mathbf{A}\mathbf{v} &= \mathbf{u}^\top \mathbf{A}(\mathbf{v} - \mathbf{v}_k) + (\mathbf{u} - \mathbf{u}_j)^\top \mathbf{A}\mathbf{v}_k + \mathbf{u}_j^\top \mathbf{A}\mathbf{v}_k \\
&\leq \frac{1}{4}\|\mathbf{A}\|_{op} + \frac{1}{4}\|\mathbf{A}\|_{op} + \mathbf{u}_j^\top \mathbf{A}\mathbf{v}_k.
\end{aligned}$$

Taking the maximum over all $\mathbf{u}_j \in \mathcal{U}$ and $\mathbf{v}_k \in \mathcal{V}$, we have

$$\|\mathbf{A}\|_{op} = \sup_{\mathbf{u} \in \mathbb{S}^{p-1}, \mathbf{v} \in \mathbb{S}^{q-1}} \mathbf{u}^\top \mathbf{A}\mathbf{v} \leq \frac{1}{2}\|\mathbf{A}\|_{op} + \max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \mathbf{u}_j^\top \mathbf{A}\mathbf{v}_k.$$

Rearranging the above term yields

$$\|\mathbf{A}\|_{op} \leq 2 \max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \mathbf{u}_j^\top \mathbf{A} \mathbf{v}_k.$$

Furthermore, applying Lemma 5.2 of Vershynin (2010), the covering number satisfies $|\mathcal{U}| \leq 9^p$ and $|\mathcal{V}| \leq 9^q$. This completes the proof. □

**Lemma 3.** *Assume condition (A2) holds. Then with probability at least* $1 - n^{-C}$ *for some* $C > 0$*, we have*

$$\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_{op} \leq C\sqrt{(p+q)\log n}.$$

**Proof of Lemma 3.** Let $\mathcal{U}$ be a 1/4-covering of $\mathbb{R}^{p-1}$ and $\mathcal{V}$ be a 1/4-covering of $\mathbb{R}^{q-1}$. For any $\mathbf{u}_j \in \mathcal{U}$, $\mathbf{v}_k \in \mathcal{V}$, by assumption (A2) that $\text{vec}(\mathbf{X}_i)$ is sub-Gaussian and the fact that $\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k = (\mathbf{v}_k \otimes \mathbf{u}_j)^\top \text{vec}(\mathbf{X}_i)$, we have

$$\mathbb{E}\left[\exp\{t(\mathbf{v}_k \otimes \mathbf{u}_j)^\top \text{vec}(\mathbf{X}_i)\}\right] \leq \exp\{Ct^2\}, \tag{S2.12}$$

for any $t > 0$. Moreover, applying Lemma 2 generates

$$\mathbb{P}(\|\mathbf{X}_i\|_{op} > 2s) \leq \mathbb{P}\left(\max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k > s\right) \leq \sum_{j=1}^{|\mathcal{U}|} \sum_{k=1}^{|\mathcal{V}|} \mathbb{P}(\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k > s)$$

with $|\mathcal{U}| \leq 9^p$ and $|\mathcal{V}| \leq 9^q$. By Markov's inequality, we have

$$\begin{aligned} \sum_{j=1}^{|\mathcal{U}|} \sum_{k=1}^{|\mathcal{V}|} \mathbb{P}(\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k > s) &\leq \sum_{j=1}^{|\mathcal{U}|} \sum_{k=1}^{|\mathcal{V}|} e^{-ts} \mathbb{E}\left[\exp\{t(\mathbf{v}_k \otimes \mathbf{u}_j)^\top \text{vec}(\mathbf{X}_i)\}\right] \\ &\leq 9^{p+q} e^{-ts + Ct^2}, \end{aligned}$$

where the second inequality uses (S2.12). Setting $t = Cs$, $s = C\sqrt{(p+q)\log n}$ and taking the union bound over $i = 1, \dots, n$ yield

$$\mathbb{P}\left(\max_{1\leq i\leq n} \|\mathbf{X}_i\|_{op} > C\sqrt{(p+q)\log n}\right) \leq n^{-C},$$

which completes the proof. □

**Lemma 4.** *Let $w_i$ be i.i.d. Rademacher variables, that is,* $\mathbb{P}(w_i = \pm 1) = 1/2$. *Under condition (A2), we have*

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i\right\|_{op} \leq C\sqrt{\frac{p+q}{n}},$$

*for some constant $C > 0$.*

**Proof of Lemma 4.** Let $\mathcal{U}$ be a 1/4-covering of $\mathbb{R}^{p-1}$ and $\mathcal{V}$ be a 1/4-covering of $\mathbb{R}^{q-1}$. By covering arguments stated in Lemma 2, we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i\right\|_{op} \leq 2\mathbb{E}\left[\max_{\mathbf{u}_j\in\mathcal{U},\mathbf{v}_k\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k\right].$$

Since $\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k = (\mathbf{v}_k \otimes \mathbf{u}_j)^\top \mathrm{vec}(\mathbf{X}_i)$ is sub-Gaussian and $w_i$ is bounded, $w_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k$ is also sub-Gaussian. By standard bound on maxima of sub-Gaussian variables, we have

$$\mathbb{E}\left[\max_{\mathbf{u}_j\in\mathcal{U},\mathbf{v}_k\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k\right] \leq C\sqrt{\frac{\log(|\mathcal{U}||\mathcal{V}|)}{n}} \leq C\sqrt{\frac{\log(9^{p+q})}{n}},$$

which completes the proof. □

**Lemma 5.** *(Restricted strong convexity) Assume conditions (A1)-(A4),* $\mathbb{E}|\langle \mathbf{X}, \mathbf{B}\rangle|^4 \leq C\|\mathbf{B}\|_F^4$ *for any* $\mathbf{B} \in \mathbb{R}^{p\times q}$ *and some constant* $C > 0$, $\tau \geq C\max\{\sigma_\delta^{1/(1+\delta)}, \gamma\}$ *and* $n \geq C(\tau/\gamma)^2 r(p+q)\log n$ *for some sufficiently large* $C$ *and some* $\gamma > 0$. *Then with probability at least* $1 - n^{-C}$ *for some* $C > 0$, *we have*

$$\langle \nabla L(\boldsymbol{\Theta}) - \nabla L(\boldsymbol{\Theta}_0), \boldsymbol{\Theta} - \boldsymbol{\Theta}_0\rangle \geq C\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\|_F^2$$

*uniformly over* $\boldsymbol{\Delta} = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0 \in \{\|\boldsymbol{\Delta}\|_F \leq \gamma\} \cap \mathbb{C}$.

**Proof of Lemma 5.** This follows similar arguments as Lemma C.4 in Sun et al. (2020). For readability, we will split the proof into three steps.

**Step 1**. First, let $\boldsymbol{\Delta} = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$ and $f(\mathbf{X}_i) = \phi_{\tau\|\boldsymbol{\Delta}\|_F/(2\gamma)}(\langle \mathbf{X}_i, \boldsymbol{\Delta}\rangle I(|Y_i - \langle \mathbf{X}_i, \boldsymbol{\Theta}_0\rangle| \leq \tau/2))$, where $\phi_R(u) = u^2 I(|u| \leq R/2) + (|u| - R)^2 I(R/2 < |u| \leq R)$ for any $R > 0$. Note that $\phi_R$ is $R$-Lipschitz continuous and satisfies $u^2 I(|u| \leq R/2) \leq \phi_R(u) \leq u^2 I(|u| \leq R)$. We will show that

$$\langle \nabla L(\boldsymbol{\Theta}) - \nabla L(\boldsymbol{\Theta}_0), \boldsymbol{\Theta} - \boldsymbol{\Theta}_0\rangle \geq \frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i). \tag{S2.13}$$

Now we consider two cases.

Case 1: $|\langle \mathbf{X}_i, \boldsymbol{\Delta}\rangle| > \tau\|\boldsymbol{\Delta}\|_F/(2\gamma)$ or $|Y_i - \langle \mathbf{X}_i, \boldsymbol{\Theta}_0\rangle| > \tau/2$, we have $f(\mathbf{X}_i) = 0$. By the convexity of $\ell_\tau$, it is easy to see the left hand side of (S2.13) is greater than zero.

Case 2: $|\langle \mathbf{X}_i, \mathbf{\Delta}\rangle| \le \tau\|\mathbf{\Delta}\|_F/(2\gamma)$ and $|Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle| \le \tau/2$, we have

$$|Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0 + \mathbf{\Delta}\rangle| \le |Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle| + |\langle \mathbf{X}_i, \mathbf{\Delta}\rangle| \le \tau/2 + \tau\|\mathbf{\Delta}\|_F/(2\gamma) \le \tau,$$

for $\mathbf{\Delta} \in \{\|\mathbf{\Delta}\|_F \le \gamma\} \cap \mathbb{C}$. Combined with the fact that $\ell_\tau'(u) = u$ if $|u| \le \tau$, we have

$$\left\{-\ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}\rangle) + \ell_\tau'(Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle)\right\} \langle \mathbf{X}_i, \mathbf{\Delta}\rangle = |\langle \mathbf{X}_i, \mathbf{\Delta}\rangle|^2 \ge f(\mathbf{X}_i),$$

where the last inequality holds since $\phi_R(u) \le u^2$. Combining the above two cases completes the proof of the first step.

**Step 2**. We now proceed to establish the lower bound for $\mathbb{E}f(\mathbf{X})$. Note that $\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^2 \ge C\|\mathbf{\Delta}\|_F^2$ due to assumption (A3), and when $|\langle \mathbf{X}, \mathbf{\Delta}\rangle| \le \tau\|\mathbf{\Delta}\|_F/(4\gamma) \le \tau/4$ and $|Y - \langle \mathbf{X}, \mathbf{\Theta}_0\rangle| \le \tau/2$, $f(\mathbf{X}) = |\langle \mathbf{X}, \mathbf{\Delta}\rangle|^2$. Therefore,

$$\begin{aligned}
\mathbb{E}f(\mathbf{X}) &\ge \mathbb{E}[|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^2 I(|\langle \mathbf{X}, \mathbf{\Delta}\rangle| \le \tau/4) I(|Y - \langle \mathbf{X}, \mathbf{\Theta}_0\rangle| \le \tau/2)] \\
&\ge C\|\mathbf{\Delta}\|_F^2 - \sqrt{\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^4}\sqrt{\mathbb{P}(|\langle \mathbf{X}, \mathbf{\Delta}\rangle| > \tau/4)} \\
&\quad - \sqrt{\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^4}\sqrt{\mathbb{P}(|Y - \langle \mathbf{X}, \mathbf{\Theta}_0\rangle| > \tau/2)}.
\end{aligned}$$

Using Markov's inequality yields

$$\mathbb{P}(|\langle \mathbf{X}, \mathbf{\Delta}\rangle| > \tau/4) \le (4/\tau)^2\sqrt{\mathbb{E}|\langle \mathbf{X}, \mathbf{\Delta}\rangle|^4} \le C(4/\tau)^2\|\mathbf{\Delta}\|_F^2$$

and

$$\mathbb{P}(|Y - \langle \mathbf{X}, \mathbf{\Theta}_0\rangle| > \tau/2) \le (2/\tau)^{1+\delta}\mathbb{E}|e|^{1+\delta} \le (2/\tau)^{1+\delta}\sigma_\delta.$$

Therefore, combining the above inequalities with assumption (A4), $\tau \geq C\max\{\sigma_\delta^{1/(1+\delta)}, \gamma\}$ and $\|\mathbf{\Delta}\|_F \leq \gamma$, we have

$$\mathbb{E}f(\mathbf{X}) \geq C\|\mathbf{\Delta}\|_F^2 - C(4/\tau)\|\mathbf{\Delta}\|_F^3 - C\|\mathbf{\Delta}\|_F^2\sqrt{(2/\tau)^{1+\delta}\sigma_\delta} \geq C\|\mathbf{\Delta}\|_F^2,$$

which completes the proof of the second step.

**Step 3**. Finally, we consider the upper bound for $\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}|(P_n - P)f|$. It is easy to see that

$$\frac{\langle \nabla L(\mathbf{\Theta}) - \nabla L(\mathbf{\Theta}_0), \mathbf{\Theta} - \mathbf{\Theta}_0\rangle}{\|\mathbf{\Theta} - \mathbf{\Theta}_0\|_F^2} \geq \frac{\mathbb{E}f(\mathbf{X})}{\|\mathbf{\Theta} - \mathbf{\Theta}_0\|_F^2} - \sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}\frac{|(P_n - P)f|}{\|\mathbf{\Theta} - \mathbf{\Theta}_0\|_F^2}.$$

Since $\phi_{bR}(bu) = b^2\phi_R(u)$ for any $b > 0$, we have

$$f(\mathbf{X}_i) = \phi_{\tau/(2\gamma)}(\langle \mathbf{X}_i, \mathbf{\Delta}\rangle I(|Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle| \leq \tau/2)/\|\mathbf{\Delta}\|_F)\|\mathbf{\Delta}\|_F^2.$$

Define $g(\mathbf{X}_i) = f(\mathbf{X}_i)/\|\mathbf{\Delta}\|_F^2$. Then it suffices to prove $\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}|(P_n - P)g| \leq C$ for some constant $C$. Let $w_i$ be i.i.d. Rademacher variables. By the symmetrization argument as that in Pollard (1984), we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}|(P_n - P)g|\right]\\
\leq\ & 2\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}\left|\frac{1}{n}\sum_{i=1}^n w_i g(\mathbf{X}_i)\right|\right]\\
\leq\ & C\frac{\tau}{2\gamma}\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\leq\gamma\}}\left|\frac{1}{n}\sum_{i=1}^n w_i\langle \mathbf{X}_i, \mathbf{\Delta}\rangle I(|Y_i - \langle \mathbf{X}_i, \mathbf{\Theta}_0\rangle| \leq \tau/2)/\|\mathbf{\Delta}\|_F\right|\right],
\end{aligned}$$

where the last line uses the contraction inequality for the Rademacher complexity (see, for example, Theorem 2.2 of Koltchinskii (2011)) since $\phi_{\tau/2\gamma}$ is

$\tau/2\gamma$-Lipschitz continuous and $\phi_{\tau/2\gamma}(0)=0$. Since $|\langle \mathbf{X}_i, \mathbf{\Delta}\rangle| \le \|\mathbf{X}_i\|_{op}\|\mathbf{\Delta}\|_*$ and $I(\cdot)$ is bounded, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}\left|\frac{1}{n}\sum_{i=1}^n w_i\langle \mathbf{X}_i,\mathbf{\Delta}\rangle I(|Y_i-\langle \mathbf{X}_i,\mathbf{\Theta}_0\rangle|\le\tau/2)/\|\mathbf{\Delta}\|_F\right|\right]\\ \le\ &\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n w_i\mathbf{X}_i\right\|_{op}\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}\frac{\|\mathbf{\Delta}\|_*}{\|\mathbf{\Delta}\|_F}\le C\sqrt{r}\sqrt{\frac{p+q}{n}},\end{aligned}$$

where the last inequality uses Lemma 4 and $\|\mathbf{\Delta}\|_*\le 4\|\mathbf{\Delta}_r\|_*\le C\sqrt{r}\|\mathbf{\Delta}\|_F$. Therefore, we have

$$\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}|(P_n-P)g|\right]\le C\frac{\tau\sqrt{r}}{\gamma}\sqrt{\frac{p+q}{n}}.$$

Recall that $0\le\phi_R(u)\le\min\{R^2/4,u^2\}$, we have $g\le\tau^2/(16\gamma^2)$ and $\mathbb{E}g^2\le\mathbb{E}(\langle\mathbf{X},\mathbf{\Delta}\rangle/\|\mathbf{\Delta}\|_F)^4\le C$. Applying Talagrand's concentration inequality yields

$$\begin{aligned}\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}|(P_n-P)g| &\le C\mathbb{E}\left[\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}|(P_n-P)g|\right]+C\sqrt{\frac{t}{n}}+C\frac{\tau^2t}{\gamma^2n}\\ &\le C\frac{\tau\sqrt{r}}{\gamma}\sqrt{\frac{p+q}{n}}+C\sqrt{\frac{t}{n}}+C\frac{\tau^2t}{\gamma^2n},\end{aligned}$$

with probability at least $1-\exp(-t)$ for any $t>0$. Let $t=C(p+q)\log n$ and assume $n\ge C(\tau/\gamma)^2r(p+q)\log n$ as well as $\tau\ge C\max\{\sigma_\delta^{1/(1+\delta)},\gamma\}$, then we have

$$\sup_{\mathbf{\Delta}\in\mathbb{C}\cap\{\|\mathbf{\Delta}\|_F\le\gamma\}}|(P_n-P)g|\le C\left(\frac{1}{\sqrt{\log n}}+\frac{1}{\sqrt{r}}+\frac{1}{r}\right)\le C,$$

with probability at least $1-n^{-C}$.

Combining **Steps 1-3**, we have

$$\langle \nabla L(\boldsymbol{\Theta}) - \nabla L(\boldsymbol{\Theta}_0), \boldsymbol{\Theta} - \boldsymbol{\Theta}_0 \rangle \geq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i) \geq C\|\boldsymbol{\Delta}\|_F^2,$$

with probability at least $1 - n^{-C}$. This completes the proof. □

**Lemma 6.** *Assume conditions (A1)-(A4) hold, then with probability at least $1 - n^{-C}$ for some $C > 0$, we have*

$$\|\nabla L(\boldsymbol{\Theta}_0)\|_{op} \leq C\sqrt{\frac{\sigma_\delta \tau^{1-\delta}(p+q)\log n}{n}} + C\frac{\tau(p+q)\log n}{n} + C\sigma_\delta \tau^{-\delta}.$$

*In particular, assume $\tau = Cv_\delta\{n/((p+q)\log n)\}^{1/(1+\delta)}$, then we have*

$$\|\nabla L(\boldsymbol{\Theta}_0)\|_{op} \leq Cv_\delta \left(\frac{(p+q)\log n}{n}\right)^{\delta/(1+\delta)}.$$

**Proof of Lemma 6.** By triangular inequality, we have

$$\|\nabla L(\boldsymbol{\Theta}_0)\|_{op} \leq \|\nabla L(\boldsymbol{\Theta}_0) - \mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op} + \|\mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op}.$$

Therefore, it remains to bound $\|\nabla L(\boldsymbol{\Theta}_0) - \mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op}$ and $\|\mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op}$, respectively. Let $\xi_i = \ell'_\tau(Y_i - \langle \mathbf{X}_i, \boldsymbol{\Theta}_0\rangle) = \ell'_\tau(e_i)$, then we have

$$\nabla L(\boldsymbol{\Theta}_0) - \mathbb{E}\nabla L(\boldsymbol{\Theta}_0) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\xi_i \mathbf{X}_i - \mathbb{E}(\xi_i \mathbf{X}_i)\right\}.$$

Let $\mathcal{U}$ be a 1/4-covering of $\mathbb{R}^{p-1}$ and $\mathcal{V}$ be a 1/4-covering of $\mathbb{R}^{q-1}$. By standard covering arguments as stated in Lemma 2, we have

$$\|\nabla L(\boldsymbol{\Theta}_0) - \mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op} \leq 2\max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \left| \frac{1}{n}\sum_{i=1}^{n}\left\{\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \mathbb{E}[\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k]\right\}\right|,$$

with $|\mathcal{U}| \leq 9^p$ and $|\mathcal{V}| \leq 9^q$. It follows from assumption (A2) that $\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k = (\mathbf{v}_k \otimes \mathbf{u}_j)^\top \mathrm{vec}(\mathbf{X}_i)$ is sub-Gaussian and satisfies

$$\mathbb{E}|\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k|^l \leq 2l \int_0^\infty t^{l-1} e^{-t^2/c_0^2} dt \leq c_0^l l \Gamma(l/2), l \geq 1 \tag{S2.14}$$

where $\Gamma(\cdot)$ denotes the Gamma function. By assumption (A4) and the fact that $\xi_i = \mathrm{sign}(e_i)\min(|e_i|, \tau)$, we have

$$\mathbb{E}(\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k)^2 \leq \tau^{1-\delta} \mathbb{E}\left[(\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k)^2 \mathbb{E}(\xi_i^{1+\delta} | \mathbf{X}_i)\right] \leq 2c_0^2 \tau^{1-\delta} \sigma_\delta,$$

$$\mathbb{E}|\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k|^l \leq \tau^{l-1-\delta} \mathbb{E}\left[|\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k|^l \mathbb{E}(\xi_i^{1+\delta} | \mathbf{X}_i)\right] \leq (l!/2)(c_0\tau/2)^{l-2} 2c_0^2 \tau^{1-\delta} \sigma_\delta,$$

for $l \geq 3$. Then applying Bernstein's inequality yields

$$\left| \frac{1}{n} \sum_{i=1}^n \left\{ \xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \mathbb{E}[\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k] \right\} \right| \leq C\sqrt{\frac{\sigma_\delta \tau^{1-\delta} z}{n}} + C\frac{\tau z}{n},$$

with probability at least $1 - 2e^{-z}$ for any $z > 0$. Taking the union bound over $\mathbf{u}_j \in \mathcal{U}$ and $\mathbf{v}_k \in \mathcal{V}$, we have

$$\begin{aligned} \|\nabla L(\mathbf{\Theta}_0) - \mathbb{E}\nabla L(\mathbf{\Theta}_0)\|_{op} &\leq 2 \max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \mathbb{E}[\xi_i \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k] \right\} \right| \\ &\leq C\sqrt{\frac{\sigma_\delta \tau^{1-\delta} z}{n}} + C\frac{\tau z}{n}, \end{aligned} \tag{S2.15}$$

with probability at least $1 - 9^{p+q} \cdot 2e^{-z}$. Setting $z = C(p+q)\log n$, we have

$$\|\nabla L(\mathbf{\Theta}_0) - \mathbb{E}\nabla L(\mathbf{\Theta}_0)\|_{op} \leq C\sqrt{\frac{\sigma_\delta \tau^{1-\delta}(p+q)\log n}{n}} + C\frac{\tau(p+q)\log n}{n},$$

with probability at least $1 - n^{-C}$. Moreover, it is easy to see

$$\|\mathbb{E}\nabla L(\mathbf{\Theta}_0)\|_{op} = \sup_{\mathbf{u} \in \mathbb{S}^{p-1}, \mathbf{v} \in \mathbb{S}^{q-1}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}|\xi_i \mathbf{u}^\top \mathbf{X}_i \mathbf{v}| \leq C\sigma_\delta \tau^{-\delta}.$$

Combining the above two inequalities completes the proof. □

**Lemma 7.** *Assume conditions (A1)-(A4) hold and* $\mathbb{E}|\langle\mathbf{X},\mathbf{B}\rangle|^4 \leq C\|\mathbf{B}\|_F^4$ *for any* $\mathbf{B} \in \mathbb{R}^{p\times q}$ *and some positive constant* $C$*. Then with probability at least* $1-n^{-C}$ *for some* $C>0$*, we have*

$$\begin{aligned} & \|\nabla L(\mathbf{\Theta}) - \nabla L(\mathbf{\Theta}_0) - \mathbb{E}\nabla L(\mathbf{\Theta}) + \mathbb{E}\nabla L(\mathbf{\Theta}_0)\|_{op} \\ \leq \;& Ca_n\sqrt{\frac{r(p+q)\log n}{n}} + Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n} \end{aligned}$$

*uniformly over* $\Omega = \{\mathbf{\Theta}\in\mathbb{R}^{p\times q} : \|\mathbf{\Theta}-\mathbf{\Theta}_0\|_F \leq a_n, \mathrm{rank}(\mathbf{\Theta}) \leq Cr\}$.

**Proof of Lemma 7.** Let $\ell_\tau'(\mathbf{\Theta}) = \ell_\tau'(Y_i - \langle\mathbf{X}_i,\mathbf{\Theta}\rangle)$ and $g(\mathbf{X}_i,\mathbf{\Theta}) = \ell_\tau'(\mathbf{\Theta})\mathbf{X}_i - \ell_\tau'(\mathbf{\Theta}_0)\mathbf{X}_i - \mathbb{E}[\ell_\tau'(\mathbf{\Theta})\mathbf{X}_i] + \mathbb{E}[\ell_\tau'(\mathbf{\Theta}_0)\mathbf{X}_i]$. We consider $\Omega = \{\mathbf{\Theta}\in\mathbb{R}^{p\times q} : \|\mathbf{\Theta}-\mathbf{\Theta}_0\|_F \leq a_n, \mathrm{rank}(\mathbf{\Theta}) \leq Cr\}$. Let $\mathcal{N}_\Omega$ be the $a_n n^{-M}$-covering of $\Omega$ with sufficiently large $M$, then we have

$$\begin{aligned} & \sup_{\mathbf{\Theta}\in\Omega}\left\|\frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i,\mathbf{\Theta})\right\|_{op} \\ \leq \;& \max_{\bar{\mathbf{\Theta}}\in\mathcal{N}_\Omega}\left\|\frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i,\bar{\mathbf{\Theta}})\right\|_{op} + \max_{\bar{\mathbf{\Theta}}\in\mathcal{N}_\Omega}\sup_{\|\bar{\mathbf{\Theta}}-\mathbf{\Theta}\|_F\leq a_n n^{-M}}\left\|\frac{1}{n}\sum_{i=1}^n(g(\mathbf{X}_i,\mathbf{\Theta}) - g(\mathbf{X}_i,\bar{\mathbf{\Theta}}))\right\|_{op} \\ := \;& T_1 + T_2. \end{aligned}$$

In the following, we will separately bound $T_1$ and $T_2$. Let $\mathcal{U}$ be a 1/4-covering of $\mathbb{R}^{p-1}$ and $\mathcal{V}$ be a 1/4-covering of $\mathbb{R}^{q-1}$. Since $\ell_\tau'$ is Lipschitz continuous, for any $\mathbf{u}_j \in \mathcal{U}$ and $\mathbf{v}_k \in \mathcal{V}$, we have

$$\begin{aligned}\mathbb{E}\left(\ell_\tau^{'}(\bar{\boldsymbol{\Theta}})\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \ell_\tau^{'}(\boldsymbol{\Theta}_0)\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k\right)^2 &\leq C\sqrt{\mathbb{E}(\langle \mathbf{X}_i, \bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\rangle)^4}\sqrt{\mathbb{E}(\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k)^4} \\ &\leq C\|\bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2 \leq Ca_n^2,\end{aligned}$$

where the first line uses Cauchy-Schwartz inequality and the second inequality uses $\mathbb{E}(\langle \mathbf{X}_i, \bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\rangle)^4 \leq C\|\bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^4$ and (S2.14). Moreover,

$$\begin{aligned}\left|\ell_\tau^{'}(\bar{\boldsymbol{\Theta}})\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \ell_\tau^{'}(\boldsymbol{\Theta}_0)\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k\right| &\leq C(\max_{1\leq i\leq n}\|\mathbf{X}_i\|_{op})^2\|\bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_* \\ &\leq C\sqrt{r}a_n(p+q)\log n,\end{aligned}$$

where the last inequality uses $\|\bar{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_* \leq C\sqrt{r}a_n$ and Lemma 3 which states that $\max_{1\leq i\leq n}\|\mathbf{X}_i\|_{op} \leq C\sqrt{(p+q)\log n}$ with high probability. Therefore, applying Bernstein's inequality yields

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n(\ell_\tau^{'}(\bar{\boldsymbol{\Theta}}) - \ell_\tau^{'}(\boldsymbol{\Theta}_0))\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k - \mathbb{E}\left[(\ell_\tau^{'}(\bar{\boldsymbol{\Theta}}) - \ell_\tau^{'}(\boldsymbol{\Theta}_0))\mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k\right]\right| \\ \leq\ & C\sqrt{\frac{za_n^2}{n}} + C\frac{z\sqrt{r}a_n(p+q)\log n}{n} := \phi(z), \qquad \text{(S2.16)}\end{aligned}$$

with probability at least $1 - 2e^{-z}$ for any $z > 0$. By applying Lemma 5.2 of Vershynin (2010), the covering number satisfies $|\mathcal{N}_\Omega| \leq (1 + 2n^M)^{Cr(p+q)}$. Combining with the union bound generates

$$\begin{aligned}\mathbb{P}\{T_1 \geq \phi(z)\} &= \mathbb{P}\left(\max_{\bar{\boldsymbol{\Theta}}\in\mathcal{N}_\Omega}\left\|\frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i, \bar{\boldsymbol{\Theta}})\right\|_{op} \geq \phi(z)\right) \\ &\leq Cn^{CMr(p+q)}\max_{\bar{\boldsymbol{\Theta}}\in\mathcal{N}_\Omega}\mathbb{P}\left(\max_{\mathbf{u}_j\in\mathcal{U},\mathbf{v}_k\in\mathcal{V}}\frac{1}{n}\sum_{i=1}^n\mathbf{u}_j^\top g(\mathbf{X}_i, \bar{\boldsymbol{\Theta}})\mathbf{v}_k \geq \phi(z)\right) \\ &\leq Cn^{CMr(p+q)}9^{p+q}e^{-z},\end{aligned}$$

where the second inequality uses Lemma 2 and the last line uses (S2.16). Setting $z = Cr(p+q)\log n$, we have

$$T_1 \leq Ca_n\sqrt{\frac{r(p+q)\log n}{n}} + Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n}, \qquad \text{(S2.17)}$$

with probability at least $1 - n^{-C}$.

Next we continue to bound $T_2$. Define $h(\mathbf{X}_i) = \sup_{\|\bar{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F \leq a_n n^{-M}}[g(\mathbf{X}_i, \boldsymbol{\Theta}) - g(\mathbf{X}_i, \bar{\boldsymbol{\Theta}})]$. Similarly, we have

$$|\mathbf{u}_j^\top h(\mathbf{X}_i)\mathbf{v}_k| \leq \left| \sup_{\|\bar{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F \leq a_n n^{-M}} \langle \boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}, \mathbf{X}_i \rangle \mathbf{u}_j^\top \mathbf{X}_i \mathbf{v}_k \right| \leq C\sqrt{r}a_n n^{-M}(p+q)\log n,$$

which, together with the union bound implies that

$$T_2 \leq C\frac{\sqrt{r}a_n(p+q)\log n}{n^M}, \qquad \text{(S2.18)}$$

with high probability. Therefore, combining (S2.17)-(S2.18) and $M$ sufficiently large yield

$$\sup_{\boldsymbol{\Theta}\in\Omega} \left\| \frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i, \boldsymbol{\Theta}) \right\|_{op} \leq Ca_n\sqrt{\frac{r(p+q)\log n}{n}} + Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n},$$

with probability at least $1 - n^{-C}$. This completes the proof. □

**Lemma 8.** *Assume conditions (A1)-(A5) hold and* $\tau = Cv_\delta\{N/((p+q)\log N)\}^{1/(1+\delta)}$ *for sufficiently large constant* $C$*. Then with probability at* *least* $1-n^{-C}$ *for some* $C>0$*, we have*

$$\begin{aligned}\|\nabla\widetilde{L}(\boldsymbol{\Theta}_0)\|_{op} &\le Ca_n\sqrt{\frac{r(p+q)\log n}{n}}+Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n}\\ &\quad +Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\delta/(1+\delta)}.\end{aligned}$$

**Proof of Lemma 8.** Recall that $\widetilde{L}(\boldsymbol{\Theta})=L_1(\boldsymbol{\Theta})-\langle\boldsymbol{\Theta},\nabla L_1(\widehat{\boldsymbol{\Theta}})-\nabla L(\widehat{\boldsymbol{\Theta}})\rangle$, then direct calculation generates

$$\nabla\widetilde{L}(\boldsymbol{\Theta}_0) = (\nabla L(\widehat{\boldsymbol{\Theta}})-\nabla L(\boldsymbol{\Theta}_0))-(\nabla L_1(\widehat{\boldsymbol{\Theta}})-\nabla L_1(\boldsymbol{\Theta}_0))+\nabla L(\boldsymbol{\Theta}_0).$$

By triangle inequality, it holds that

$$\begin{aligned}\|\nabla\widetilde{L}(\boldsymbol{\Theta}_0)\|_{op} &\le \|\nabla L(\widehat{\boldsymbol{\Theta}})-\nabla L(\boldsymbol{\Theta}_0)-\mathbb{E}\nabla L(\widehat{\boldsymbol{\Theta}})+\mathbb{E}\nabla L(\boldsymbol{\Theta}_0)\|_{op}\\ &\quad +\|\nabla L_1(\widehat{\boldsymbol{\Theta}})-\nabla L_1(\boldsymbol{\Theta}_0)-\mathbb{E}\nabla L_1(\widehat{\boldsymbol{\Theta}})+\mathbb{E}\nabla L_1(\boldsymbol{\Theta}_0)\|_{op}\\ &\quad +\|\nabla L(\boldsymbol{\Theta}_0)\|_{op}\\ &:= I_1+I_2+I_3.\end{aligned}$$

Next we will bound $I_1$-$I_3$ separately. By invoking Lemma 7, we have

$$I_1+I_2\le Ca_n\sqrt{\frac{r(p+q)\log n}{n}}+Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n},$$

with probability at least $1-n^{-C}$. By Lemma 6 and assumption $\tau = Cv_\delta(N/((p+q)\log N))^{1/(1+\delta)}$, we have

$$I_3\le Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\delta/(1+\delta)},$$

with probability at least $1-n^{-C}$. Therefore, combining the above inequalities yields

$$\begin{aligned}\|\nabla\widetilde{L}(\mathbf{\Theta}_0)\|_{op} &\leq Ca_n\sqrt{\frac{r(p+q)\log n}{n}}+Ca_n\frac{r^{3/2}(p+q)^2(\log n)^2}{n}\\ &\quad +Cv_\delta\left(\frac{(p+q)\log N}{N}\right)^{\delta/(1+\delta)},\end{aligned}$$

with probability at least $1-n^{-C}$. This completes the proof. □

**Lemma 9.** *Assume conditions (A1)-(A4) hold,* $\mathbb{E}|\langle\mathbf{u},\mathbf{x}\rangle|^4\leq C\|\mathbf{u}\|_2^4$ *for any* $\mathbf{u}\in\mathbb{R}^{pq}$ *and some constant* $C$*, and* $\tau\geq C\sigma_\delta^{1/(1+\delta)}$ *for sufficiently large* $C$*. Define* $\mathbf{\Theta}_{0,\tau}=\arg\min_{\mathbf{\Theta}\in\mathbb{R}^{p\times q}}\mathbb{E}L(\mathbf{\Theta})$*. Then we have*

$$\|\mathbf{\Theta}_0-\mathbf{\Theta}_{0,\tau}\|_F\leq C\sigma_\delta\tau^{-\delta}.$$

**Proof of Lemma 9.** Let $\mathbf{\Delta}_0=\mathbf{\Theta}_0-\mathbf{\Theta}_{0,\tau}$, $\boldsymbol{\delta}_0=\mathrm{vec}(\mathbf{\Delta}_0)$ and $h(\mathbf{\Theta})=\mathbb{E}[\ell_\tau(Y-\langle\mathbf{X},\mathbf{\Theta}\rangle)]$. By the first order condition, we have $\nabla h(\mathbf{\Theta}_{0,\tau})=\mathbf{0}$. Then it follows the mean value theorem that

$$\langle\boldsymbol{\delta}_0,\nabla^2h(\mathbf{\Theta}_t)\boldsymbol{\delta}_0\rangle=\langle\boldsymbol{\delta}_0,\nabla h(\mathbf{\Theta}_0)-\nabla h(\mathbf{\Theta}_{0,\tau})\rangle=\langle\boldsymbol{\delta}_0,\nabla h(\mathbf{\Theta}_0)\rangle=-\mathbb{E}[\ell_\tau'(e)\boldsymbol{\delta}_0^\top\mathbf{x}],$$

where $\mathbf{\Theta}_t=t\mathbf{\Theta}_0+(1-t)\mathbf{\Theta}_{0,\tau}$ for some $t\in[0,1]$. Since $\mathbb{E}(e|\mathbf{x})=0$, we have

$$\begin{aligned}\mathbb{E}[\ell_\tau'(e)|\mathbf{x}] &= \mathbb{E}[\{eI(|e|\leq\tau)+\tau\mathrm{sign}(e)I(|e|>\tau)\}|\mathbf{x}]-\mathbb{E}(e|\mathbf{x})\\ &= -\mathbb{E}[\{eI(|e|>\tau)-\tau\mathrm{sign}(e)I(|e|>\tau)\}|\mathbf{x}]\end{aligned}$$

and thus

$$|\mathbb{E}[\ell_\tau'(e)|\mathbf{x}]| \le \mathbb{E}[(|e|-\tau)I(|e|>\tau)|\mathbf{x}] \le \frac{\mathbb{E}[(|e|^{1+\delta}-\tau^{1+\delta})I(|e|>\tau)|\mathbf{x}]}{\tau^\delta} \le \sigma_\delta \tau^{-\delta}.$$

Since $\mathbf{x}$ is sub-Gaussian, we have

$$|\mathbb{E}[\ell_\tau'(e)\boldsymbol{\delta}_0^\top \mathbf{x}]| \le |\mathbb{E}[\boldsymbol{\delta}_0^\top \mathbf{x}\mathbb{E}(\ell_\tau'(e)|\mathbf{x})]| \le C\sigma_\delta \tau^{-\delta}\|\boldsymbol{\delta}_0\|_2. \qquad \text{(S2.19)}$$

Next, we continue to derive the lower bound for $\langle \boldsymbol{\delta}_0, \nabla^2 h(\boldsymbol{\Theta}_t)\boldsymbol{\delta}_0\rangle$. Let $\widetilde{e} = Y - \langle \mathbf{X}, \boldsymbol{\Theta}_t\rangle$ and $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}\mathbf{x}^\top)$, then we have

$$\nabla^2 h(\boldsymbol{\Theta}_t) = \mathbb{E}[I(|\widetilde{e}| \le \tau)\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} - \mathbb{E}[I(|\widetilde{e}| > \tau)\mathbf{x}\mathbf{x}^\top].$$

By the definition of $\ell_\tau$, we have

$$\begin{aligned}\mathbb{E}[\ell_\tau(e)|\mathbf{x}] &\le \mathbb{E}\left[\left\{\frac{\tau^{1-\delta}}{2}|e|^{1+\delta}I(|e|\le\tau) + (\tau^{1-\delta}|e|^{1+\delta} - \frac{\tau^{2-\delta}}{2}|e|^\delta)I(|e|>\tau)\right\}|\mathbf{x}\right] \\ &\le C\sigma_\delta\tau^{1-\delta}.\end{aligned}$$

Combining the above inequality with the convexity of $h$, we have

$$h(\boldsymbol{\Theta}_t) \le th(\boldsymbol{\Theta}_0) + (1-t)h(\boldsymbol{\Theta}_{0,\tau}) \le h(\boldsymbol{\Theta}_0) \le C\sigma_\delta\tau^{1-\delta}.$$

Furthermore, for all $\boldsymbol{\Theta} \in \mathbb{R}^{p\times q}$, we have

$$h(\boldsymbol{\Theta}) \ge \mathbb{E}[(\tau|Y - \langle \mathbf{X}, \boldsymbol{\Theta}\rangle| - \tau^2/2)I(|Y - \langle \mathbf{X}, \boldsymbol{\Theta}\rangle| > \tau)].$$

Combining the above two inequalities, we have

$$\tau\mathbb{E}[|\widetilde{e}|I(|\widetilde{e}|>\tau)] - (\tau^2/2)\mathbb{P}(|\widetilde{e}|>\tau) \le C\sigma_\delta\tau^{1-\delta}. \qquad \text{(S2.20)}$$

Since $\tau\mathbb{E}[|\widetilde{e}|I(|\widetilde{e}| > \tau)] \geq \tau\mathbb{E}[|\widetilde{e}|I(|\widetilde{e}| > \tau)\tau/|\widetilde{e}|] = \tau^2\mathbb{P}(|\widetilde{e}| > \tau)$, (S2.20) deduces that $\mathbb{P}(|\widetilde{e}| > \tau) \leq C\sigma_\delta\tau^{-1-\delta}$. Therefore, we have

$$\begin{aligned}\langle \boldsymbol{\delta}_0, \nabla^2 h(\boldsymbol{\Theta}_t)\boldsymbol{\delta}_0\rangle &\geq \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_0\|_2^2 - \mathbb{E}[I(|\widetilde{e}| > \tau)\langle \mathbf{x}, \boldsymbol{\delta}_0\rangle^2] \\ &\geq \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_0\|_2^2 - \sqrt{\mathbb{E}|\langle \boldsymbol{\delta}_0, \mathbf{x}\rangle|^4}\sqrt{\mathbb{P}(|\widetilde{e}| > \tau)} \\ &\geq C\|\boldsymbol{\delta}_0\|_2^2 - C\|\boldsymbol{\delta}_0\|_2^2\sigma_\delta\tau^{-1-\delta}, \qquad \text{(S2.21)}\end{aligned}$$

where the third inequality uses assumption (A3) and $\mathbb{E}|\langle \boldsymbol{\delta}_0, \mathbf{x}\rangle|^4 \leq C\|\boldsymbol{\delta}_0\|_2^4$. Finally, combining (S2.19), (S2.21) and assumption $\tau \geq C\sigma_\delta^{1/(1+\delta)}$ complete the proof. □

**Remark 1.** It is worth noting that $\boldsymbol{\Theta}_{0,\tau}$ is different from $\boldsymbol{\Theta}_0$ generally and $\tau$ plays a critical role in the approximation bias $\|\boldsymbol{\Theta}_{0,\tau} - \boldsymbol{\Theta}_0\|_F$. However, as stated in Remark 2.2 of Pan et al. (2021), there are several scenarios that this approximation bias can vanish. In particular, when the conditional distribution $e|\mathbf{x}$ is symmetric around zero, $\boldsymbol{\Theta}_{0,\tau} = \boldsymbol{\Theta}_0$ for any $\tau$.

## S3. Simulation studies

In this section, we investigate the finite sample performances of the proposed method through two simulation examples. The initial estimator is obtained by Algorithm 1 on $\mathcal{M}_1$ and Algorithm 2 is used to obtain the distributed estimator. We compare five estimators:

(a) LHuber: the local Huber estimator using the only data on $\mathcal{M}_1$;

(b) NHuber: the naive average of all local Huber estimators calculated on $\{\mathcal{M}_j\}_{j=1}^m$;

(c) DHuber: the proposed distributed Huber estimator;

(d) DLS: the distributed least squares estimator using surrogate loss, that is, replacing Huber loss $\ell_\tau$ in (3.5) with least squares loss $\ell(\cdot) = (\cdot)^2$ and still applying the CSL framework for the purpose of fair comparison;

(e) DMed: the distributed median estimator using surrogate loss, that is, replacing Huber loss $\ell_\tau$ in (3.5) with median/absolute value loss $\ell(\cdot) = |\cdot|$ and still applying the CSL framework.

Particularly, in the case where only one machine ($m = 1$) is available, DLS will reduce to the central estimator in Negahban and Wainwright (2011b)/Zhou and Li (2014), DMed will degenerate to the central estimator in Elsener and van de Geer (2018), and LHuber, DHuber and NHuber all naturally correspond to the central Huber estimator. The regularization parameter $\lambda$ and robustification parameter $\tau$ are selected by 5-fold cross-validation.

### S3.1 Simulation example 1

We consider the heterogeneous model $Y_{ij} = \langle \mathbf{X}_{ij}, \mathbf{\Theta}_0 \rangle + (1 + |z_{ij}|)e_{ij}$, where $z_{ij}$ is the first component of $\mathrm{vec}(\mathbf{X}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Similar to Zhou and Li (2014), we generate a rank $r$ matrix $\mathbf{\Theta}_0 = \mathbf{B}_1 \mathbf{B}_2^\top$, with $\mathbf{B}_1 \in \mathbb{R}^{p \times r}$ and $\mathbf{B}_2 \in \mathbb{R}^{q \times r}$. Specifically, we first generate each entry of $\mathbf{B}_k$, $k = 1, 2$ from $N(0, 1)$, then control the percentage of non-zero entries using a Bernoulli distribution with probability $\sqrt{1 - 0.5^{1/r}}$ of being 1. Each entry of $\mathbf{X}_i$ is generated from $N(0, 1)$ and the noises follow three different distributions: the standard normal distribution, the $t$-distribution with 3 degrees of freedom, and the Pareto distribution with scale parameter 3 and shape parameter 2. Throughout the simulations, we consider rank $r = 2, 5$ and run different methods on a desktop computer with 32 GB RAM and Intel(R) Core(TM) i7-9700 CPU (3.00 GHz). The performances are evaluated by the estimation error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F$ and running time (in minutes), based on 100 replications.

In the first scenario, we fix the sample size $N = 10000$, dimension $p = 20$, $q = 20$ and vary the number of machines $m$ in $\{1, 5, 10, 20, 25\}$. Note that $m = 1$ corresponds to the central estimator using the full data. The errors are summarized in Figure 1 and the logarithm of running time is reported in Figure 2. For the second scenario, we fix $N = 10000$, $m = 10$,

$q = 20$ and vary $p$ in $\{10, 30, 50, 80, 100\}$. We represent the estimation errors in Figure 3.

Figure 1 shows that the errors generally increase with the number of machines (or, decrease with the local sample size since $N$ is fixed). The distributed estimators DHuber and NHuber always outperform the local estimator LHuber since the latter only uses local data on the first machine. Besides, DHuber can further improve the estimation performances over NHuber. In terms of noise distribution, (i) for normal errors (light-tailed and symmetric), all distributed estimators are comparable with each other; (ii) when the noises follow the $t$ distribution (heavy-tailed and symmetric), DHuber, NHuber and DMed perform better than DLS since in this case conditional mean and conditional median are the same and both Huber loss and median loss are robust to heavy-tailed errors, compared to DLS; (iii) when noises follow the Pareto distribution (heavy-tailed and asymmetric), the proposed method significantly outperforms the alternatives. DMed has large errors in mean estimation since it estimates the conditional median which is different from conditional mean. DLS does not perform well due to the sensitivity to heavy-tailed noises.

Regarding the running time presented in Figure 2, it is seen that the distributed methods can significantly reduce the computation cost compared to the central estimator ($m = 1$). Specifically, the running time decreases as the number of machines $m$ increases. Compared with NHuber, our approach takes slightly longer time. This is because our procedure requires solving at least two $l_1$ penalized optimization problems while NHuber takes almost the same time as LHuber. Combined with the error results in Figure 1, we indeed sacrifice a little bit time to achieve better estimation performances than NHuber, from the accuracy-efficiency trade-off perspective.

From Figure 3, we observe that the errors increase with dimension as expected and show similar phenomena in Figure 1 with regard to various noises. In summary, the proposed DHuber method enjoys appreciable improvement over NHuber and exhibits more stable estimation performances over various types of noises in contrast with DLS and DMed.

### S3.2 Simulation example 2

In the second simulation example, we use some geometric shapes to visualize the competitiveness of the proposed method. We consider the same heterogeneous model in simulation example 1. The coefficient matrix $\mathbf{\Theta}_0 \in \mathbb{R}^{64\times 64}$ is binary: entries of $\mathbf{\Theta}_0$ in the true signal region, including square ($r = 2$), cross ($r = 3$), circle ($r = 9$) and butterfly ($r = 30$) equals one and the entries in the remaining region equals zero (see the first column in Figure 4).

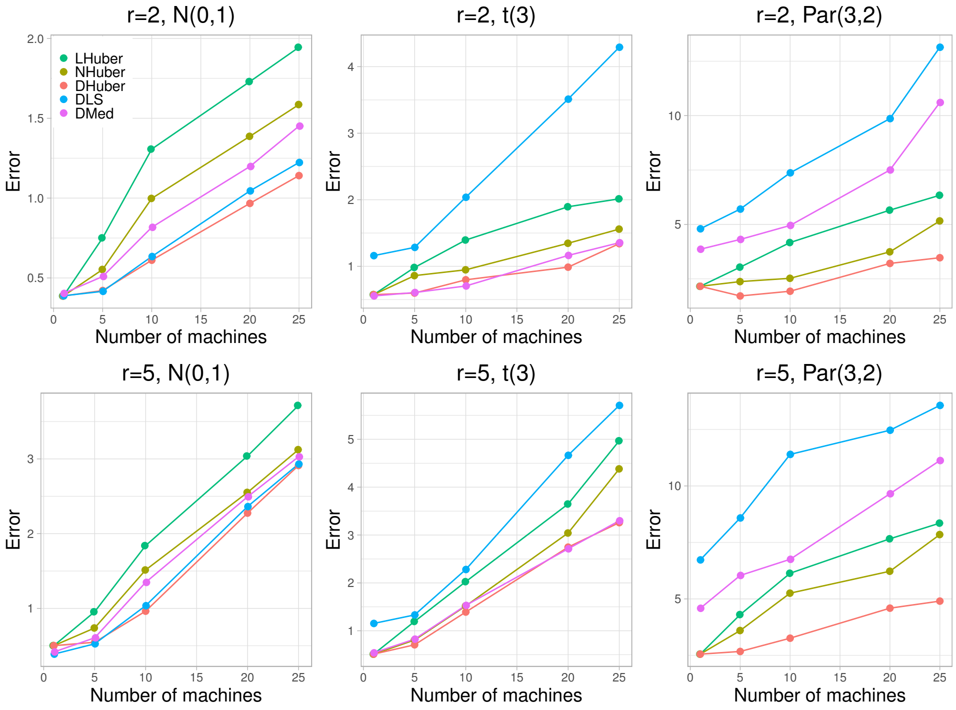

Figure 1: Averaged estimation errors of different methods over 100 replications with fixed $N = 10000$, $p = 20$, $q = 20$ and $m \in \{1, 5, 10, 20, 25\}$ ($m = 1$ corresponds to the central estimator) in the first scenario.

Each entry of $\mathbf{X}_i$ is generated from $N(0, 1)$. We generate noises from the standard normal distribution and Pareto distribution $Par(3, 2)$. Throughout the simulation, we use $N = 20000, m = 10$. The estimation errors are summarized in Table 1 and the estimated coefficient matrices for different noises are presented in Figures 4-5.

In Table 1, when noises are normal, the performances of different distributed methods are similar; when noises follow the Pareto distribution, the

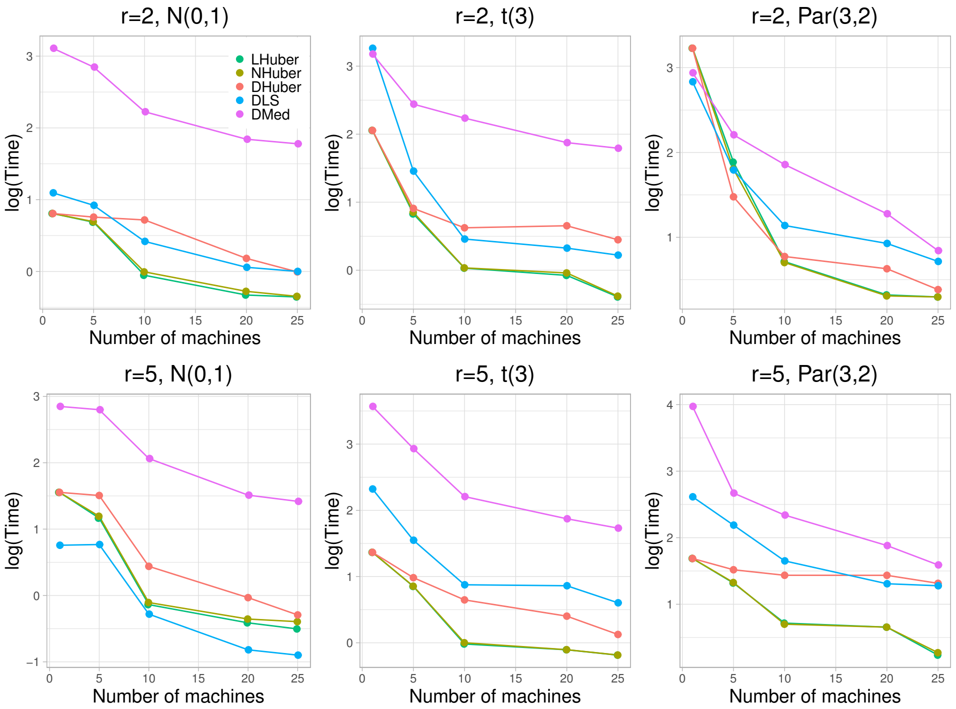

Figure 2: The logarithm of running time (in minutes) for different methods over 100 replications with fixed $N = 10000$, $p = 20$, $q = 20$ and $m \in \{1, 5, 10, 20, 25\}$ ($m = 1$ corresponds to the central estimator) in the first scenario.

proposed DHuber achieves the smallest errors compared to other methods. Besides, DHuber shows a significant advantage over LHuber and NHuber. Visually, in view of Figure 4 (normal noises), all five methods can recover the signal regions. However, from Figure 5 (Pareto noises), it is obvious that the DHuber outperforms the competitors. For example, when estimating the butterfly, the DLS generates a blurred image estimate and the DMed may yield some fuzzy shapes, both failing to identify the important

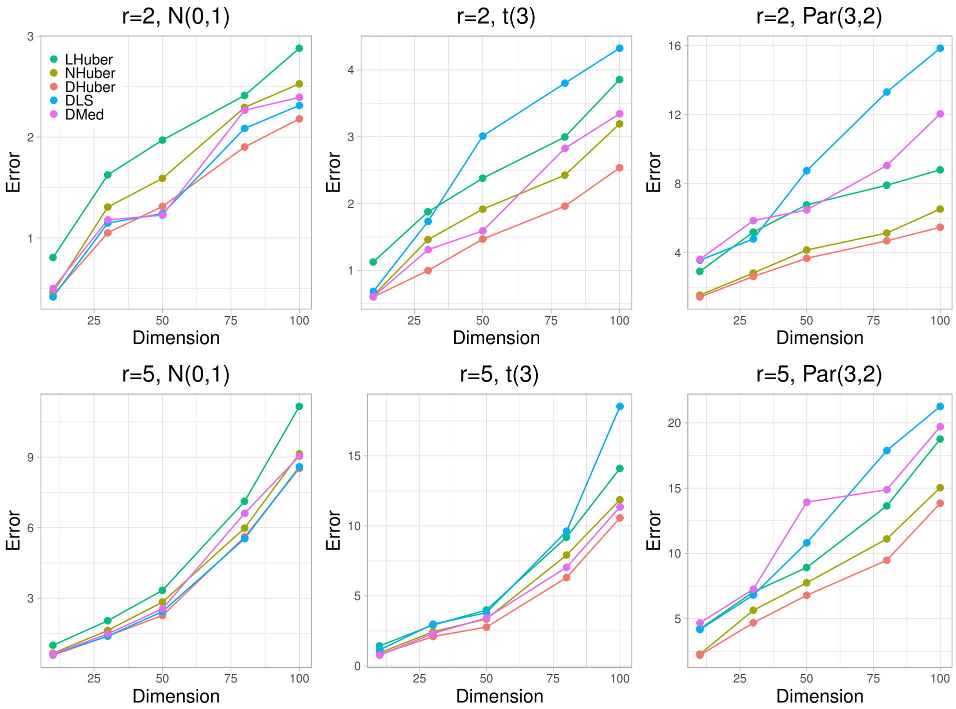

Figure 3: Averaged estimation errors of different methods over 100 replications with fixed $N = 10000$, $m = 10$, $q = 20$ and $p \in \{10, 30, 50, 80, 100\}$ in the second scenario.

signals clearly.

# Bibliography

Elsener, A. and S. van de Geer (2018). Robust low-rank matrix estimation. *The Annals of Statistics 46*(6), 3481–3509.

Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. New York: Springer.

Table 1: Estimation errors of different methods with various geometric shapes under normal and pareto noises.

| Geometric shape | Noise | LHuber | NHuber | DHuber | DLS | DMed |
|---|---|---|---|---|---|---|
| Square | N(0, 1) | 2.887 | 1.941 | 1.895 | 1.716 | 2.047 |
| | Par(3, 2) | 6.706 | 4.087 | 3.163 | 14.539 | 15.012 |
| Cross | N(0, 1) | 2.421 | 2.066 | 1.634 | 1.673 | 1.851 |
| | Par(3, 2) | 6.650 | 5.331 | 3.471 | 12.853 | 15.316 |
| Circle | N(0, 1) | 3.857 | 3.448 | 2.483 | 2.641 | 4.047 |
| | Par(3, 2) | 7.452 | 6.395 | 4.699 | 14.104 | 19.591 |
| Butterfly | N(0, 1) | 10.072 | 9.536 | 8.514 | 8.185 | 10.102 |
| | Par(3, 2) | 12.234 | 11.304 | 10.630 | 19.418 | 20.147 |

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Negahban, S. and M. J. Wainwright (2011a). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics 39*(2), 1069–1097.

Negahban, S. and M. J. Wainwright (2011b). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics 39*(2), 1069–1097.

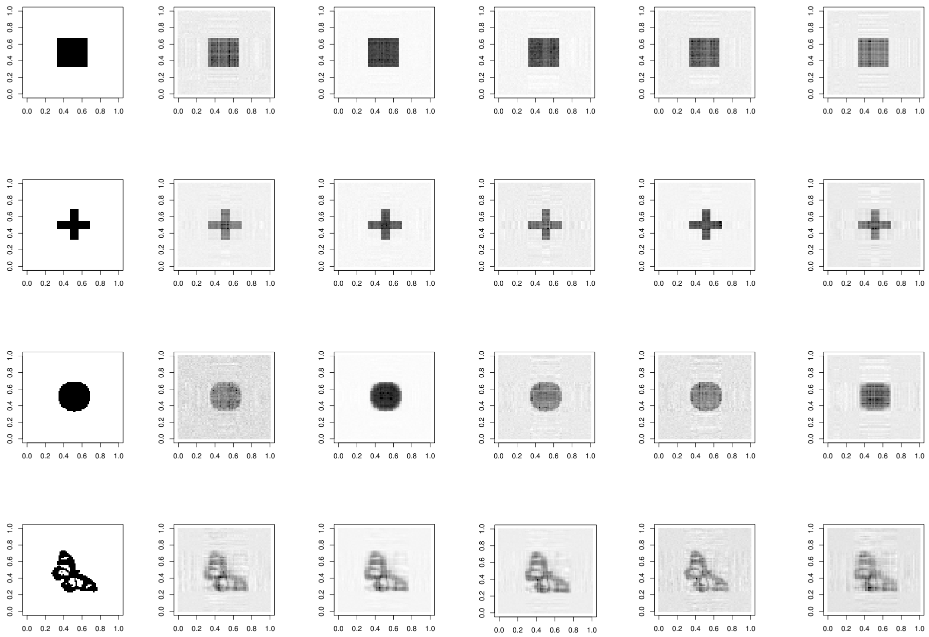

Figure 4: True signals, local Huber estimates, naive average of local Huber estimates, the proposed distributed Huber estimates, distributed least squares estimates and distributed median estimates (from left to right) under standard normal noises.

Pan, X., Q. Sun, and W.-X. Zhou (2021). Iteratively reweighted l1-penalized robust regression. *Electronic Journal of Statistics 15*(1), 3287–3348.

Pollard, D. (1984). *Convergence of stochastic processes.* New York: Springer-Verlag.

Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive huber regression. *Journal of the American Statistical Association 115*(529), 254–265.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of ran-

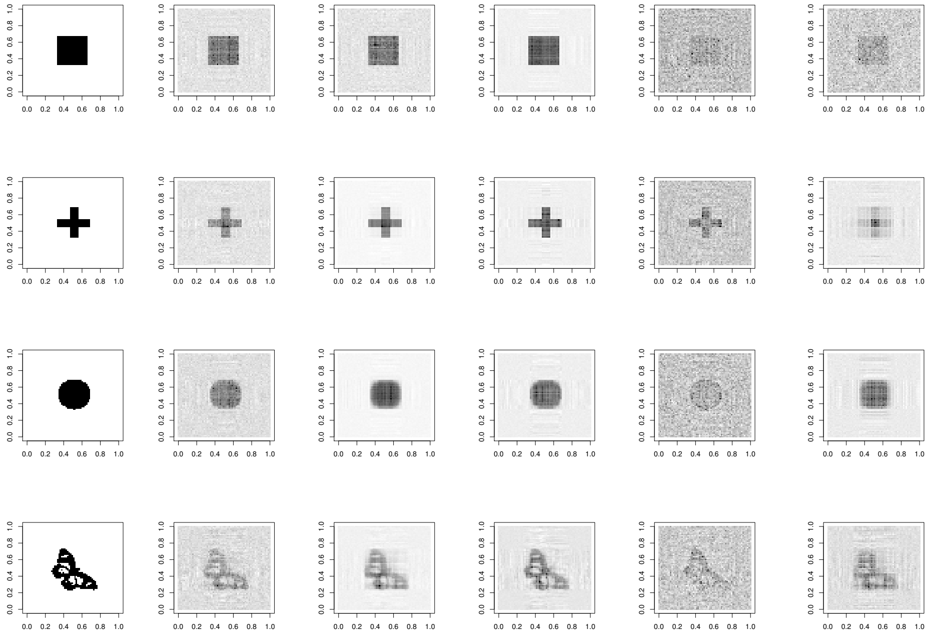

Figure 5: True signals, local Huber estimates, naive average of local Huber estimates, the proposed distributed Huber estimates, distributed least squares estimates and distributed median estimates (from left to right) under Pareto noises.

dom matrices. *arXiv preprint arXiv:1011.3027*.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B-Statistical Methodology 76*(2), 463–483.