Sparse and debiased adaptive Huber regression in distributed data: aggregated and communication-efficient approaches

Wei Ma[1], Junzhuo Gao[1], Lei Wang[1] and Heng Lian[2]

[1]*Nankai University and* [2]*City University of Hong Kong*

Supplementary Material

The Supplementary Material contains two algorithms for computing the SADL and SCDL estimators in Sections 2 and 3; additional simulation results under $\chi^2_3/LogN(0,1)$ errors as and numerical performance when $n = 200$; the simulation results in the presence of heteroskedastic error and outliers; the computing time comparison and variable selection results. All the proofs of Theorems and Corollaries are also provided.

S1 Two algorithms

Algorithm 1: The SADL adaptive Huber estimator

(i) For $1 \leq m \leq M$, obtain the mth local ℓ_1-penalized estimator $\hat{\boldsymbol{\beta}}_m$ by (2.3).

(ii) Construct the debiased lasso estimator $\hat{\boldsymbol{\beta}}^{\mathbf{d}}_m = (\hat{\beta}^{\mathbf{d}}_{m,1}, \cdots, \hat{\beta}^{\mathbf{d}}_{m,p})^\top$ and obtain the aggregated debiased lasso adaptive Huber estimator $\bar{\boldsymbol{\beta}}^{\mathbf{d}} = M^{-1}\sum_{m=1}^{M} \hat{\boldsymbol{\beta}}^{\mathbf{d}}_m$.

(iii) Apply the hard-thresholding procedure with $\nu = C_0\sqrt{\log p/N}$ and then finally output the SADL estimator $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) = (\mathcal{T}_\nu(\bar{\beta}_1^{\mathbf{d}}), \cdots, \mathcal{T}_\nu(\bar{\beta}_p^{\mathbf{d}}))^\top$.

Algorithm 2 : The SCDL and aggregated SCDL estimators

(i) Use $\hat{\boldsymbol{\beta}}_1$ as the initial estimator $\tilde{\boldsymbol{\beta}}^{[0]}$, for $1 \le t \le T$, broadcast $\tilde{\boldsymbol{\beta}}^{[t-1]}$ to each local site and compute $\nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}^{[t-1]})$ for $m = 1, \ldots, M$; transmit the gradient information to the central site and then obtain $\tilde{\boldsymbol{\beta}}^{[t]}$ by solving (3.8); Construct the multi-round communication-efficient debiased lasso estimator $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]} = (\tilde{\beta}_1^{\mathbf{d}[t]}, \cdots, \tilde{\beta}_p^{\mathbf{d}[t]})^\top$.

(ii) Apply the hard-thresholding procedure with $\nu = C_0\sqrt{\log p/N}$ and then finally output the SCDL estimator $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}[t]}) = (\mathcal{T}_\nu(\tilde{\beta}_1^{\mathbf{d}[t]}), \cdots, \mathcal{T}_\nu(\tilde{\beta}_p^{\mathbf{d}[t]}))^\top$.

(iii) Repeat the procedures (i)-(ii) using $\tilde{\mathcal{L}}_m(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}_m^{[t-1]})$ and obtain $\tilde{\boldsymbol{\beta}}_m^{\mathbf{d}[t]}$ for $m = 1, \ldots, M$; compute the aggregated estimator $\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[t]} = M^{-1}\sum_{m=1}^{M}\tilde{\boldsymbol{\beta}}_m^{\mathbf{d}[t]}$ and apply hard-thresholding procedure in (ii) successively to obtain the aggregated SCDL estimator.

S2 Additional simulation results

S2.1 Simulation results under $\chi_3^2/LogN(0,1)$ errors and $n = 200$

In this part, we show the simulation results in Section 4.2 for the other two errors: χ_3^2 and $LogN(0,1)$. The setting is the same as Section 4.2 except the error term. The simulation results when $(n, M) = (200, 5)$ for five errors are also provided. The results are shown in **Figures S1-S4**. The two columns in **Figures S1**, **S2** and **S4** correspond to the errors χ_3^2 and $LogN(0,1)$, respectively. The performance of two proposed estimators have similar performance in Section 4.2 and are comparable with the golden estimator $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$, which implies the proposed estimators are robust to various error distributions. As shown in **Figures S3 and S4**, the ℓ_∞ and ℓ_2 errors of all estimators increase when the dimension p increases. Compared with the simulation results when $n = 100$, both ℓ_∞ and ℓ_2 errors of all estimators decrease as the local sample size n increases.

S2.2 Effect of heteroskedastic error and outliers

Based on the similar simulation settings as in Sections 4.2 and 4.3, we generate the error from a mixture distribution, i.e., $\varepsilon_i = 0.5N(0,1) + 0.5t_3$, to evaluate the robustness of our proposed two distributed estimators to heteroskedasticity. The simulated ℓ_∞ and ℓ_2 errors are shown in **Figure S5** and we have the similar conclusions as in Sections 4.2 and 4.3. In addition, we consider the t_3 error but generate data based on the following three cases: (1) no outliers; (2) randomly choose 5% of response y_i to be $y_i + 5$; (3) randomly choose 5% of response y_i to be $y_i + 10$, to evaluate the robustness of our proposed estimators. The simulated results are shown in **Figures S6-S7**. The three columns correspond to the cases (1)-(3), respectively. (i)

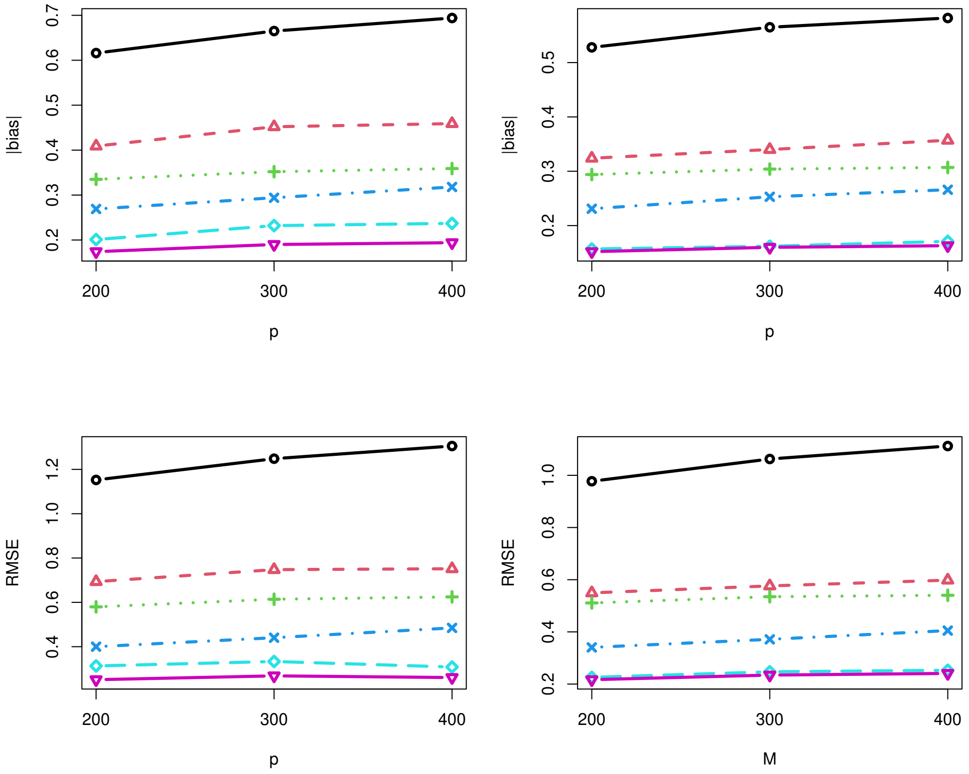

Figure S1: The ℓ_∞ and ℓ_2 errors for χ^2_3 and $LogN(0,1)$ with varying $p = 200$, 300, 400 when $(n, M) = (100, 5)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

Compared the results with no outliers, the ℓ_∞ and ℓ_2 errors of all estimators increases in the presence of the outliers, i.e., the cases (2)-(3), especially for $\bar{\boldsymbol{\beta}}$. Compared with the case (2), when the contamination level increases to case (3), the ℓ_∞ and ℓ_2 errors also increase, but the proposed two estimators have much smaller magnitudes of increase. (ii) The similar trends of the

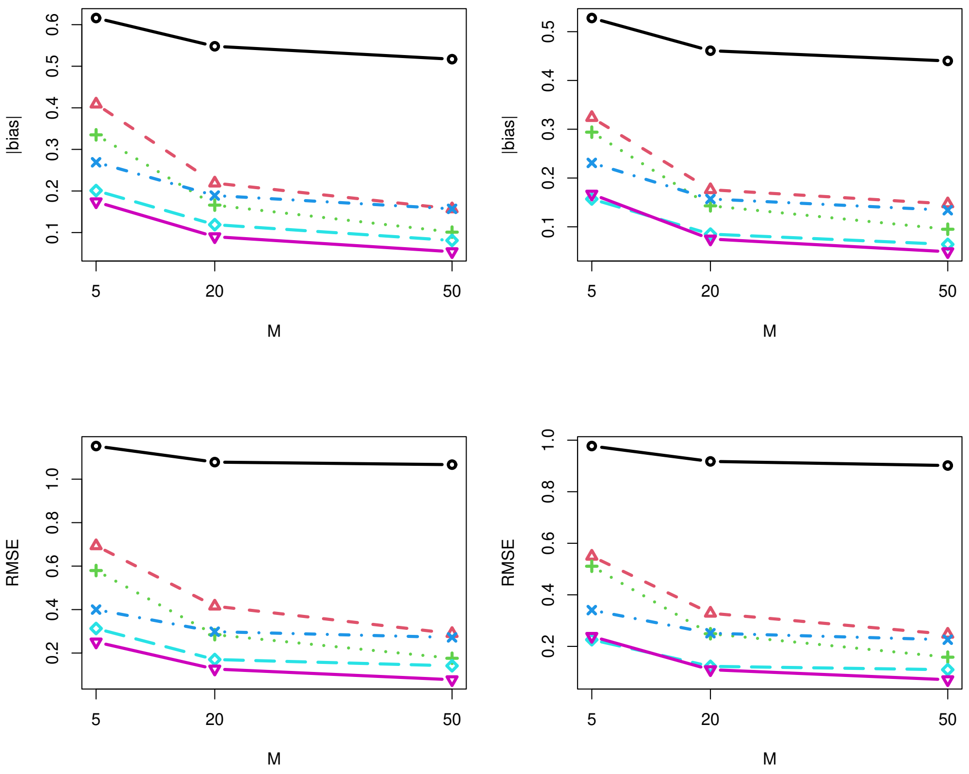

Figure S2: The ℓ_∞ and ℓ_2 errors for χ_3^2 and $LogN(0,1)$ with varying number of sites $M = 5, 20, 50$ when $(n,p) = (100, 200)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

ℓ_∞ and ℓ_2 errors as in Sections 4.2 and 4.3 still hold with respect to the varying p and M, respectively. Hence, it can be seen that the proposed two estimators $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ do not sacrifice much with the outliers and the performance of $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ is closer to the golden standard $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$.

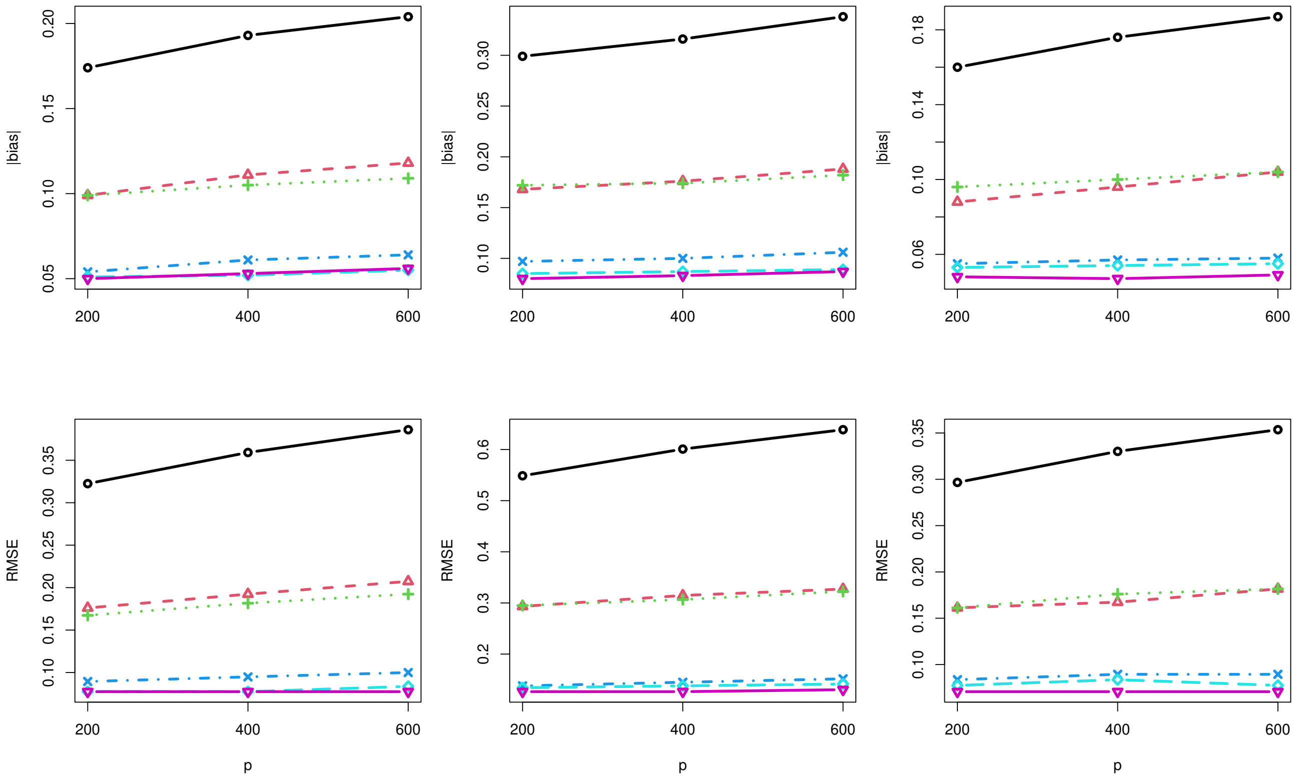

Figure S3: ℓ_∞ and ℓ_2 errors for $N(0,1)$, t_3 and $Pareto(2,4)$ with varying $p = 200$, 400, 600 when $(n, M) = (200, 5)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

S2.3 Computation time

For fixed $M = 5$ with the varying $p = 200$, 300, 400 as well as the fixed $p = 200$ with the varying $M = 5$, 20, 50, we consider the t_3 error with the local sample size $n = 100$ and compare the computation time of the different estimators over 200 replications, respectively. The results are shown

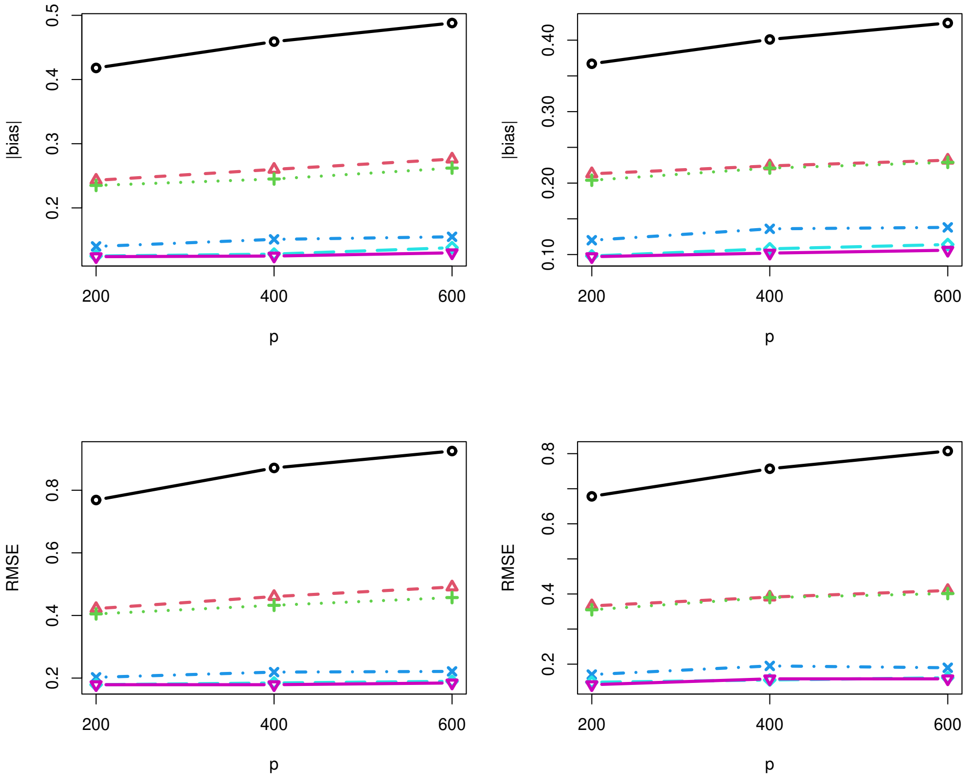

Figure S4: The ℓ_∞ and ℓ_2 errors for χ_3^2 and $LogN(0,1)$ with varying $p = 200, 400, 600$ when $(n, M) = (200, 5)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

in **Table S1**. (1) Under $M = 5$, for any fixed p, due to estimating the unknown projection vector $\boldsymbol{\gamma}_j^*$ in the debiasing procedure, the debiased and sparse estimators (**b**), (**d**) and (**f**) take much longer time than the estimators (**a**), (**c**) and (**e**). However, it can be seen that the computation time of the global estimators (**a**) and (**b**) are the largest ones, respectively. As p in-

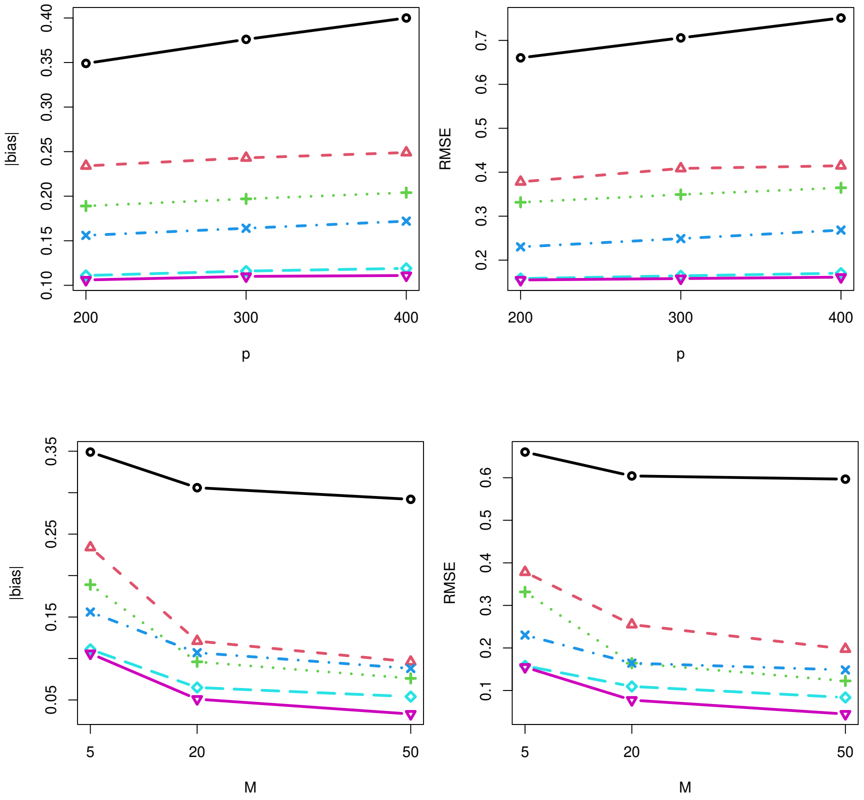

Figure S5: The ℓ_∞ and ℓ_2 errors with varying p and M for heteroskedastic error. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

creases, all estimators take longer time, especially for the global estimators $\hat{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$. (2) Under $p = 200$, when M increases, the computation time of the aggregated estimator $\bar{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ do not change significantly; the computation time of the communication-efficient estimators $\tilde{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ increases slightly due to the transmitting gradients procedure; the computation time of $\hat{\boldsymbol{\beta}}$ and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}})$ increases.

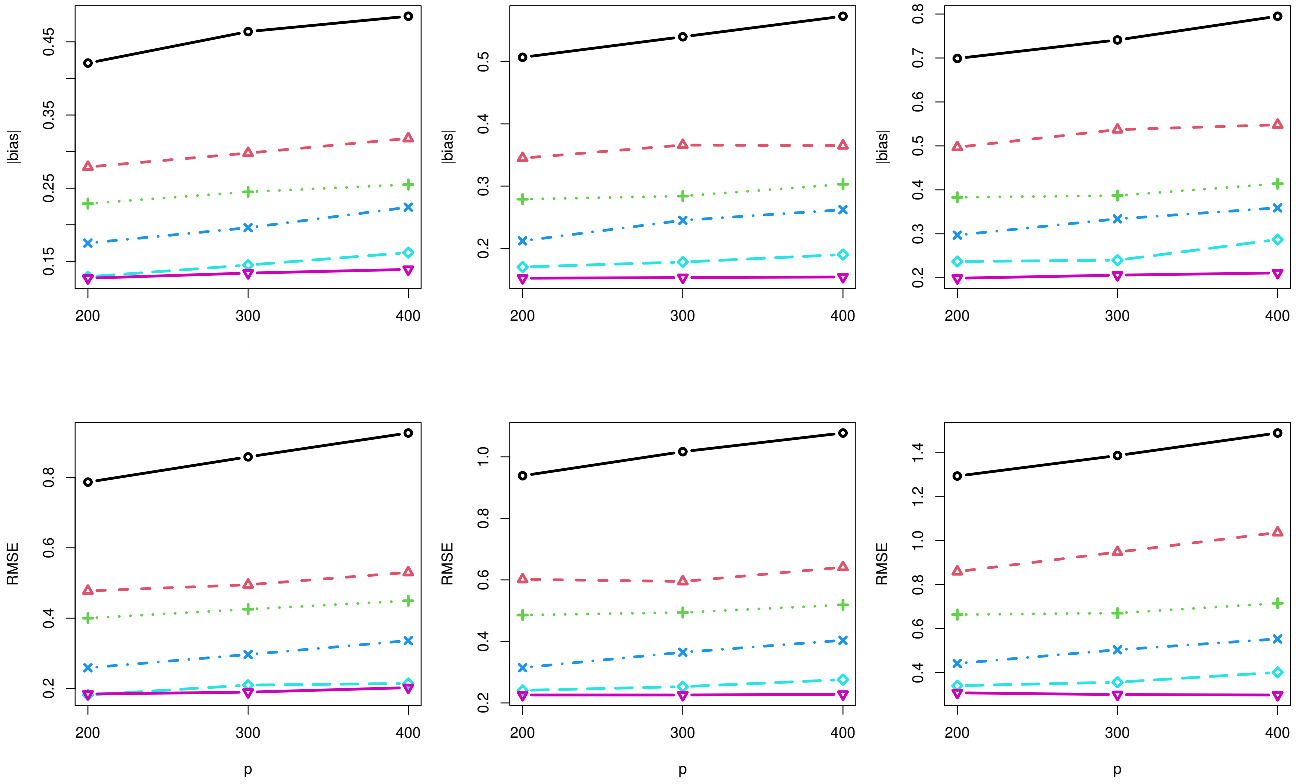

Figure S6: The ℓ_∞ and ℓ_2 errors for cases (1)-(3) with varying $p = 200, 300, 400$ when $(n, M) = (100, 5)$. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

S2.4 Variable selection

We consider the same simulation setting as in Section 4.4 based on 500 repetitions and the results are shown in the following Table S2. To implement our proposed methods, we apply five-fold cross-validation to choose C_0 of the hard-thresholding parameter $\nu = C_0\sqrt{\log p/N}$. The variable selection results are shown in **Table S2**. Columns “C” and “IC” are measures of

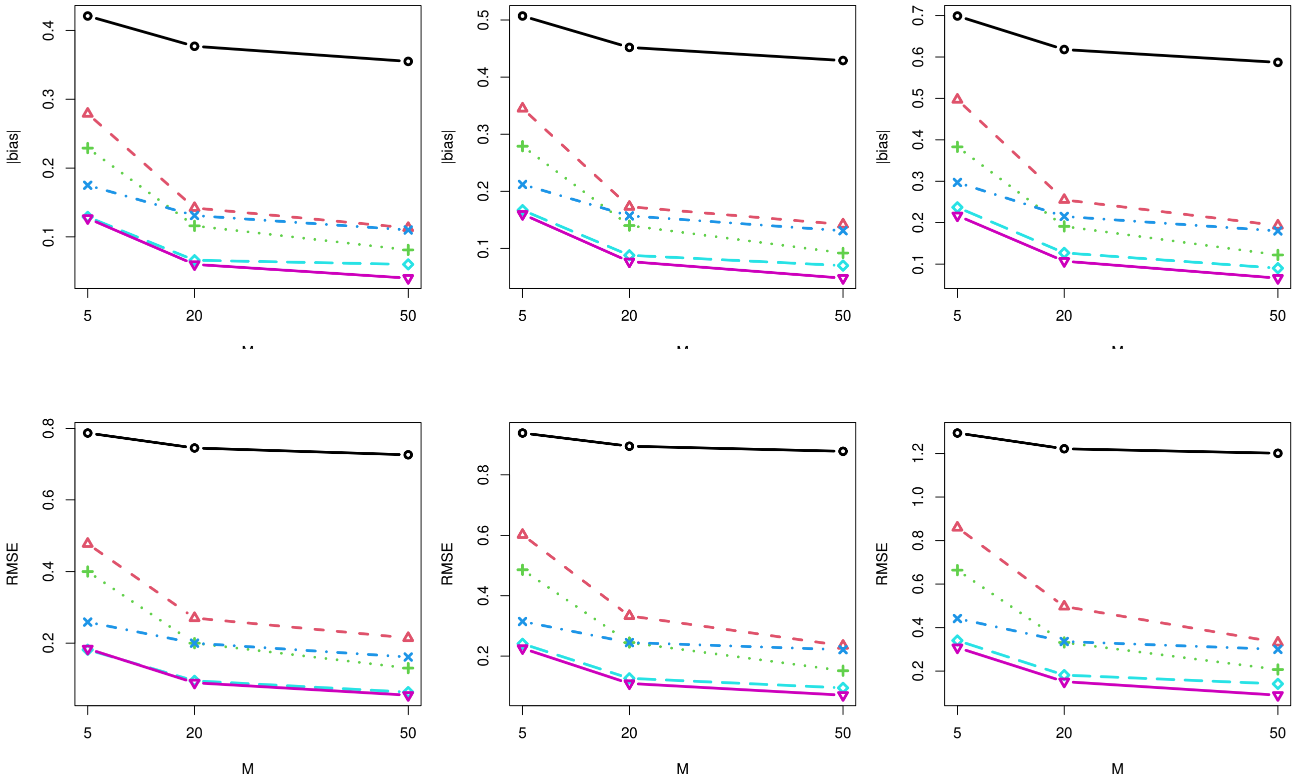

Figure S7: The ℓ_∞ and ℓ_2 errors for cases (1)-(3) with varying $M = 5,\ 20,\ 50$ when $(n,p) = (100, 200)$ with outliers. Here, $\bar{\boldsymbol{\beta}}$ (○), $\tilde{\boldsymbol{\beta}}$ (△), $\hat{\boldsymbol{\beta}}$(+), $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$ (×), $\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$ (◇) and $\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$ (▽).

model complexity, with “C” representing the average number of nonzero coefficients correctly estimated to be nonzero, and “IC” representing the average number of zero coefficients incorrectly estimated to be nonzero. From the variable selection results, we see that all methods can select all five true predictors in all settings. The average numbers of zero coefficient

Table S1: Average computation time (in seconds) comparison of different estimates with changing dimension p and changing number of sites M for t_3 error with local sample size $n = 100$.

	p			M		
Methods	200	300	400	5	20	50
$\bar{\boldsymbol{\beta}}$	0.385	0.727	1.205	0.385	0.386	0.341
$\tilde{\boldsymbol{\beta}}$	0.187	0.250	0.290	0.187	0.423	0.524
$\hat{\boldsymbol{\beta}}$	0.641	1.110	1.370	0.641	1.877	4.218
$\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$	12.013	30.336	57.371	12.013	11.767	11.527
$\mathcal{T}_\nu(\tilde{\boldsymbol{\beta}}^{\mathbf{d}})$	11.363	24.065	27.358	11.363	12.375	13.513
$\mathcal{T}_\nu(\hat{\boldsymbol{\beta}}^{\mathbf{d}})$	21.376	82.519	525.312	21.376	34.744	60.045

incorrectly estimated to be nonzero of our proposed methods are zero in most of the cases, which implies that the hard-thresholding method performs well.

Proof of Theorem 1.

For convenience, for any given $\tau > 0$, define

$$\boldsymbol{\beta}^*_\tau \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} E[\ell_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})],$$

thus $\boldsymbol{\beta}^*_\tau$ satisfies $E[\boldsymbol{x}_i \psi_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*_\tau)] = 0$. Based on the proof of Theorem 1 in Han et al. (2022), with $s^2 M \log p / N = o(1)$, by multiplying the regularization parameters λ_m used in Han et al. (2022) with an arbitrarily large

Table S2: Variable selection results of the sparse and debiased estimators with varying p and M.

	C	IC	C	IC	C	IC	C	IC	C	IC	C	IC
			$\boldsymbol{M}=5$						$\boldsymbol{p}=200$			
	$\boldsymbol{p}=200$		$\boldsymbol{p}=400$		$\boldsymbol{p}=600$		$\boldsymbol{M}=5$		$\boldsymbol{M}=10$		$\boldsymbol{M}=20$	
$\mathcal{T}_c(\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{ols})$	6	0.008	6	0	6	0.002	6	0.008	6	0	6	0
$\mathcal{T}_c(\tilde{\boldsymbol{\beta}}^{d}_{ols})$	6	0.006	6	0.002	6	0.014	6	0.006	6	0	6	0
$\mathcal{T}_c(\hat{\boldsymbol{\beta}}^{d}_{ols})$	6	0.002	6	0	6	0.004	6	0.002	6	0	6	0
$\mathcal{T}_c(\bar{\boldsymbol{\beta}}^{\mathbf{d}})$	6	0	6	0	6	0.006	6	0	6	0	6	0
$\mathcal{T}_c(\tilde{\boldsymbol{\beta}}^{d})$	6	0	6	0.004	6	0.002	6	0	6	0	6	0
$\mathcal{T}_c(\hat{\boldsymbol{\beta}}^{d})$	6	0.004	6	0	6	0	6	0.004	6	0	6	0

positive constant independent of m, we can show that with probability at least $1-(e+1)p^{-c}$, there exists a universal constant $C>0$ independent of m, such that

$$\|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}\|_2 \leq C\sqrt{sM\log p/N} \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}\|_1 \leq Cs\sqrt{M\log p/N},$$

where $c>0$ can be arbitrarily large by adjusting the constant C. Since the samples are independent and identically distributed among the M sites (homogeneous), taking $\lambda_m \asymp \sqrt{M\log p/N}$ uniformly in m and applying the union bound, we obtain that

$$\max_m \|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}\|_2 \leq C\sqrt{sM\log p/N} \text{ and } \max_m \|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}\|_1 \leq Cs\sqrt{M\log p/N},$$

hold with probability at least $1 - M(e+1)p^{-c}$. Under $\log M = O(\log p)$, $1 - M(e+1)p^{-c}$ can be written again as $1-(e+1)p^{-c}$ (with a different c). The similar argument can be found in Lian and Fan (2018). This technique will be used to derive uniform convergence rates in the following proof. Similar to the proof of Theorem 2 in Han et al. (2022), denote $\hat{S}_m = \{1 \le j \le p \mid \hat{\beta}_{m,j} \neq 0\}$ for the mth site. For any $j \in \hat{S}_m$, let

$$\hat{x}_j \in \begin{cases} \mathrm{sgn}[\hat{\beta}_{m,j}] & \text{if } \hat{\beta}_{m,j} \neq 0; \\ [-1,1] & \text{if } \hat{\beta}_{m,j} = 0. \end{cases}$$

From KKT conditions for $\hat{\boldsymbol{\beta}}_m$,

$$\frac{1}{n}\sum_{i\in\mathcal{I}_m} \boldsymbol{x}_{i,\hat{S}_m}[(\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m) - \psi_{\tau_n}(\varepsilon_{i,\tau_n})] = \lambda_m \hat{\boldsymbol{x}}_{\hat{S}_m} - \frac{1}{n}\sum_{i\in\mathcal{I}_m} \boldsymbol{x}_{i,\hat{S}_m}\psi_{\tau_n}(\varepsilon_{i,\tau_n}), \tag{S2.1}$$

where $\boldsymbol{x}_{i,\hat{S}_m} := \{x_{i,j} : j \in \hat{S}_m\}$, $\hat{\boldsymbol{x}}_{\hat{S}_m} := \{\hat{x}_j : j \in \hat{S}_m\}$ and $\varepsilon_{i,\tau_n} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^*_{\tau_n}$. It is easy to know that

$$|\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m) - \psi_{\tau_n}(\varepsilon_{i,\tau_n})| \le |\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n})|. \tag{S2.2}$$

With the bounded covariates assumption and (ii) in (C2), similar to the derivations of Lemma 6 in Belloni and Chernozhukov (2011), it can be proved that with high probability, $|\hat{S}_m| \le N/(M\log p)$. Combining Condition (C3) and (S2.2), it leads to

$$\begin{aligned}&\Big\|\frac{1}{n}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_{\hat{S}_m}[\psi_{\tau_n}(y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)-\psi_{\tau_n}(\varepsilon_{i,\tau_n})]\Big\|_2\\ \leq&\sup_{\mathrm{supp}(\boldsymbol{z})\subset\hat{S}_m,\|\boldsymbol{z}\|_2=1}\frac{1}{n}\sum_{i\in\mathcal{I}_m}|\boldsymbol{z}^\top\boldsymbol{x}_i||\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}^*_{\tau_n})|\leq C\sqrt{s}\lambda_m, \qquad \text{(S2.3)}\end{aligned}$$

where $\lambda_m \asymp \sqrt{M\log p/N}$ uniformly in m. Taking the ℓ_2-norm on both sides of (S2.1), we get

$$\Big\|\frac{1}{n}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_{i,\hat{S}_m}[\psi_{\tau_n}(y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)-\psi_{\tau_n}(\varepsilon_{i,\tau_n})]\Big\|_2=\Big\|\lambda_m\hat{\boldsymbol{x}}_{\hat{S}_m}-\frac{1}{n}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_{i,\hat{S}_m}\psi_{\tau_n}(\varepsilon_{i,\tau_n})\Big\|_2. \qquad \text{(S2.4)}$$

From (S2.3), (S2.4) and triangle inequality, we obtain with high probability, $|\hat{S}_m|\leq Cs$ uniformly in m.

Denote $\boldsymbol{\Theta}_j$ as the jth row of $\boldsymbol{\Theta}$ with $\boldsymbol{\Theta}$ be the inverse matrix of $\boldsymbol{\Sigma}=E\{\boldsymbol{x}_i\boldsymbol{x}_i^\top I(|\varepsilon_{i,\tau_n}|\leq\tau_n)\}$, where $\varepsilon_{i,\tau_n}=y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}^*_{\tau_n}$. From (iii) in (C1), $E[I(|\varepsilon_{i,\tau_n}|\leq\tau_n)]>0$. It is easy to verify that

$$\boldsymbol{\Theta}_j=\boldsymbol{\rho}_j/\{E[x_{i,j}(x_{i,j}-\boldsymbol{x}_{i,-j}^\top\boldsymbol{\gamma}_j^*)]E[I(|\varepsilon_{i,\tau_n}|\leq\tau_n)]\}.$$

Under Condition (C3) and noticing that ε_{i,τ_n} is independence of $\boldsymbol{x}_i$ (Proposition 5, Wang et al., 2021), we have $\max_j\|\boldsymbol{\Theta}_j\|_0\leq s_1$. Note that

$$\hat{\boldsymbol{\Theta}}_j^{(m)}=\hat{\boldsymbol{\rho}}_j^{(m)}/\Big\{n^{-2}\sum_{i\in\mathcal{I}_m}x_{i,j}(x_{i,j}-\boldsymbol{x}_{i,-j}^\top\hat{\boldsymbol{\gamma}}_j^{(m)})\sum_{i\in\mathcal{I}_m}I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\leq\tau_n)\Big\},$$

and

$$
\begin{aligned}
&\|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1 \\
\le\ & \underbrace{\Big\|\frac{\boldsymbol{\rho}_j}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]}\Big\|_1}_{A_j} \underbrace{\Big|\frac{1}{n^{-1}\sum_{i\in\mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m| \le \tau_n)} - \frac{1}{E[I(|\varepsilon_{i,\tau_n}| \le \tau_n)]}\Big|}_{B^{(m)}} \\
& + \underbrace{\Big\|\frac{\hat{\boldsymbol{\rho}}_j^{(m)}}{n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})} - \frac{\boldsymbol{\rho}_j}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]}\Big\|_1}_{C_j^{(m)}} \\
& \quad \times \underbrace{\Big|\frac{1}{n^{-1}\sum_{i\in\mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m| \le \tau_n)}\Big|}_{D^{(m)}},
\end{aligned}
$$

To proceed, we have to consider the uniform convergence rate of $\|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1$ both in j and m, i.e., $\max_m \max_j \|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1$, which is crucial for the proof.

By $\boldsymbol{\gamma}_j^* \equiv \operatorname{argmin}_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}} E(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j)^2$, we have $E[\boldsymbol{x}_{i,-j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)] = 0$. With $\lambda_{\min}\{E(\boldsymbol{x}_i\boldsymbol{x}_i^\top)\} > C_{\min} > 0$ in (C2) and denoting $\tau_j^2 = E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)] = E[(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)^2]$, we know $\tau_j^2 = 1/\Omega_{jj} \ge C_{\min} > 0$ stays away from zero and $\tau_j^2 \le \max_{1\le j\le p} \Xi_{jj} = O(1)$ uniformly in j. With $\max_j \|\boldsymbol{\Theta}_j\|_0 \le s_1$, it can be seen that $A_j \le C\sqrt{s_1}$ uniformly in j. The denominators of $B^{(m)}$ and $D^{(m)}$ are greater than 0 uniformly in m by $E[I(|\varepsilon_{i,\tau_n}| \le \tau_n)] > 0$.

For $C_j^{(m)}$, we prove that with probability at least $1 - 2p^{-c}$, there exists a universal constant $C > 0$ independent of j and m, such that

$$\max_m \max_j \Big\| \frac{\hat{\boldsymbol{\rho}}_j^{(m)}}{n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})} - \frac{\boldsymbol{\rho}_j}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]} \Big\|_1 \leq C s_1 \sqrt{\frac{M \log p}{N}}, \tag{S2.5}$$

where $c > 0$ can be arbitrarily large by adjusting the constant C. Since $2p^{-c}$ can be as small as possible, the left side of (S2.5) is bounded in probability at the rate $s_1\sqrt{M \log p/N}$ and can be further written as

$$\max_m \max_j \Big\| \frac{\hat{\boldsymbol{\rho}}_j^{(m)}}{n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})} - \frac{\boldsymbol{\rho}_j}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]} \Big\|_1 = O_p\Big(s_1 \sqrt{\frac{M \log p}{N}}\Big).$$

For the mth local site and $1 \leq j \leq p$, noting $\hat{\boldsymbol{\rho}}_j^{(m)} = (-\hat{\gamma}_{j,1}^{(m)}, \ldots, -\hat{\gamma}_{j,(j-1)}^{(m)}, 1, -\hat{\gamma}_{j,j}^{(m)}, \ldots, -\hat{\gamma}_{j,(p-1)}^{(m)})$, we can decompose it as follows

$$\begin{aligned}
&\Big\| \frac{\hat{\boldsymbol{\rho}}_j^{(m)}}{n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})} - \frac{\boldsymbol{\rho}_j}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]} \Big\|_1 \qquad \text{(S2.6)}\\
\leq{}& \|\boldsymbol{\gamma}_j^*\|_1 \Big| \frac{1}{n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})} - \frac{1}{E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]} \Big| \\
&+ \frac{\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1}{|n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})|} \\
\leq{}& \|\boldsymbol{\gamma}_j^*\|_1 |1/\hat{\tau}_{m,j}^2 - 1/\tau_j^2| + \|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j\|_1/\hat{\tau}_{m,j}^2,
\end{aligned}$$

where $\tau_j^2 = E[x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)]$ and $\hat{\tau}_{m,j}^2 = n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)})$. The uniform convergence rate of (S2.6) in j and m depends on the uniform convergence rates of $\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j\|_1$, $1/\hat{\tau}_{m,j}^2$ and $|1/\hat{\tau}_{m,j}^2 - 1/\tau_j^2|$ in j and m.

Note $\hat{\boldsymbol{\gamma}}_j^{(m)} \in \operatorname*{argmin}_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}} \{h_{m,j}(\boldsymbol{\gamma}_j) + \omega_{jm}\|\boldsymbol{\gamma}_j\|_1\}$ with $h_{m,j}(\boldsymbol{\gamma}_j) = \sum_{i\in\mathcal{I}_m}(x_{i,j} -$

$\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j)^2/(2n)$. By the definition of $\hat{\boldsymbol{\gamma}}_j^{(m)}$,

$$h_{m,j}(\hat{\boldsymbol{\gamma}}_j^{(m)}) + \omega_{jm}\|\hat{\boldsymbol{\gamma}}_j^{(m)}\|_1 \le h_{m,j}(\boldsymbol{\gamma}_j^*) + \omega_{jm}\|\boldsymbol{\gamma}_j^*\|_1, \tag{S2.7}$$

where $\boldsymbol{\gamma}_j^* \equiv \text{argmin}_{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}} E(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j)^2$. By the convexity of $h_{m,j}(\boldsymbol{\gamma}_j)$,

$$h_{m,j}(\hat{\boldsymbol{\gamma}}_j^{(m)}) \ge h_{m,j}(\boldsymbol{\gamma}_j^*) + \nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*). \tag{S2.8}$$

In the event $\{\|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty \le \omega_{jm}/2\}$, we have

$$|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)| \le \frac{\omega_{jm}}{2}\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1. \tag{S2.9}$$

Combining (S2.7)-(S2.9), we obtain

$$\|\boldsymbol{\gamma}_j^*\|_1 \ge -\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1/2 + \|\hat{\boldsymbol{\gamma}}_j^{(m)}\|_1. \tag{S2.10}$$

Let $\mathcal{S}_j$ be the set of indices of nonzero components of $\boldsymbol{\gamma}_j^*$ and $\mathcal{S}_j^c$ be the complement of $\mathcal{S}_j$. Then by triangle inequality, we have

$$\begin{aligned}\|\hat{\boldsymbol{\gamma}}_j^{(m)}\|_1 &= \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^* + \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1 + \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^* + \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j^c}\|_1 \\ &\ge \|\boldsymbol{\gamma}_j^*\|_1 - \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1 + \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j^c}\|_1,\end{aligned}$$

which leads to

$$\|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j^c}\|_1 \le 3\|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1, \tag{S2.11}$$

by (S2.10). Under $\lambda_{\min}\{E(\boldsymbol{x}_i\boldsymbol{x}_i^\top)\} \ge C_{\min} > 0$ and $s_1^2 M \log p/N = o(1)$, there exists a universal constant $C_{\lambda_{\min}} > 0$, such that

$$\begin{aligned} C_{\lambda_{\min}}\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2^2 &\leq h_{m,j}(\hat{\boldsymbol{\gamma}}_j^{(m)}) - h_{m,j}(\boldsymbol{\gamma}_j^*) - \nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*) \\ &\leq \omega_{jm}\|\boldsymbol{\gamma}_j^*\|_1 - \omega_{jm}\|\hat{\boldsymbol{\gamma}}_j^{(m)}\|_1 + \frac{\omega_{jm}}{2}\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1 \\ &\leq \frac{3\omega_{jm}}{2}\|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1 - \frac{\omega_{jm}}{2}\|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j^c}\|_1. \end{aligned}$$

Since $\|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1 \leq s_1^{1/2}\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2$ holds by Cauchy-Schwarz inequality, we conclude that $\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2 \leq 3s_1^{1/2}\omega_{jm}/(2C_{\lambda_{\min}})$ and by (S2.11)

$$\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1 = \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j}\|_1 + \|(\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*)_{\mathcal{S}_j^c}\|_1 \leq 6s_1\omega_{jm}/C_{\lambda_{\min}}.$$

Therefore, in the event $\{\|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty \leq \omega_{jm}/2\}$, we have

$$\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1 \leq 6s_1\omega_{jm}/C_{\lambda_{\min}} \text{ and } \|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2 \leq 3s_1^{1/2}\omega_{jm}/(2C_{\lambda_{\min}}).$$

Now we show that with probability at least $1-2p^{-c}$, there exists a universal constant $C > 0$ independent of j and m, such that

$$\|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty \leq C\sqrt{M\log p/N}, \tag{S2.12}$$

where $c > 0$ can be arbitrarily large by adjusting the constant C. To prove (S2.12), we firstly calculate

$$P(|\{\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\}_l| > \omega_{jm}/2),$$

where $\{\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\}_l$ is the lth component of $\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)$ and then apply the union bound in l. This trick is also used in the proof of Lemma 23 in Javanmard and Montanari (2014).

By Theorem 2.10 in Boucheron et al. (2013), we apply the Bernstein's inequality to

$$\{x_{i,l}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)\}_{i\in\mathcal{I}_m} \quad \text{for} \quad 1 \le j \le p \quad \text{and} \quad 1 \le l \ne j \le p.$$

Note τ_j^2 stays away from zero and $\tau_j^2 \le \xi \equiv \max_{1\le j\le p} \Xi_{jj} = O(1)$ uniformly in j. With the bounded covariate assumption $\max_{i,j}|x_{i,j}| \le B$, we have

$$\sum_{i\in\mathcal{I}_m} E[x_{i,l}^2(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)^2] \le nB^2 E[(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)^2] \le nB^2\xi.$$

In addition, with $\max_{i,j}|\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*| \le B$ in (C3),

$$\max_{i,j,l}|x_{i,l}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*)| \le \max_{i,l}|x_{i,l}| \max_{i,j}|x_{i,j}| + \max_{i,l}|x_{i,l}| \max_{i,j}|\boldsymbol{x}_{i,-j}^\top \boldsymbol{\gamma}_j^*| \le 2B^2.$$

Taking $\omega_{jm}/2 = 2\sqrt{2B^2\xi(c+1)M\log p/N}$ with an arbitrarily large constant $c > 0$ independent of j and m, such that

$$\begin{aligned}
&P(|\{\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\}_l| > \omega_{jm}/2) \\
\le& P\Big(|n^{-1}\sum_{i\in\mathcal{I}_m} x_{i,l}(x_{i,j} - \boldsymbol{x}_{i,-j}\boldsymbol{\gamma}_j^*)| > \sqrt{2B^2\xi(c+1)\log p/n} + 2B^2(c+1)\log p/n\Big) \\
=& P\Big(|\sum_{i\in\mathcal{I}_m} x_{i,l}(x_{i,j} - \boldsymbol{x}_{i,-j}\boldsymbol{\gamma}_j^*)| > \sqrt{2nB^2\xi(c+1)\log p} + 2B^2(c+1)\log p\Big) \\
\le& 2\exp(-(c+1)\log p) = 2p^{-(c+1)}.
\end{aligned}$$

The first inequality is due to that $\sqrt{\log p/n}$ is the leading term when $M\log p/N = o(1)$ and the last inequality is obtained by applying the Bernstein's inequality with $v = nB^2\xi$, $b = 2B^2$ and $t = (c+1)\log p$. Taking a union bound over the $(p-1)$ components of $\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)$ with

$\omega_{jm}/2 = 2\sqrt{2B^2\xi(c+1)M\log p/N}$, we get

$$P(\|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty > \omega_{jm}/2) \leq 2p^{-(c+1)}p = 2p^{-c}.$$

Hence, (S2.12) is proved.

For the uniform convergence rates of $\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1$ and $\|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2$ in j and m, note the samples are independent and identically distributed and the M local sites are homogeneous. From the above discussion, we can take ω_{jm} uniformly in j and m, i.e.,

$$\omega_{jm}/2 = C\sqrt{M\log p/N} \;\; \text{with} \;\; C = 2\sqrt{2B^2\xi(c+1)},$$

such that

$$P(\|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty \leq C\sqrt{M\log p/N}) \geq 1 - 2p^{-c},$$

holds for $1 \leq j \leq p$ and $1 \leq m \leq M$ simultaneously. Then by the union bound both in j and m, we have

$$P(\max_m \max_j \|\nabla_{\boldsymbol{\gamma}_j} h_{m,j}(\boldsymbol{\gamma}_j^*)\|_\infty \leq C\sqrt{M\log p/N}) \geq 1 - 2pMp^{-c}. \quad \text{(S2.13)}$$

Since $c > 0$ is an arbitrarily large constant, with $\log M = O(\log p)$, $1 - 2pMp^{-c}$ can be written again as $1 - 2p^{-c}$ (with a different c). Noting (S2.13) and ω_{jm} is independent of j and m, we know that

$$\max_m \max_j \|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_1 \leq Cs_1\sqrt{M\log p/N} \text{ and } \max_m \max_j \|\hat{\boldsymbol{\gamma}}_j^{(m)} - \boldsymbol{\gamma}_j^*\|_2 \leq C\sqrt{s_1 M\log p/N},$$

hold with probability at least $1-2p^{-c}$ for an another constant C independent of m and j.

For the uniform convergence rates of $|1/\hat{\tau}^2_{m,j} - 1/\tau^2_j|$ and $1/\hat{\tau}^2_{m,j}$ in j and m, note $\|\hat{\boldsymbol{\gamma}}^{(m)}_j\|_1 \leq \|\boldsymbol{\gamma}^*_j\|_1 + \|\hat{\boldsymbol{\gamma}}^{(m)}_j - \boldsymbol{\gamma}^*_j\|_1$,

$$\begin{aligned}
\hat{\tau}^2_{m,j} =& n^{-1} \sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j) \\
=& n^{-1} \sum_{i\in\mathcal{I}_m} (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j)^2 + n^{-1} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j) \\
=& n^{-1} \sum_{i\in\mathcal{I}_m} (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j)^2 + n^{-1} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\boldsymbol{\gamma}^*_j) \\
& + n^{-1} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j \boldsymbol{x}^\top_{i,-j}(\boldsymbol{\gamma}^*_j - \hat{\boldsymbol{\gamma}}^{(m)}_j),
\end{aligned}$$

and

$$\begin{aligned}
|\hat{\tau}^2_{m,j} - \tau^2_j| \leq & \left| n^{-1} \sum_{i\in\mathcal{I}_m} (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j)^2 - \tau^2_j \right| + \left| n^{-1} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j (x_{i,j} - \boldsymbol{x}^\top_{i,-j}\boldsymbol{\gamma}^*_j) \right| \\
& + \left| n^{-1} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}^\top_{i,-j}\hat{\boldsymbol{\gamma}}^{(m)}_j \boldsymbol{x}^\top_{i,-j}(\boldsymbol{\gamma}^*_j - \hat{\boldsymbol{\gamma}}^{(m)}_j) \right|.
\end{aligned}$$

Note τ^2_j stays away from zero and bounded uniformly in j. Along the lines in the proof of Lemma 5.3 and Theorem 2.4 in van de Geer et al. (2014), these three terms of decomposition can be tackled. The uniform convergence rates of $|\hat{\tau}^2_{m,j} - \tau^2_j|$ and $1/\hat{\tau}^2_{m,j}$ in j and m come from the uniform convergence rates of $\|\hat{\boldsymbol{\gamma}}^{(m)}_j - \boldsymbol{\gamma}^*_j\|_1$ and $\|\hat{\boldsymbol{\gamma}}^{(m)}_j - \boldsymbol{\gamma}^*_j\|_2$ in j and m. By (S2.6) and the above discussion, (S2.5) is proved.

For $B^{(m)}$, it can be seen that

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\\
=&\underbrace{\left\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)-E[I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)]\right\}}_{I_1^{(m)}}+\underbrace{E[I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]}_{I_2^{(m)}}\\
=&\underbrace{\left\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)-\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|\varepsilon_{i,\tau_n}|\le\tau_n)-\big[E[I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\big]\right\}}_{I_{1,1}^{(m)}}\\
&+\underbrace{\left\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|\varepsilon_{i,\tau_n}|\le\tau_n)-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\right\}}_{I_{1,2}^{(m)}}+\underbrace{\left\{E[I(|y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\right\}}_{I_2^{(m)}}.
\end{aligned}$$

According to the equation (52) in Han et al. (2022), it can be proved that $I_{1,1}^{(m)}=o_p(\sqrt{M/N})$ and $I_{1,2}^{(m)}=O_p(\sqrt{M/N})$. Subsequently, $I_2^{(m)}=O_p(\sqrt{sM\log p/N})$ can be obtained by the fact that $\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{\tau_n}^*\|_2\le C\sqrt{sM\log p/N}$ holds with probability at least $1-(e+1)p^{-c}$ on the mth site. The expression $o_p(1)$ denotes a sequence that converges in probability to zero. To proceed, we still need to consider the uniform convergence rates of $I_{1,1}^{(m)}$, $I_{1,2}^{(m)}$ and $I_2^{(m)}$ in m.

Among $I_{1,1}^{(m)}$, $I_{1,2}^{(m)}$ and $I_2^{(m)}$, it can be verified that $I_2^{(m)}$ is the leading term such that we only need to discuss the uniform convergence rate of $I_2^{(m)}$ in m. From the proof of Han et al. (2022), the unifrom convergence rate of $I_2^{(m)}$ is determined by the unifrom convergence rate of $\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{\tau_n}^*\|_2$ in m, i.e., $\max_m\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{\tau_n}^*\|_2$. As we discussed at the beginning of the proof, $\max_m\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{\tau_n}^*\|_2\le C\sqrt{sM\log p/N}$ and $\max_m\|\hat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_{\tau_n}^*\|_1\le$

$Cs\sqrt{M\log p/N}$ hold with probability at least $1-(e+1)p^{-c}$ when $\log M = O(\log p)$. Based on the above results, the uniform convergence rates of $I_{1,1}^{(m)}$, $I_{1,2}^{(m)}$, and $I_2^{(m)}$ in m are controlled by $\max_m I_2^{(m)} = O_p(\sqrt{sM\log p/N})$ and the uniform convergence rate of $B^{(m)}$ in m is obtained.

Based on the above discussion about A_j, $B^{(m)}$, $C_j^{(m)}$ and $D^{(m)}$, we have with probability at least $1-2(e+1)p^{-c}$, there exists a universal constant $C>0$ independent of j and m, such that

$$\max_m \max_j \|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1 \le C(s_1 \vee \sqrt{ss_1})\sqrt{\frac{M\log p}{N}}, \tag{S2.14}$$

and (S2.14) can be further written as

$$\max_m \max_j \|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1 = O_p\Big((s_1 \vee \sqrt{ss_1})\sqrt{\frac{M\log p}{N}}\Big), \tag{S2.15}$$

for simplicity. In the following text, the notation O_p can be interpreted similarly.

From the construction of $\bar{\boldsymbol{\beta}}^{\mathbf{d}}$, we can decompose $\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*_{\tau_n}$ as follows:

$$\begin{aligned}\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*_{\tau_n} &= \frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{\beta}}^{\mathbf{d}}_m - \boldsymbol{\beta}^*_{\tau_n})\\ &= \frac{1}{M}\sum_{m=1}^{M}[(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}) + \frac{1}{n}\hat{\boldsymbol{\Theta}}^{(m)}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)],\end{aligned}$$

where $\hat{\mathbf{\Theta}}^{(m)} = \{\hat{\mathbf{\Theta}}_j^{(m)} : 1 \le j \le p\}$ and note

$$\begin{aligned}\hat{\boldsymbol\beta}^{\mathbf{d}}_m - \boldsymbol\beta^*_{\tau_n} &= \hat{\boldsymbol\beta}_m - \boldsymbol\beta^*_{\tau_n} + \frac{1}{n}\hat{\mathbf{\Theta}}^{(m)} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m)\\ &= \frac{1}{n}\hat{\mathbf{\Theta}}^{(m)} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n}) + \frac{1}{n}(\hat{\mathbf{\Theta}}^{(m)} - \mathbf{\Theta}) \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) + \hat{\boldsymbol\beta}_m - \boldsymbol\beta^*_{\tau_n}\\ &\quad + \frac{1}{n}\mathbf{\Theta} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i[\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) - \psi_{\tau_n}(\varepsilon_{i,\tau_n})] + \frac{1}{n}(\mathbf{\Theta} - \hat{\mathbf{\Theta}}^{(m)}) \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n}).\end{aligned}$$

It can be seen that

$$\begin{aligned}\bar{\boldsymbol\beta}^{\mathbf{d}} - \boldsymbol\beta^*_{\tau_n} &= \big[\frac{1}{M}\sum_{m=1}^M \frac{1}{n}\hat{\mathbf{\Theta}}^{(m)} \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})\big] + \big[\frac{1}{M}\sum_{m=1}^M \frac{1}{n}(\hat{\mathbf{\Theta}}^{(m)} - \mathbf{\Theta}) \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m)\big]\\ &\quad + \frac{1}{M}\sum_{m=1}^M (\hat{\boldsymbol\beta}_m - \boldsymbol\beta^*_{\tau_n}) + \big\{\frac{1}{M}\sum_{m=1}^M \frac{1}{n}\mathbf{\Theta} \sum_{i\in\mathcal{I}_m} [\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})]\big\}\\ &\quad + \big[\frac{1}{M}\sum_{m=1}^M \frac{1}{n}(\mathbf{\Theta} - \hat{\mathbf{\Theta}}^{(m)}) \sum_{i\in\mathcal{I}_m} \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})\big]\\ &= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5\\ &= \Delta_1 + \frac{1}{M}\sum_{m=1}^M [\Delta_2^{(m)} + \Delta_3^{(m)} + \Delta_4^{(m)} + \Delta_5^{(m)}].\end{aligned}$$

Moreover, denote $\Delta_4^{(m)} = \Delta_{4,1} + \Delta_{4,2}^{(m)}$, where $\Delta_{4,1} = \mathbf{\Theta} E[\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})]$ and

$$\Delta_{4,2}^{(m)} = \frac{1}{n}\mathbf{\Theta} \sum_{i\in\mathcal{I}_m} [\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})] - \mathbf{\Theta} E[\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol\beta}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})].$$

Next, we will derive the bounds of Δ_1, Δ_2, $\Delta_3 + \Delta_{4,1}$, $\Delta_{4,2} = M^{-1}\sum_{m=1}^M \Delta_{4,2}^{(m)}$ and Δ_5, respectively. Under Conditions (C2), (C3) and Cauchy-Schwarz inequality,

$$\max_{1\le j\le p}\big\|\boldsymbol{\Theta}_j\big\|_1 \le \max_{1\le j\le p}\sqrt{s_1}\big\|\boldsymbol{\Theta}_j\big\|_2 \le C\sqrt{s_1}. \tag{S2.16}$$

For $\Delta_2^{(m)}$, by KKT conditions and $\log M = O(\log p)$, noting $\lambda_m \asymp \sqrt{M\log p/N}$ uniformly in m and (S2.15), we have

$$\begin{aligned}\max_m \|\Delta_2^{(m)}\|_\infty =& \max_m \Big\|(\hat{\boldsymbol{\Theta}}^{(m)} - \boldsymbol{\Theta})\frac{1}{n}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)\Big\|_\infty\\ \le& \max_{j,m}\|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1\Big\|\frac{1}{n}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)\Big\|_\infty\\ \le& \max_{j,m}\|\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1\ \lambda_m\\ =& O_p\Big(\frac{(s_1\vee\sqrt{ss_1})M\log p}{N}\Big). \end{aligned}\tag{S2.17}$$

By Conditions (C1)-(C2) and equation (47) in the Supplement of Han et al. (2022), noting the uniform convergence of $\hat{\boldsymbol{\beta}}_m$ discussed at the beginning of the proof, we have

$$\max_m\|\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}) + E[\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m) - \boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})]\|_\infty = O_p\Big(\frac{sM\log p}{N}\Big). \tag{S2.18}$$

Thus for $\Delta_3^{(m)} + \Delta_{4,1}$, by Hölder's inequality, (S2.16) and (S2.18), we obtain

$$\begin{aligned}
&\max_m \|\Delta_3^{(m)} + \Delta_{4,1}\|_\infty \\
=& \max_m \|\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n} + \boldsymbol{\Theta} E[\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})]\|_\infty \\
=& \max_m \|\boldsymbol{\Theta}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}) + \boldsymbol{\Theta} E[\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})]\|_\infty \\
\le& \max_{j,m} \|\boldsymbol{\Theta}_j\|_1 \|\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*_{\tau_n}) + E[\boldsymbol{x}_i \psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m) - \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})]\|_\infty \\
=& O_p\Big(\frac{s\sqrt{s_1}M\log p}{N}\Big). \qquad \text{(S2.19)}
\end{aligned}$$

For $\Delta_{4,2} = M^{-1}\sum_{m=1}^M \Delta_{4,2}^{(m)}$, noting the uniform convergence of $\hat{\boldsymbol{\beta}}_m$, applying the bounds (37) and (45) in Han et al. (2022), it leads to

$$\begin{aligned}
\max_m \|\Delta_{4,2}^{(m)}\|_\infty =& O_p\left(\sqrt{\frac{sM\log p}{N}} + \sqrt{s_1^{1/2}s^{1/2}(N/M)^{-1/2}\sqrt{\log p}}\right) O_p\left(\sqrt{\frac{ss_1M\log p}{N}}\right) \\
=& O_p\left(\frac{(s\vee s_1)M\log p}{N}\right). \qquad \text{(S2.20)}
\end{aligned}$$

Applying (S2.14) and Hölder's inequality, we can show that $\Delta_5^{(m)}$ also holds with the same bound as $\Delta_2^{(m)}$. Combining (S2.17), (S2.19), (S2.20) and $s_1 \asymp s$ in Condition (C3), we get

$$\begin{aligned}
&\|\Delta_2 + \Delta_3 + \Delta_{4,1} + \Delta_{4,2} + \Delta_5\|_\infty = \Big\|\frac{1}{M}\sum_{m=1}^M [\Delta_2^{(m)} + \Delta_3^{(m)} + \Delta_4^{(m)} + \Delta_5^{(m)}]\Big\|_\infty \\
\le& \max_m \|\Delta_2^{(m)} + \Delta_3^{(m)} + \Delta_{4,1} + \Delta_{4,2}^{(m)} + \Delta_5^{(m)}\|_\infty = O_p\left(\frac{s^{3/2}M\log p}{N}\right).
\end{aligned}$$

For Δ_1, note that (S2.15) and $E[\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})] = 0$. Applying Hoeffding's inequality, we get

$$\begin{aligned}\|\Delta_1\|_\infty &= \Big\|\frac{1}{N}\sum_{m=1}^{M}\hat{\mathbf{\Theta}}^{(m)}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})\Big\|_\infty \\ &\le \Big\|\frac{1}{N}\sum_{m=1}^{M}(\hat{\mathbf{\Theta}}^{(m)}-\mathbf{\Theta})\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})\Big\|_\infty + \Big\|\frac{1}{N}\sum_{m=1}^{M}\mathbf{\Theta}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})\Big\|_\infty \\ &= O_p\Big(\sqrt{\frac{\log p}{N}}\Big).\end{aligned}$$

Under Conditions (i), (iii) in (C1) and Condition (ii) in (C2), along the lines in the proof of the Proposition 5 in Wang et al. (2021), the slope parts of $\boldsymbol{\beta}^*_\tau$ and $\boldsymbol{\beta}^*$ are the same but the intercept terms have a constant difference depending on τ, i.e., $\boldsymbol{\beta}^*_\tau = (\beta^*_1 + \alpha_\tau, \boldsymbol{\beta}^{*\top}_{-1})^\top$. Moreover, α_τ with $\tau > \sigma$ satisfies the bound

$$|\alpha_\tau| \le \frac{\sigma^2 - E[\psi^2_\tau(\varepsilon)]}{1-\tau^{-2}\sigma^2}\frac{1}{\tau}. \tag{S2.21}$$

From the above discussion, with $\tau_n \asymp \sigma\sqrt{N/(M\log p)}$ and (S2.21), we have

$$\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{-1} - \boldsymbol{\beta}^*_{-1}\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big), |\bar{\beta}^{\mathbf{d}}_1 - \beta^*_1| = O_p\Big(\sqrt{\frac{M\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big). \tag{S2.22}$$

However, $\sqrt{M\log p/N}$ in $|\bar{\beta}^{\mathbf{d}}_1 - \beta^*_1|$ is slower than the optimal convergence rate. We refer to the following result to control $|\alpha_{\tau_n}|$ better.

Proposition B.1 (Sun et al., 2020) *Assume that* $E(\varepsilon) = 0, \sigma^2 = E(\varepsilon^2) > 0$ *and* $E(|\varepsilon|^{2+\kappa}) < \infty$ *from some* $\kappa \ge 0$. *Then we have*

$$|E\psi_\tau(\varepsilon)| \le \min\{\tau^{-1}\sigma^2, \tau^{-1-\kappa}E(|\varepsilon|^{2+\kappa})\}.$$

Moreover, if $\kappa > 0$*,*

$$\sigma^2 - 2\kappa^{-1}\tau^{-\kappa}E(|\varepsilon|^{2+\kappa}) \le E\{\psi_\tau^2(\varepsilon)\} \le \sigma^2. \tag{S2.23}$$

If $\kappa > 0$, from (S2.21) and (S2.23), we get

$$|\alpha_\tau| \le \frac{\sigma^2 - E\{\psi_\tau^2(\varepsilon)\}}{1-\tau^{-2}\sigma^2}\frac{1}{\tau} \le \frac{2\kappa^{-1}\tau^{-(\kappa+1)}E\left(|\varepsilon|^{2+\kappa}\right)}{1-\tau^{-2}\sigma^2}.$$

Thus with condition $E\left(|\varepsilon|^3\right) < \infty$, we get

$$|\alpha_{\tau_n}| \le \frac{2{\tau_n}^{-2}E\left(|\varepsilon|^3\right)}{1-{\tau_n}^{-2}\sigma^2} \asymp \frac{M\log p}{N}. \tag{S2.24}$$

Finally, combining (S2.22) and (S2.24), we get

$$\|\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}} + \frac{s^{3/2}M\log p}{N}\Big).$$

Proofs of Theorem 2 and Corollary 1.

The strategy of proving this part is similar to the proof of Theorem 1. For $1 \le j \le p$, decompose the jth component $\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*_{\tau_n}$ as

$$\begin{aligned}
\bar{\beta}^{\mathbf{d}}_j - \beta^*_{\tau_n,j} =& \big[\frac{1}{M}\sum_{m=1}^M \frac{1}{n}\hat{\boldsymbol{\Theta}}_j^{(m)}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})\big] + \big[\frac{1}{M}\sum_{m=1}^M\frac{1}{n}(\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j)\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m)\big] \\
&+ \big[\frac{1}{M}\sum_{m=1}^M(\hat{\beta}_{m,j} - \beta_j^*)\big] + \big\{\frac{1}{M}\sum_{m=1}^M\frac{1}{n}\boldsymbol{\Theta}_j\sum_{i\in\mathcal{I}_m}[\boldsymbol{x}_i\psi_{\tau_n}(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_m) - \boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})]\big\} \\
&+ \big[\frac{1}{M}\sum_{m=1}^M\frac{1}{n}(\boldsymbol{\Theta}_j - \hat{\boldsymbol{\Theta}}_j^{(m)})\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})\big] \\
=&\Omega_1 + \Omega_2 + \Omega_3 + \Omega_4 + \Omega_5 \\
=&\Omega_1 + \frac{1}{M}\sum_{m=1}^M[\Omega_2^{(m)} + \Omega_3^{(m)} + \Omega_4^{(m)} + \Omega_5^{(m)}].
\end{aligned}$$

Similar to the decomposition in the proof of Theorem 1, $\Omega_4^{(m)} = \Omega_{4,1} + \Omega_{4,2}^{(m)}$. With $M = o(\sqrt{N}/(s^{3/2}\log p))$, from the proof of Theorem 1, we have Ω_2, $\Omega_3 + \Omega_{4,1}$, $\Omega_{4,2} = M^{-1}\sum_{m=1}^{M}\Omega_{4,2}^{(m)}$ and Ω_5 equal $o_p(N^{-1/2})$. Moreover, with the conditions $M = o(\sqrt{N}/(s^{3/2}\log p))$ and $E(|\varepsilon|^3) < \infty$, $\sqrt{N}|\alpha_{\tau_n}| \lesssim \sqrt{N}M\log p/N = o_p(s^{-3/2}) \leq o_p(1)$ by (S2.24). Noting $\boldsymbol{\beta}^*_{\tau_n} = (\beta_1^* + \alpha_{\tau_n}, \boldsymbol{\beta}^{*\top}_{-1})^\top$, the proof is completed.

For each $j \in \{1, \ldots, p\}$ and $i \in \mathcal{I}_m, m \in \{1, \ldots, M\}$, denote $\zeta_{i,j} = \boldsymbol{\Theta}_j \boldsymbol{x}_i \psi_{\tau_n}(\varepsilon_{i,\tau_n})/(\sqrt{N}\sigma_j)$, by the definition of $\boldsymbol{\beta}^*_{\tau_n}$, we can verify that

$$E(\zeta_{i,j}) = 0, \quad Var(\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\zeta_{i,j}) \ = 1.$$

Moreover, for all $\eta > 0$, with the finite variance assumption in (C1) and $E(|\varepsilon|^3) < \infty$,

$$\lim_{M\to\infty}\lim_{n\to\infty}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m} E[(\zeta_{i,j})^2 I\{|\zeta_{i,j}| > \eta\}] = 0. \tag{S2.25}$$

From Theorem 2, for $1 \leq j \leq p$, with the condition $M = o(\sqrt{N}/(s^{3/2}\log p))$, we can write $\sqrt{N}(\bar{\beta}_j^{\mathbf{d}} - \beta_j^*)/\sigma_j$ as

$$\begin{aligned}\sqrt{N}\frac{\bar{\beta}_j^{\mathbf{d}} - \beta_j^*}{\sigma_j} &= \frac{1}{\sqrt{N}}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\frac{\hat{\boldsymbol{\Theta}}_j^{(m)}\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})}{\sigma_j} + o_p(1)\\ &= \{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\frac{\boldsymbol{\Theta}_j\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})}{\sqrt{N}\sigma_j}\} + \{\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\frac{(\hat{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j)\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})}{\sqrt{N}\sigma_j}\} + o_p(1).\end{aligned}$$

Using (S2.25), we know that $\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\zeta_{i,j}^{(m)} \to N(0,1)$ in distribution by the Lindeberg-Feller central limit theorem. In addition, with (S2.14) and $M = o(\sqrt{N}/(s^{3/2}\log p))$, we have

$$\begin{aligned}|\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\frac{(\hat{\boldsymbol{\Theta}}_j^{(m)}-\boldsymbol{\Theta}_j)\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})}{\sqrt{N}\sigma_j}| &\leq \max_{j,m}\|\hat{\boldsymbol{\Theta}}_j^{(m)}-\boldsymbol{\Theta}_j\|_1\|\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\frac{\boldsymbol{x}_i\psi_{\tau_n}(\varepsilon_{i,\tau_n})}{\sqrt{N}\sigma_j}\|_\infty \\ &= O_p\Big((s_1\vee\sqrt{ss_1})\sqrt{\frac{M\log p}{N}}\Big)O_p(\sqrt{\log p}) = o_p(1).\end{aligned}$$

Thus the proof is completed.

Proof of Theorem 3.

From the results in Theorem 1 and $M = O(\sqrt{N/(s^3\log p)})$, we have $P(\mathcal{E})$ happens with high probability for the event $\mathcal{E} := \{\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}-\boldsymbol{\beta}^*\|_\infty \leq C_0\sqrt{\log p/N}\}$ with a sufficiently large constant C_0, i.e., under $\mathcal{E}$, with $\nu = C_0\sqrt{\log p/N}$, $\nu \geq \|\bar{\boldsymbol{\beta}}^{\mathbf{d}} - \boldsymbol{\beta}^*\|_\infty$ holds. Since $\mathcal{S}$ is the support of $\boldsymbol{\beta}^*$, then for the event $\mathcal{E}$, we have $\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{\mathcal{S}^c}) = 0$ as $\|\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{\mathcal{S}^c}\|_\infty \leq \nu$. For $j \in \mathcal{S}$, if $|\beta^*_j| \geq 2\nu$, note $|\bar{\beta}^{\mathbf{d}}_j| \geq |\beta^*_j| - \nu \geq \nu$, and we get $|\mathcal{T}_\nu(\bar{\beta}^{\mathbf{d}}_j) - \beta^*_j| = |\bar{\beta}^{\mathbf{d}}_j - \beta^*_j| \leq \nu$. When $|\beta^*_j| < 2\nu$, $|\mathcal{T}_\nu(\bar{\beta}^{\mathbf{d}}_j) - \beta^*_j| \leq |\beta^*_j| \vee |\bar{\beta}^{\mathbf{d}}_j - \beta^*_j| \leq 2\nu$. In the event $\mathcal{E}$, we have

$$\begin{aligned}\|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) - \boldsymbol{\beta}^*\|_2 &= \|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_2 \leq 2\sqrt{s}\nu, \\ \|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}) - \boldsymbol{\beta}^*\|_\infty &= \|\mathcal{T}_\nu(\bar{\boldsymbol{\beta}}^{\mathbf{d}}_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty \leq 2\nu.\end{aligned}$$

Proof of Theorem 4.

The proof of Theorem 4 is similar with the proof of Theorem 1. We refer Corollary 3.1 in Luo et al.(2022) as a start point. By multiplying the regularization parameters $\tilde{\lambda}^{[t]}_m$ used in Luo et al. (2022) by an arbitrarily large positive constant independent of m for $t = 1, \ldots, T$, after $T \asymp \lceil \log M \rceil$ rounds of communication, we can show that with probability at least $1 - \log(M)p^{-c}$, there exists a universal constant $C > 0$ independent of m, such that $\|\tilde{\boldsymbol{\beta}}_m^{[T]} - \boldsymbol{\beta}^*\|_1 \leq Cs\sqrt{\log p/N}$ and $\|\tilde{\boldsymbol{\beta}}_m^{[T]} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s \log p/N}$, where $c > 0$ can be arbitrarily large by adjusting the constant C. Similar the proof of Theorem 1, taking $\tilde{\lambda}_m^{[T]}$ uniformly in m and applying the union bound, we obtain that

$$\max_m \|\tilde{\boldsymbol{\beta}}_m^{[T]} - \boldsymbol{\beta}^*\|_1 \leq Cs\sqrt{\log p/N} \text{ and } \max_m \|\tilde{\boldsymbol{\beta}}_m^{[T]} - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s \log p/N},$$

hold with probability at least $1 - \log(M)Mp^{-c}$. Under $\log M = O(\log p)$, $1 - \log(M)Mp^{-c}$ can be written again as $1 - p^{-c}$ (with a different c). Noting $\boldsymbol{\beta}_{\tau_N}^* = (\beta_1^* + \alpha_{\tau_N}, \boldsymbol{\beta}_{-1}^{*\top})^\top$ and $\|\boldsymbol{\beta}_{\tau_N}^* - \boldsymbol{\beta}^*\|_\infty = |\alpha_{\tau_N}| \leq \sqrt{\log p/N}$ by (S2.21), the above results also holds for $\boldsymbol{\beta}_{\tau_N}^*$.

From the construction of $\tilde{\beta}_j^{\mathbf{d}[T]}$, we have

$$\tilde{\beta}_j^{\mathbf{d}[T]} - \beta_{\tau_N,j}^* = \tilde{\beta}_j^{[T]} - \frac{\nabla_{\beta_j} \tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[T]}, \tilde{\boldsymbol{\beta}}_{-j}^{[T]} | \tilde{\boldsymbol{\beta}}^{[T]}) - \hat{\boldsymbol{\gamma}}_j^{(1)\top} \nabla_{\boldsymbol{\beta}_{-j}} \tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[T]}, \tilde{\boldsymbol{\beta}}_{-j}^{[T]} | \tilde{\boldsymbol{\beta}}^{[T]})}{n^{-2} \sum_{i \in \mathcal{I}_1} (x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(1)}) \sum_{i \in \mathcal{I}_1} I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| \leq \tau_n)} - \beta_{\tau_N,j}^*.$$

Note that

$$\begin{aligned} \tilde{\mathcal{L}}_1(\boldsymbol{\beta} | \tilde{\boldsymbol{\beta}}^{[T]}) &= L_{1,\tau_n}(\boldsymbol{\beta}) - \langle \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[T]}) - \nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]}), \boldsymbol{\beta} \rangle \\ &= L_{1,\tau_n}(\boldsymbol{\beta}) - \langle \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[T]}) - \frac{1}{M} \sum_{m=1}^M \nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]}), \boldsymbol{\beta} \rangle. \end{aligned}$$

From the definition of $L_{\tau_N}(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N} \ell_{\tau_N}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ and noting

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[T]}, \tilde{\boldsymbol{\beta}}_{-j}^{[T]} | \tilde{\boldsymbol{\beta}}^{[T]}) =& \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\boldsymbol{\beta}^{[T]}) - \Big(\nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[T]}) - \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\beta}} L_{m,\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]})\Big) \\
=& \nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[T]}) - \Big(\nabla_{\boldsymbol{\beta}} L_{1,\tau_n}(\tilde{\boldsymbol{\beta}}^{[T]}) - \nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]})\Big) \\
=& \nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]}) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}).
\end{aligned}$$

Denote $\tilde{\boldsymbol{\Theta}}_j^{(m)} = \hat{\boldsymbol{\rho}}_j^{(m)} / \{n^{-2}\sum_{i\in\mathcal{I}_m} x_{i,j}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(m)}) \sum_{i\in\mathcal{I}_m} I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| \le \tau_n)\}$, where $\hat{\boldsymbol{\rho}}_j^{(m)} = (-\hat{\gamma}_{j,1}^{(m)}, \ldots, -\hat{\gamma}_{j,(j-1)}^{(m)}, 1, -\hat{\gamma}_{j,j}^{(m)}, \ldots, -\hat{\gamma}_{j,(p-1)}^{(m)})$. Then

$$\begin{aligned}
\tilde{\beta}_j^{\mathbf{d}[T]} - \beta_{\tau_N,j}^* =& \tilde{\beta}_j^{[T]} - \frac{\nabla_{\beta_j} \tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[T]}, \tilde{\boldsymbol{\beta}}_{-j}^{[T]} | \tilde{\boldsymbol{\beta}}^{[T]}) - \hat{\boldsymbol{\gamma}}_j^{(1)\top} \nabla_{\boldsymbol{\beta}_{-j}} \tilde{\mathcal{L}}_1(\tilde{\beta}_j^{[T]}, \tilde{\boldsymbol{\beta}}_{-j}^{[T]} | \tilde{\boldsymbol{\beta}}^{[T]})}{n^{-2}\sum_{i\in\mathcal{I}_1}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(1)}) \sum_{i\in\mathcal{I}_1} I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| \le \tau_n)} - \beta_{\tau_N,j}^* \\
=& \tilde{\beta}_j^{[T]} - \beta_{\tau_N,j}^* - \frac{\hat{\boldsymbol{\rho}}_j^{(1)} \nabla_{\boldsymbol{\beta}} L_{\tau_N}(\tilde{\boldsymbol{\beta}}^{[T]})}{n^{-2}\sum_{i\in\mathcal{I}_1}(x_{i,j} - \boldsymbol{x}_{i,-j}^\top \hat{\boldsymbol{\gamma}}_j^{(1)}) \sum_{i\in\mathcal{I}_1} I(|y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}| \le \tau_n)} \\
=& \tilde{\beta}_j^{[T]} - \beta_{\tau_N,j}^* + \frac{1}{N}\tilde{\boldsymbol{\Theta}}_j^{(1)} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]} - \boldsymbol{\beta}_{\tau_N}^* =& \tilde{\boldsymbol{\beta}}^{[T]} - \boldsymbol{\beta}_{\tau_N}^* + \frac{1}{N}\tilde{\boldsymbol{\Theta}}^{(1)} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) \\
=& \big[\frac{1}{N}\tilde{\boldsymbol{\Theta}}^{(1)} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})\big] + \big[\frac{1}{N}(\tilde{\boldsymbol{\Theta}}^{(1)} - \boldsymbol{\Theta}) \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]})\big] \\
&+ \big[\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\tau_N}^*\big] + \Big\{\frac{1}{N}\boldsymbol{\Theta} \sum_{i=1}^{N} [\boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) - \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})]\Big\} \\
&+ \big[\frac{1}{N}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}^{(1)}) \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})\big] \\
=& \Pi_1 + \Pi_2 + \Pi_3 + \Pi_4 + \Pi_5. \qquad \text{(S2.26)}
\end{aligned}$$

Similarly the proof of Theorem 1,

$$
\begin{aligned}
\Pi_4 =& \frac{1}{N}\boldsymbol{\Theta}\sum_{i=1}^{N}[\boldsymbol{x}_i\psi_{\tau_N}(y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]})-\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})] = \boldsymbol{\Theta}E[\boldsymbol{x}_i\psi_{\tau_N}(y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]})-\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})] \\
&+\Big\{\frac{1}{N}\boldsymbol{\Theta}\sum_{i=1}^{N}[\boldsymbol{x}_i\psi_{\tau_N}(y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]})-\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})]-\boldsymbol{\Theta}E[\boldsymbol{x}_i\psi_{\tau_N}(y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]})-\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})]\Big\} \\
=& \Pi_{4,1}+\Pi_{4,2}.
\end{aligned}
$$

Moreover, the uniform convergence of $\tilde{\boldsymbol{\Theta}}_j^{(m)}$ in j and m can be discussed as in the proof of Theorem 1 with $s_1^2 M\log p/N = o(1)$. From the construction of $\tilde{\boldsymbol{\Theta}}_j^{(m)}$, the only difference between $\tilde{\boldsymbol{\Theta}}_j^{(m)}$ and $\hat{\boldsymbol{\Theta}}_j^{(m)}$ is the indicator part. From (iii) in (C1), $E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]>0$. Similar to the decomposition in the proof of Theorem 1,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)] \\
=&\underbrace{\Big\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)-E[I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)]\Big\}}_{I_1^{'(m)}}+\underbrace{E[I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]}_{I_2^{'}} \\
=&\underbrace{\Big\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)-\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|\varepsilon_{i,\tau_n}|\le\tau_n)-\big[E[I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\big]\Big\}}_{I_{1,1}^{'(m)}} \\
&+\underbrace{\Big\{\frac{1}{n}\sum_{i\in\mathcal{I}_m} I(|\varepsilon_{i,\tau_n}|\le\tau_n)-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]\Big\}}_{I_{1,2}^{'(m)}}+\underbrace{E[I(|y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}^{[T]}|\le\tau_n)]-E[I(|\varepsilon_{i,\tau_n}|\le\tau_n)]}_{I_2^{'}}.
\end{aligned}
$$

Compared with the part of the proof in Theorem 1, $I_{1,1}^{'(m)}$, $I_2^{'}$ depend on $\tilde{\boldsymbol{\beta}}^{[T]}$ rather than $\hat{\boldsymbol{\beta}}_m$. Similarly, it can be proved that $I_{1,1}^{'(m)} = o_p(\sqrt{M/N})$ and $I_{1,2}^{'(m)} = O_p(\sqrt{M/N})$. $I_2^{'} = O_p(\sqrt{s\log p/N})$ can be obtained by the fact that $\|\tilde{\boldsymbol{\beta}}^{[T]}-\boldsymbol{\beta}_{\tau_N}^*\|_2\le C\sqrt{s\log p/N}$ holds with probability at least $1-p^{-c}$.

To proceed, we still need to consider the uniform convergence rates of $I_{1,1}^{(m)'}$ and $I_{1,2}^{(m)'}$ in m.

Note that $I_{1,2}^{(m)'}$ is the leading term and we only need to discuss the uniform convergence rate of $I_{1,2}^{(m)'}$ in m. For $I_{1,2}^{(m)'}$, since $I_{1,2}^{(m)'}$ is a mean of i.i.d. random variables and $0 \leq I(|\varepsilon_{i,\tau_n}| \leq \tau_n) \leq 1$, we have $P(|nI_{1,2}^{(m)'}| \leq t) \geq 1 - 2\exp(-2t^2/n)$ by applying Hoeffding's inequality directly. Taking $t = \sqrt{(cn\log p)/2}$ with an arbitrarily large constant $c > 0$ independent of m and applying the union bound, we have

$$\begin{aligned} P(\max_m |I_{1,2}^{(m)'}| \leq \sqrt{(c\log p)/2n}) &= P(\max_m |nI_{1,2}^{(m)'}| \leq \sqrt{(cn\log p)/2}) \\ &\geq 1 - 2M\exp(-c\log p) = 1 - 2Mp^{-c}. \end{aligned}$$

With $\log M = O(\log p)$, $1 - 2Mp^{-c}$ can be written again as $1 - 2p^{-c}$ (with a different c), thus $\max_m |I_{1,2}^{(m)'}| = O_p(\sqrt{(M\log p)/N})$ and $\max_m |I_{1,1}^{(m)'} + I_{1,2}^{(m)'} + I_2'| = O_p(\sqrt{M\log p/N} \vee \sqrt{s\log p/N})$. Therefore, the uniform convergence rate of $\tilde{\boldsymbol{\Theta}}_j^{(m)}$ in j and m is obtained.

$$\max_m \max_j \|\tilde{\boldsymbol{\Theta}}_j^{(m)} - \boldsymbol{\Theta}_j\|_1 = O_p\Big(s_1\sqrt{\frac{M\log p}{N}} \vee \sqrt{\frac{s_1 s\log p}{N}}\Big) = O_p\Big(s_1\sqrt{\frac{M\log p}{N}}\Big). \tag{S2.27}$$

Using the similar arguments in the proof of Theorem 1, we can derive the bounds of the terms of Π_1, Π_2, $\Pi_3 + \Pi_{4,1}$, $\Pi_{4,2}$ and Π_5 respectively. Unlike the proof of Theorem 1, $\tilde{\boldsymbol{\beta}}^{[T]}$ has a faster convergence rate than $\hat{\boldsymbol{\beta}}_m$ and leads to a better performance. For Π_2, noting (S2.27) and $\tilde{\lambda}_1^{[T]} \asymp \sqrt{\log p/N}$

with $T \asymp \lceil \log M \rceil$, we have

$$\begin{aligned}
\|\Pi_2\|_\infty &= \Big\| \frac{1}{N}(\tilde{\boldsymbol{\Theta}}^{(1)} - \boldsymbol{\Theta}) \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) \Big\|_\infty \\
&\le \max_{1\le j\le p} \|(\tilde{\boldsymbol{\Theta}}_j^{(1)} - \boldsymbol{\Theta}_j)\|_1 \Big\| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) \Big\|_\infty \\
&\le \tilde{\lambda}_1^{[T]} \max_{1\le j\le p} \|\tilde{\boldsymbol{\Theta}}_j^{(1)} - \boldsymbol{\Theta}_j\|_1 \\
&= O_p\Big(\frac{s_1 \sqrt{M} \log p}{N} \Big). \qquad \text{(S2.28)}
\end{aligned}$$

The same bound also holds for $\|\Pi_5\|_\infty$. Then for $\Pi_3 + \Pi_{4,1}$, noting the convergence rate of $\tilde{\boldsymbol{\beta}}^{[T]}$ we discussed at the beginning of the proof, we can obtain

$$\begin{aligned}
&\|\Pi_3 + \Pi_{4,1}\|_\infty \\
&= \|\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\tau_N}^* + \boldsymbol{\Theta} E[\boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) - \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})]\|_\infty \\
&= \|\boldsymbol{\Theta}\boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{[T]} - \boldsymbol{\beta}_{\tau_N}^*) + \boldsymbol{\Theta} E[\boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) - \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})]\|_\infty \\
&\le \max_{1\le j\le p} \|\boldsymbol{\Theta}_j\|_1 \|\boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{[T]} - \boldsymbol{\beta}_{\tau_N}^*) + E[\boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) - \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})]\|_\infty \\
&= O_p\Big(\frac{s\sqrt{s_1} \log p}{N} \Big). \qquad \text{(S2.29)}
\end{aligned}$$

Similarly,

$$\|\Pi_{4,2}\|_\infty = O_p\Big(\frac{(s_1 \vee s_1) \log p}{N} \Big). \qquad \text{(S2.30)}$$

For Π_1, applying Hoeffding's inequality, we get

$$\begin{aligned}\|\Pi_1\|_\infty &= \Big\|\frac{1}{N}\sum_{m=1}^{M}\sum_{i\in\mathcal{I}_m}\tilde{\mathbf{\Theta}}^{(1)}\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})\Big\|_\infty \\ &\le \Big\|\frac{1}{N}\sum_{m=1}^{M}(\tilde{\mathbf{\Theta}}^{(1)}-\mathbf{\Theta})\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})\Big\|_\infty + \Big\|\frac{1}{N}\sum_{m=1}^{M}\mathbf{\Theta}\sum_{i\in\mathcal{I}_m}\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N})\Big\|_\infty \\ &= O_p\Big(\sqrt{\frac{\log p}{N}}\Big).\end{aligned}$$

Combining (S2.28), (S2.29) and (S2.30), $\|\Pi_2\|_\infty$, $\|\Pi_3+\Pi_{4,1}\|_\infty$, $\|\Pi_{4,2}\|$ and $\|\Pi_5\|_\infty$ have a uniform bound. Moreover, noting $\boldsymbol{\beta}^*_{\tau_N} = (\beta^*_1+\alpha_{\tau_N}, \boldsymbol{\beta}^{*\top}_{-1})^\top$ and $\|\boldsymbol{\beta}^*_{\tau_N}-\boldsymbol{\beta}^*\|_\infty = |\alpha_{\tau_N}| \le \sqrt{\log p/N}$ by (S2.21). We get

$$\|\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]}-\boldsymbol{\beta}^*\|_\infty = O_p\Big(\sqrt{\frac{\log p}{N}}+\frac{s\sqrt{M}\log p}{N}\Big).$$

In addition, since $\tilde{\mathbf{\Theta}}_j^{(1)}$ is computed based on the central site and every site can be regarded as a central site and optimize their corresponding optimization problem in parallel, thus

$$\begin{aligned}\tilde{\beta}^{\mathbf{d}[T]}_{all,j}-\beta^*_{\tau_N,j} &= \frac{1}{M}\sum_{m=1}^{M}(\tilde{\beta}^{\mathbf{d}[T]}_{m,j}-\beta^*_{\tau_N,j}) \\ &= \frac{1}{M}\sum_{m=1}^{M}[\tilde{\beta}^{\mathbf{d}[T]}_{m,j}-\beta^*_{\tau_N,j}+\frac{1}{N}\tilde{\mathbf{\Theta}}_j^{(m)}\sum_{i=1}^{N}\boldsymbol{x}_i\psi_{\tau_N}(y_i-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}_m^{[T]})] \\ &= \frac{1}{M}\sum_{m=1}^{M}[\Pi_1^{(m)}+\Pi_2^{(m)}+\Pi_3^{(m)}+\Pi_4^{(m)}+\Pi_5^{(m)}],\end{aligned}$$

where $\Pi_t^{(m)}$, $t=1,2,3,4,5$, are the corresponding terms in (S2.26) by replacing $\tilde{\mathbf{\Theta}}_j^{(1)}$ with $\tilde{\mathbf{\Theta}}_j^{(m)}$. Note that (S2.27) and $\max_m\|\tilde{\boldsymbol{\beta}}_m^{[T]}-\boldsymbol{\beta}^*_{\tau_N}\|_1 \le Cs\sqrt{\log p/N}$, $\max_m\|\tilde{\boldsymbol{\beta}}_m^{[T]}-\boldsymbol{\beta}^*_{\tau_N}\|_2 \le C\sqrt{s\log p/N}$ hold with probability at least $1 - p^{-c}$, as we discussed at the beginning of the proof. Using the similar technique in the proof of Theorem 1, with $|\alpha_{\tau_N}| \leq \sqrt{\log p/N}$, the proof is completed.

Proof of Theorem 5.

For $1 \leq j \leq p$, decompose the jth component $\tilde{\boldsymbol{\beta}}^{\mathbf{d}[T]} - \boldsymbol{\beta}^*_{\tau_N}$ as

$$\begin{aligned}
\tilde{\beta}^{\mathbf{d}[T]}_j - \beta^*_{\tau_N,j} =& \tilde{\beta}^{[T]}_j - \beta^*_{\tau_N,j} + \frac{1}{N}\tilde{\boldsymbol{\Theta}}^{(1)}_j \sum_{i=1}^N \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) \\
=& \Big[\frac{1}{N}\tilde{\boldsymbol{\Theta}}^{(1)}_j \sum_{i=1}^N \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})\Big] + \Big[\frac{1}{N}(\tilde{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \sum_{i=1}^N \boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]})\Big] \\
&+ [\tilde{\beta}^{[T]}_j - \beta^*_{\tau_N,j}] + \Big\{\frac{1}{N}\boldsymbol{\Theta}_j \sum_{i=1}^N [\boldsymbol{x}_i \psi_{\tau_N}(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{[T]}) - \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})]\Big\} \\
&+ \Big[\frac{1}{N}(\boldsymbol{\Theta}_j - \tilde{\boldsymbol{\Theta}}^{(1)}_j) \sum_{i=1}^N \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N})\Big] \\
=& \Xi_1 + \Xi_2 + \Xi_3 + \Xi_4 + \Xi_5,
\end{aligned}$$

Similarly, $\Xi_4 = \Xi_{4,1} + \Xi_{4,2}^{(m)}$. With $M = o(N/(s^2 \log^2 p))$, from the proof of Theorem 4, we have Ξ_2, $\Xi_3 + \Xi_{4,1}$, $\Xi_{4,2}$ and Ξ_5 equal $o_p(N^{-1/2})$. Noting $\boldsymbol{\beta}^*_{\tau_N} = (\beta^*_1 + \alpha_{\tau_N}, \boldsymbol{\beta}^{*\top}_{-1})^\top$, under the conditions $E(|\varepsilon|^3) < \infty$ and $M = o(N/(s^2 \log^2 p))$ in Theorem 4, $\sqrt{N}|\alpha_{\tau_N}| \lesssim \sqrt{N}\log p/N = O_p(\log p/\sqrt{N}) \leq o_p(1)$ by (S2.24). For $1 \leq j \leq p$, we have

$$\tilde{\beta}^{\mathbf{d}[T]}_j - \beta^*_j = \frac{1}{N}\tilde{\boldsymbol{\Theta}}^{(1)}_j \sum_{i=1}^N \boldsymbol{x}_i \psi_{\tau_N}(\varepsilon_{i,\tau_N}) + o_p(N^{-1/2}).$$

Thus from $\tilde{\boldsymbol{\beta}}_{all}^{\mathbf{d}[T]} = M^{-1}\sum_{m=1}^{M}\tilde{\boldsymbol{\beta}}_{m}^{\mathbf{d}[T]}$ and the uniform bounds of $\Pi_t^{(m)}$, $t = 1, 2, 3, 4, 5$, we get

$$\tilde{\beta}_{all,j}^{\mathbf{d}[T]} - \beta_j^* = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\tilde{\boldsymbol{\Theta}}_j^{(m)}\sum_{i=1}^{N}\boldsymbol{x}_i\psi_{\tau_N}(\varepsilon_{i,\tau_N}) + o_p(N^{-1/2}).$$

Proof of Corollary 2 and Theorem 6.

With the results in Theorem 4 and Theorem 5, the proof in this part is similar with the proofs of Corollary 1 and Theorem 3 and thus are omitted for saving space.

References

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high dimensional sparse models. *The Annals of Statistics*, **39**, 82-130.

Boucheron, S., Lugosi, G. and Massart, P. (2013). Concentration inequalities: A nonasymptotic theory of independence. Oxford.

Han, D., Huang, J., Lin, Y. and Shen, G. (2022). Robust post-selection inference of high dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors. *Journal of Econometrics*, **230**, 416–431.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, **15**, 2869–2909.

Lian, H. and Fan, Z. (2018). Divide-and-Conquer for debiased l_1-norm support vector machine in ultra-high dimensions. *Journal of Machine Learning Research*, **18**, 1–26.

Luo, J., Sun, Q. and Zhou, W. X. (2022). Distributed adaptive Huber regression. *Computational Statistics and Data Analysis*, **169**, 107419.

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

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166–1202.

Wang, L., Zheng, C. and Zhou, W. X. (2021). A new principle for tuning-free Huber regression. *Statistica Sinica*, **31**, 2153-2177.

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: maweiha@gmail.com

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: junzhuogao1012@163.com

School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University, Tianjin, 300071, China.

E-mail: lwangstat@nankai.edu.cn

Department of Mathematics, City University of Hong Kong, China.

E-mail: henglian@cityu.edu.hk