# One-Round Communication Efficient Distributed M-Estimation: Supplementary Materials

## A Proof of Main Results

### A.1 Proof of Theorem 5.1

First, we denote another $\ell_2$ ball with smaller radius $U_\delta := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \le \delta_\rho\}$ where $\delta_\rho = \min\{\rho, \rho\lambda_-/(2M)\}$. Then define some good events:

$$
\begin{aligned}
E_{0k} &:= \left\{ \|\nabla L_k(\boldsymbol{\theta}^*)\|_2 \le \frac{(1-\rho)\lambda_-\delta_\rho}{2} \right\} \\
E_{1k} &:= \left\{ \frac{1}{n} \sum_{i \in \mathcal{H}_k} M(\boldsymbol{X}_i) \le M \right\}, \text{ and} \\
E_{2k} &:= \left\{ \|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\|_2 \le \frac{\rho\lambda_-}{2} \right\}.
\end{aligned} \tag{A.1}
$$

**Lemma A.1.** *Suppose the condition (***C**5*) holds, then*

$$\max_{0 \le k \le m} \|\nabla L_k(\boldsymbol{\theta}^*)\|_2 \le C_2 \sqrt{\frac{p}{n}} + t$$

*and*

$$\left\| \frac{1}{m+1} \sum_{k=0}^{m} \nabla L_k(\boldsymbol{\theta}^*) \right\|_2 \le C_2 \sqrt{\frac{p}{N}} + t$$

*hold with probability at least $1 - C_3 m \exp(-C_4 n t^2)$ and $1 - C_3 \exp(-C_4 N t^2)$ respectively, where $C_2, C_3$ and $C_4$ are three universal positive constants.*

Using the Lemma 6 nad 7 in Zhang et al. (2013) and Lemma A.1, we can obtain Lemma A.2 and Lemma A.3.

**Lemma A.2.** *Under event $E = \bigcap_{k=1}^m E_{0k} \cap E_{1k} \cap E_{2k}$, for $\boldsymbol{\theta} \in U_\delta$ we have*

$$\lambda_{\min}(\mathbf{H}_k(\boldsymbol{\theta})) \ge (1-\rho)\lambda_-. \tag{A.2}$$

*And for each local estimator $\widehat{\boldsymbol{\theta}}_k$ we have*

$$\left\| \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \right\|_2 \le \frac{2\|\nabla L_k(\boldsymbol{\theta}^*)\|_2}{(1-\rho)\lambda_-}. \tag{A.3}$$

**Lemma A.3.** *Assume condition (***C**4*) hold, there exists some positive constant $C_1$ such that*

$$\mathbb{E}\left( \|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\|_2^K \right) \le C_1 L^K \left( \frac{\log 2p}{n} \right)^{K/2}.$$

**Lemma A.4.** *Denote the event $E = \bigcap_{k=1}^m E_{0k} \cap E_{1k} \cap E_{2k}$, then there exist three positive constants $c_0, c_1$ and $c_2$ such that*

$$\mathbb{P}(E^c) \le m \left( 2e^{(-c_0 n + 2p)} + c_1 n^{-K/2} + c_2 \left( \frac{\log 2p}{n} \right)^{K/2} \right).$$

*Proof of Theorem 5.1.* Using the fact $\nabla L_k(\widehat{\boldsymbol{\theta}}_k) = 0$, we have

$$\nabla L_k(\boldsymbol{\theta}^*) = \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\left(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\right)$$

where $\widetilde{\boldsymbol{\theta}}_k$ lies between $\widehat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}^*$. It implies that

$$\begin{aligned}
\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{\text{CASE}} &= \left(\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)\right)^{-1}\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)^{-1}\nabla L_k(\boldsymbol{\theta}^*)\\
&= \left(\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)\right)^{-1}\frac{1}{m+1}\sum_{k=0}^{m}\nabla L_k(\boldsymbol{\theta}^*)\\
&+ \left(\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)\right)^{-1}\frac{1}{m+1}\sum_{k=0}^{m}\left(\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right)^{-1}\left(\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right)\nabla L_k(\boldsymbol{\theta}^*),
\end{aligned}$$

Note that under event $E$, we have

$$\begin{aligned}
&\left\|\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right\|_2\\
\leq &\left\|\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_0(\boldsymbol{\theta}^*)\right\|_2 + \|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2 + \left\|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right\|_2\\
\leq &M\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_2 + M\|\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_2 + \|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2\\
\leq &2M\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_2 + \|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2\\
\leq &\frac{4M\|\nabla L_k(\boldsymbol{\theta}^*)\|_2}{(1-\rho)\lambda_-} + \|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2\,,
\end{aligned} \tag{A.4}$$

where the second inequality follows from the definition of $E_{1k}$ and the third inequality follows from (A.3) in Lemma A.2. Combining with (A.2) in Lemma A.2, we have

$$\begin{aligned}
\left\|\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)^{-1}\left(\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right)\nabla L_k(\boldsymbol{\theta}^*)\right\|_2 &\leq \left\|\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)^{-1}\left(\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right)\right\|_2\|\nabla L_k(\boldsymbol{\theta}^*)\|_2\\
&\leq \frac{4M\|\nabla L_k(\boldsymbol{\theta}^*)\|_2^2}{(1-\rho)^2\lambda_-^2} + \frac{\|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2\|\nabla L_k(\boldsymbol{\theta}^*)\|_2}{(1-\rho)\lambda_-}.
\end{aligned}$$

On the other hand, we make use of the fact that: for any matrix $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times p}$,

$$\left\|(\boldsymbol{A}+\boldsymbol{B})^{-1} - \boldsymbol{A}^{-1}\right\|_2 \leq \left\|\boldsymbol{A}^{-1}\right\|_2^2\|\boldsymbol{B}\|_2\,.$$

Now let $\boldsymbol{A} = \mathbf{I}(\boldsymbol{\theta}^*)$ and $\boldsymbol{B} = \mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{I}(\boldsymbol{\theta}^*)$ and using Lemma A.2 again, we have

$$\begin{aligned}
\left\|\left(\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)\right)^{-1} - \mathbf{I}(\boldsymbol{\theta}^*)^{-1}\right\|_2 &\leq \left\|\mathbf{I}(\boldsymbol{\theta}^*)^{-1}\right\|_2^2\left\|\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{I}(\boldsymbol{\theta}^*)\right\|_2\\
&\leq \max_{0\leq k\leq m}\left\|\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{I}(\boldsymbol{\theta}^*)\right\|_2\left\|\mathbf{I}(\boldsymbol{\theta}^*)^{-1}\right\|_2^2\\
&\leq \max_{0\leq k\leq m}\left(\left\|\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k) - \mathbf{H}_0(\boldsymbol{\theta}^*)\right\|_2 + \|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\|_2\right)\left\|\mathbf{I}(\boldsymbol{\theta}^*)^{-1}\right\|_2^2\\
&\leq \frac{2M\|\nabla L_k(\boldsymbol{\theta}^*)\|_2}{(1-\rho)\lambda_-^3} + \lambda_-^2\max_{0\leq k\leq m}\|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\|_2\\
&= o(1).
\end{aligned}$$

Then under event $E$, there exists some positive constant $C$ such that

$$\|u_n\|_2 \leq C\max_{0\leq k\leq m}\left(\frac{4M\|\nabla L_k(\boldsymbol{\theta}^*)\|_2^2}{(1-\rho)^2\lambda_-^3} + \frac{\|\mathbf{H}_0(\boldsymbol{\theta}^*) - \mathbf{H}_k(\boldsymbol{\theta}^*)\|_2\|\nabla L_k(\boldsymbol{\theta}^*)\|_2}{(1-\rho)\lambda_-^2}\right).$$

The result follows from Lemma A.4. ∎

### A.2 Proof of Corollary 5.1

*Proof.* For any $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^{N} \mathbb{E}\left( \left\| \nabla l(\boldsymbol{X}_i, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2^2 \mathbb{I}\left\{ \left\| \nabla l(\boldsymbol{X}_i, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2 > \varepsilon \right\} \right) \\
=&\mathbb{E}\left( \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2^2 \mathbb{I}\left\{ \left\| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2 > \varepsilon \right\} \right) \\
\leq& \left[ \mathbb{E} \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2^4 \right]^{1/2} \left[ \mathbb{P}\left( \left\| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2 > \varepsilon \right) \right]^{1/2}.
\end{aligned}$$

Using condition (**C**5), we can prove for any $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[ \exp\left( \lambda \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2 \right) \right] \leq 2e^{\lambda^2}. \tag{A.5}$$

Then using Markov's inequality and (A.5), we have

$$\mathbb{P}\left( \left\| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2 > \varepsilon \right) \leq 2\exp\left( \frac{\lambda^2}{N} - \lambda\varepsilon \right) \leq 2\exp\left( -\frac{N\varepsilon}{4} \right).$$

Moreover note that

$$\begin{aligned}
\mathbb{E} \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2^4 &= \int_0^\infty \mathbb{P}\left( \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2^4 > t \right) dt \\
&= 4 \int_0^\infty \mathbb{P}\left( \| \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*) \|_2 > t \right) t^3 dt \\
&\leq 8 \int_0^\infty e^{-t^2/2} t^3 dt = 4\Gamma(2) < \infty,
\end{aligned}$$

where $\Gamma(x)$ is Gamma function. Thus

$$\sum_{i=1}^{N} \mathbb{E}\left( \left\| \nabla l(\boldsymbol{X}_i, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2^2 \mathbb{I}\left\{ \left\| \nabla l(\boldsymbol{X}_i, \boldsymbol{\theta}^*)/\sqrt{N} \right\|_2 > \varepsilon \right\} \right) = o(1).$$

Then the result follows from Lindeberg–Feller central limit theorem. □

### A.3 Proof of Theorem 5.2

*Proof.* The gradient of logistic regression is $\nabla l(\boldsymbol{X}, \boldsymbol{\theta}) = -Y\boldsymbol{X}/(1 + \exp(Y\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}))$, then we have for any $\lambda \in \mathbb{R}$

$$\sup_{\|\boldsymbol{u}\|_2 = 1} \mathbb{E}\left[ \exp\left( \lambda |\boldsymbol{u}^{\mathrm{T}} \nabla l(\boldsymbol{X}, \boldsymbol{\theta}^*)| \right) \right] \leq \sup_{\|\boldsymbol{u}\|_2 = 1} \mathbb{E}\left[ \exp\left( \lambda |\boldsymbol{u}^{\mathrm{T}} \boldsymbol{X}| \right) \right] \leq \exp(\lambda^2).$$

It implies that condition (**C**5) holds. Let $p(\boldsymbol{\theta}) = (1 + \exp(-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}))$ and using $p(\boldsymbol{\theta})(1 - p(\boldsymbol{\theta})) < 1$, we have

$$\sup_{\|\boldsymbol{u}\|_2 = 1} \mathbb{E}\left[ \exp\left( \lambda |\boldsymbol{u}^{\mathrm{T}} \boldsymbol{X} \sqrt{p(\boldsymbol{\theta})(1 - p(\boldsymbol{\theta}))}| \right) \right] \leq \exp(\lambda^2).$$

Let $\boldsymbol{Z} = \boldsymbol{X}\sqrt{p(\boldsymbol{\theta})(1 - p(\boldsymbol{\theta}))}$ then $\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})$. Then according to Theorem 6.5 in Wainwright (2019), we can prove that

$$\max_k \| \mathbf{H}_k(\boldsymbol{\theta}) - \mathbf{I}(\boldsymbol{\theta}) \|_2 = O_{\mathbb{P}}\left( \sqrt{\frac{p}{n}} \right).$$

From Proposition D.1 in Chen et al. (2018), we can verify condition (**C**3) and (**C**4). Therefore, the $\ell_2$ error bound can be obtained from the proof of Theorem 5.1.

□

### A.4 Proof of Theorem 5.3

*Proof.* For the ease of the representation, we use $\widehat{\boldsymbol{\theta}}$ to denote $\widehat{\boldsymbol{\theta}}_{\text{Pen-CASE}}$. By the optimality of $\widehat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}&\frac{1}{2(m+1)}\widehat{\boldsymbol{\theta}}^{\mathrm{T}}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\right)\widehat{\boldsymbol{\theta}}-\frac{1}{m+1}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\widehat{\boldsymbol{\theta}}_k\right)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}+\lambda_n\|\widehat{\boldsymbol{\theta}}\|_1\\ \leq&\frac{1}{2(m+1)}\boldsymbol{\theta}^{*\mathrm{T}}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\right)\boldsymbol{\theta}^*-\frac{1}{m+1}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\widehat{\boldsymbol{\theta}}_k\right)^{\mathrm{T}}\boldsymbol{\theta}^*+\lambda_n\|\boldsymbol{\theta}^*\|_1,\end{aligned}$$

which implies that

$$\begin{aligned}&\frac{1}{2(m+1)}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)^{\mathrm{T}}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\right)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)+\lambda_n\|\widehat{\boldsymbol{\theta}}\|_1\\ \leq&\frac{1}{m+1}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}^*)\right)^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)+\lambda_n\|\boldsymbol{\theta}^*\|_1.\end{aligned}\tag{A.6}$$

Note that,

$$\begin{aligned}&\left\|\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}^*)\right\|_\infty\\ \leq&\left\|\frac{1}{m+1}\sum_{k=0}^{m}\nabla L_k(\boldsymbol{\theta}^*)\right\|_\infty+\max_{0\leq k\leq m}\left\|\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)^{-1}\left(\mathbf{H}_0(\widehat{\boldsymbol{\theta}}_k)-\mathbf{H}_k(\widetilde{\boldsymbol{\theta}}_k)\right)\right\|_2\|\nabla L_k(\boldsymbol{\theta})\|_2.\end{aligned}\tag{A.7}$$

From the proof of Theorem 5.1 and the definition of $\lambda_n$, there exists some sufficiently large $C_3$ such that

$$\left\|\frac{1}{m+1}\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}^*)\right\|_\infty\leq\frac{\lambda_n}{2}.\tag{A.8}$$

Under event $E$, $\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)/(m+1)$ is positive definite, combining with (A.6) we have

$$\lambda_n\|\widehat{\boldsymbol{\theta}}\|_1\leq\frac{\lambda_n}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1+\lambda_n\|\boldsymbol{\theta}^*\|_1.$$

Then using the fact $\boldsymbol{\theta}^*_{S^c}=\mathbf{0}$, we can obtain $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1\leq4\|\widehat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}^*_S\|_1$. From (A.6), we also have

$$(1-\rho)\lambda_-\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2\leq\frac{3\lambda_n}{2}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1\leq6\lambda_n\sqrt{s}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2,$$

where the first inequality follows from Lemma A.2 and the second inequality follows from $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1\leq4\|\widehat{\boldsymbol{\theta}}_S-\boldsymbol{\theta}^*_S\|_1$. Under condition (**C**5), for any $\lambda>0$ there exists some positive constant $c$ such that

$$\max_{1\leq j\leq p}\mathbb{E}[\exp(\lambda\nabla_j l(\boldsymbol{X},\boldsymbol{\theta}^*))]\leq\exp(\lambda^2).$$

Then using the maximum sub-gaussian inequality, we have

$$\mathbb{P}\left(\left\|\frac{1}{m+1}\sum_{k=0}^{m}\nabla L_k(\boldsymbol{\theta})\right\|_\infty\geq c\sqrt{\frac{\log N}{N}}\right)\leq pN^{-c^2/2}.$$

□

### A.5 Proof of Theorem 5.4

*Proof.* We first define the oracle problem:

$$\widehat{\boldsymbol{\theta}}^o_S=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^s}\frac{1}{2(m+1)}\boldsymbol{\theta}^{\mathrm{T}}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\right)_{SS}\boldsymbol{\theta}-\frac{1}{m+1}\left(\sum_{k=0}^{m}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\widehat{\boldsymbol{\theta}}_k\right)_S^{\mathrm{T}}\boldsymbol{\theta}+\lambda_n\|\boldsymbol{\theta}\|_1.$$

Let $\widehat{\mathbf{H}} = \sum_{k=0}^{m} \mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)/(m+1)$ and $\boldsymbol{b} = \sum_{k=0}^{m} \mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\widehat{\boldsymbol{\theta}}_k/(m+1)$, then by the KKT condition we have

$$\widehat{\mathbf{H}}_{SS}\widehat{\boldsymbol{\theta}}_S^o - \boldsymbol{b}_S + \lambda_n \boldsymbol{z}_S = \mathbf{0},$$

which implies that $\widehat{\boldsymbol{\theta}}_S^o = \widehat{\mathbf{H}}_{SS}^{-1}(\boldsymbol{b}_S - \lambda_n \boldsymbol{z}_S)$. If we can prove there exists $\boldsymbol{z}_{S^c} \in \mathbb{R}^{p-s}$ satisfying the zero subgradient condition

$$\widehat{\mathbf{H}}_{S^cS}\widehat{\boldsymbol{\theta}}_S^o - \boldsymbol{b}_{S^c} + \lambda_n \boldsymbol{z}_{S^c} = \mathbf{0}$$

and $\|\boldsymbol{z}\|_\infty < 1$, then $\widehat{S} \subseteq S$. Substitute $\widehat{\boldsymbol{\theta}}_S^o$ to the above equation then we can write $\boldsymbol{z}_{S^c}$ as

$$\begin{aligned}
\boldsymbol{z}_{S^c} &= \lambda_n^{-1}\left\{\boldsymbol{b}_{S^c} - \widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\left(\boldsymbol{b}_S - \lambda_n \boldsymbol{z}_S\right)\right\} \\
&= \lambda_n^{-1}\left\{\boldsymbol{b}_{S^c} - \widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\boldsymbol{b}_S\right\} + \widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\boldsymbol{z}_S \\
&= \lambda_n^{-1}\left\{\left[\frac{1}{m+1}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right)\right]_{S^c} - \widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\left[\frac{1}{m+1}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right)\right]_S\right\} + \widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\boldsymbol{z}_S.
\end{aligned}$$

Let $\mathbf{I} = \mathbf{I}(\boldsymbol{\theta}^*)$, then note that

$$\widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1} = \left(\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right)\left(\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right) + \mathbf{I}_{S^cS}\left(\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right) + \left(\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right)\mathbf{I}_{SS}^{-1} + \mathbf{I}_{S^cS}\mathbf{I}_{SS}^{-1}.$$

According to Lemma A.3 we have

$$\left\|\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right\|_2 \lesssim \sqrt{\frac{p}{n}}.$$

Moreover,

$$\left\|\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right\|_2 \leq \max_{0\leq k\leq m}\left\|\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k) - \mathbf{I}(\boldsymbol{\theta}^*)\right\|_2 \lesssim \sqrt{\frac{p}{n}}.$$

Then we have

$$\left\|\left(\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right)\left(\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right)\right\|_\infty \leq s\left\|\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right\|_2\left\|\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right\|_2 \lesssim \frac{sp}{n},$$

$$\left\|\mathbf{I}_{S^cS}\left(\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right)\right\|_\infty \leq s\left\|\mathbf{I}\right\|_2\left\|\widehat{\mathbf{H}}_{SS}^{-1} - \mathbf{I}_{SS}^{-1}\right\|_2 \lesssim s\sqrt{\frac{p}{n}},$$

and

$$\left\|\left(\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right)\mathbf{I}_{SS}^{-1}\right\|_\infty \leq s\left\|\mathbf{I}_{SS}^{-1}\right\|_2\left\|\widehat{\mathbf{H}}_{S^cS} - \mathbf{I}_{S^cS}\right\|_2 \lesssim s\sqrt{\frac{p}{n}}.$$

Combing the results above, it follows that $\|\widehat{\mathbf{H}}_{S^cS}\widehat{\mathbf{H}}_{SS}^{-1}\|_\infty \lesssim \|\mathbf{I}_{S^cS}\mathbf{I}_{SS}^{-1}\|_\infty + s\sqrt{p}/\sqrt{n}$. Due to the definition of $\lambda_n$ and (A.8), there exists some sufficiently large positive constant $C_0$ such that

$$\|\boldsymbol{z}_{S^c}\|_\infty \leq \frac{1}{C_0} + \|\mathbf{I}_{S^cS}\mathbf{I}_{SS}^{-1}\|_\infty < \frac{1}{C_0} + \alpha < 1$$

with probability tending to 1. Using the fact $\boldsymbol{\theta}_S^* = \widehat{\mathbf{H}}_{SS}^{-1}(\widehat{\mathbf{H}}\boldsymbol{\theta}^*)_S$, we can obtain

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_S^o - \boldsymbol{\theta}_S^*\right\|_\infty &\leq \left\|\widehat{\mathbf{H}}_{SS}^{-1}\right\|_\infty\left\|\frac{1}{m+1}\mathbf{H}_k(\widehat{\boldsymbol{\theta}}_k)\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right)\right\|_\infty + \left\|\lambda_n\widehat{\mathbf{H}}_{SS}^{-1}\boldsymbol{z}_S\right\|_\infty \\
&\lesssim \left\|\mathbf{I}_{SS}^{-1}\right\|_\infty \lambda_n.
\end{aligned}$$

□

### A.6 Proof of Theorem 5.5

*Proof.* It suffices to prove that

$$\left\|\widehat{\mathbf{C}}_0\left(\boldsymbol{\theta}^* - \frac{1}{m+1}\sum_{k=0}^{m}\widehat{\boldsymbol{\theta}}_k\right)\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N}}\right). \tag{A.9}$$

Note that

$$\frac{1}{m+1}\sum_{k=0}^{m}\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* = \frac{1}{m+1}\sum_{k=0}^{m}\widehat{\mathbf{C}}_k^{-1}\frac{1}{n}\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i,$$

where $\widehat{\mathbf{C}}_k = \sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}/n$. Moreover, using the subgaussian property of $\boldsymbol{X}_i$ and $\epsilon_i$ we can easily obtain $\max_k\|\widehat{\mathbf{C}}_k - \mathbf{C}\|_2 = O_{\mathbb{P}}(\sqrt{p/n})$, $\max_k\|\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i/n\|_2 = O_{\mathbb{P}}(\sqrt{p/n})$ and $\|\sum_{k=0}^{m}\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i/N\|_\infty = O_{\mathbb{P}}(\sqrt{\log N/N})$. It follows that

$$\begin{aligned}
\left\|\widehat{\mathbf{C}}_0\left(\boldsymbol{\theta}^* - \frac{1}{m+1}\sum_{k=0}^{m}\widehat{\boldsymbol{\theta}}_k\right)\right\|_\infty &= \left\|\frac{1}{m+1}\sum_{k=0}^{m}\widehat{\mathbf{C}}_0\widehat{\mathbf{C}}_k^{-1}\frac{1}{n}\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i\right\|_\infty \\
&\le \left\|\frac{1}{N}\sum_{k=0}^{m}\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i\right\|_\infty + \max_k\left\|\widehat{\mathbf{C}}_k^{-1}\right\|_2\left\|\widehat{\mathbf{C}}_0 - \widehat{\mathbf{C}}_k\right\|_2\left\|\frac{1}{n}\sum_{i\in\mathcal{H}_k}\boldsymbol{X}_i\epsilon_i\right\|_2 \\
&= O_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N}} + \frac{p}{n}\right).
\end{aligned}$$

Then (A.9) follows from $m \lesssim \sqrt{N\log N}/p$. ◻

## B Proof of Auxiliary Lemmas

### B.1 Proof of Lemma A.1

*Proof.* Note that the variational representation $\|\nabla L_k(\boldsymbol{\theta}^*)\|_2 = \sup_{\boldsymbol{u}\in S^{p-1}}|\boldsymbol{u}^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)|$, where $S^{p-1}$ is the sphere in $\mathbb{R}^p$. Let $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_M\}$ be a 1/2-covering with $M \le 5^p$ vectors. For any $\boldsymbol{u}\in S^{p-1}$, there is some $\boldsymbol{u}_j$ such that $\boldsymbol{u} = \boldsymbol{u}_j + \Delta$ with $\|\Delta\|_2 \le 1/2$. Thus we have

$$\boldsymbol{u}^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*) = \boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*) + \Delta^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*),$$

which indicates that

$$|\boldsymbol{u}^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)| \le |\boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)| + \frac{1}{2}\left\|\nabla L_k(\boldsymbol{\theta}^*)\right\|_2.$$

It yields that

$$\|\nabla L_k(\boldsymbol{\theta}^*)\|_2 \le 2\max_{1\le j\le M}|\boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)|.$$

Then we have,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\left\|\nabla L_k(\boldsymbol{\theta}^*)\right\|_2\right)\right] &\le \mathbb{E}\left[\exp\left(2\lambda\max_{1\le j\le M}|\boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)|\right)\right] \\
&\le \sum_{j=1}^{M}\mathbb{E}\left[\exp\left(2\lambda|\boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)|\right)\right].
\end{aligned}$$

Due to the definition of $\nabla L_k(\boldsymbol{\theta}^*)$ and condition (**C**5), we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(2\lambda|\boldsymbol{u}_j^{\mathrm{T}}\nabla L_k(\boldsymbol{\theta}^*)|\right)\right] &\le \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{2\lambda}{n}|\boldsymbol{u}_j^{\mathrm{T}}\nabla l(\boldsymbol{X}_i, \boldsymbol{\theta}^*)|\right)\right] \\
&\le \exp\left(\frac{4\lambda^2}{n}\right).
\end{aligned}$$

Consequently, we have

$$\mathbb{E}\left[\exp\left(\lambda \left\|\nabla L_k(\boldsymbol{\theta}^*)\right\|_2\right)\right] \leq 5^p \exp\left(\frac{4\lambda^2}{n}\right) \leq \exp\left(\frac{4\lambda^2}{n} + 2p\right). \tag{B.1}$$

Then it follows that

$$\mathbb{P}\left(\left\|\nabla L_k(\boldsymbol{\theta}^*)\right\|_2 \geq C_2\sqrt{\frac{p}{n}} + t\right) \leq C_3 \exp\left(-C_4 n t^2\right)$$

for some constants $C_2, C_3$ and $C_4$. □

### B.2 Proof of Lemma A.4

*Proof.* From the proof of Lemma A.1,

$$\begin{aligned}\mathbb{P}\left(E_{0k}^c\right) = \mathbb{P}\left(\left\|\nabla L_k(\boldsymbol{\theta}^*)\right\|_2 \geq \frac{(1-\rho)\lambda_-\delta_\rho}{2}\right) &\leq 2\exp\left(\frac{4\lambda^2}{n} + 2p - \frac{\lambda(1-\rho)\lambda_-\delta_\rho}{2}\right) \\ &\leq 2\exp\left(-\frac{(1-\rho)\lambda_-\delta_\rho}{32} n + 2p\right) \leq 2\exp\left(-c_0 n + 2p\right).\end{aligned}$$

Then according to our condition (**C**4) and Markov's inequality

$$\mathbb{P}\left(E_{1k}^c\right) = \mathbb{P}\left(M_k \geq M\right) \leq \frac{\mathbb{E}(\sum_{i\in\mathcal{H}_k} M(\boldsymbol{X}_i))^K}{n^K M} \leq c_1 n^{-K/2}.$$

Using (B.1) in Lemma A.3, we have

$$\begin{aligned}\mathbb{P}\left(E_{2k}^c\right) = \mathbb{P}\left(\left\|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\right\|_2 \geq \frac{\rho\lambda_-}{2}\right) &\leq \frac{2^K \mathbb{E}\left(\left\|\mathbf{H}_k(\boldsymbol{\theta}^*) - \mathbf{I}(\boldsymbol{\theta}^*)\right\|_2^K\right)}{\rho^K \lambda_-^K} \\ &\leq (2L)^K (\rho\lambda_-)^{-K}\left(\frac{\log 2p}{n}\right)^{K/2} \leq c_2\left(\frac{\log 2p}{n}\right)^{K/2}.\end{aligned}$$

Therefore, using the union bound we can obtain that

$$\mathbb{P}(E^c) \leq m\left(2e^{(-c_0 n + 2p)} + c_1 n^{-K/2} + c_2\left(\frac{\log 2p}{n}\right)^{K/2}\right).$$

□

## References


Chen, X., Liu, W., and Zhang, Y. (2018). First-order newton-type estimator for distributed estimation and inference. *ArXiv*, abs/1811.11368.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363.