Supplementary Material for "Byzantine-tolerant Distributed Multiclass Sparse Discriminant Analysis"

Abstract

This document provides supplementary material to the article "Byzantine-tolerant Distributed Multiclass Sparse Discriminant Analysis" written by the same authors.

1 Proofs of Main Results

1.1 Proof of Theorem 3.1

Theorem 1.1. *Let* $\max_{2\le k\le K}\|\widehat{\boldsymbol{\theta}}_k^{(0)}-\boldsymbol{\theta}_k^*\|_1=O_{\mathbb{P}}(a_n)$ *and choose some sufficiently large positive constant* η_1 *such that*

$$\lambda_1=\begin{cases}\eta_1\left(\sqrt{\frac{\log p}{N}}+a_n\sqrt{\frac{\log p}{n}}\right), & \textit{System I}\\ \eta_1\left(\sqrt{\frac{\log p}{N}}+a_n\sqrt{\frac{\log p}{n}}+\frac{\alpha}{\sqrt{n}}+\frac{1}{n}\right), & \textit{System II}\end{cases}$$

under conditions $(\mathbf{C}1)$, $(\mathbf{C}3)$ *and* $(\mathbf{C}5)$, *we have*

$$\max_{2\le k\le K}\left\|\widehat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_k^*\right\|_2=O_{\mathbb{P}}(\sqrt{s}\lambda_1),\tag{1.1}$$

and

$$\max_{2\le k\le K}\left\|\widehat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_k^*\right\|_1=O_{\mathbb{P}}(s\lambda_1).\tag{1.2}$$

Note that the initial error for the t-th iteration would be $\max_{2\le k\le K}\|\widehat{\boldsymbol{\theta}}_k^{(t-1)}-\boldsymbol{\theta}_k^*\|_1$ by plug-in the initial estimator $\widehat{\boldsymbol{\theta}}_k^{(t-1)}$. Then we can obtain the ℓ_1 and ℓ_2 error bound of the $\widehat{\boldsymbol{\theta}}_k^{(t)}$ easily by induction according to (1.1) and (1.2) in Theorem 1.1. In the next, to show the proof of Theorem 1.1, we present several useful lemmas in the following.

Lemma 1.1. *For* $\boldsymbol{x}_k\in\mathbb{R}^p, k=1,...,K-1$ *such that*

$$\sum_{k=1}^{K-1}\|\boldsymbol{x}_k\|_1\le 4\sqrt{s(K-1)}\left(\sum_{k=1}^{K-1}\|\boldsymbol{x}_k\|_2^2\right)^{1/2},$$

we have

$$\sum_{k=1}^{K-1} \boldsymbol{x}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \boldsymbol{x}_k^{\mathrm{T}} \geq L \sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_2^2,$$

holds with probability tending to 1.

Lemma 1.2. *Let $\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|_1 = O_{\mathbb{P}}(a_n)$, then for $k = 2, ..., K$ we have*

$$\left\|\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \widehat{\boldsymbol{b}}_{k,0}\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + a_n\sqrt{\frac{\log p}{n}}\right),$$

under System I.

Lemma 1.3. *Let $\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|_1 = O_{\mathbb{P}}(a_n)$, then for $k = 2, ..., K$ we have*

$$\left\|\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \widetilde{\boldsymbol{b}}_{k,0}\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + a_n\sqrt{\frac{\log p}{n}} + \frac{\alpha}{\sqrt{n}} + \frac{1}{n}\right),$$

under System II.

The proofs of above lemmas are relegated to Section 2 in the following.

Proof of Theorem 3.1. For simplicity of notations, we use $\widehat{\boldsymbol{\theta}}_k$ to denote $\widehat{\boldsymbol{\theta}}_k^{(1)}$. By the optimality of $(\widehat{\boldsymbol{\theta}}_2, \cdots, \widehat{\boldsymbol{\theta}}_K)$, we have

$$\begin{aligned}
&\frac{1}{2}\widehat{\boldsymbol{\theta}}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \widehat{\boldsymbol{\theta}}_k - \left(\widehat{\boldsymbol{\Sigma}}_{(0)}\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{b}_{k,0}\right)^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_k + \lambda_1 \sum_{j=1}^{p}\left(\sum_{l=2}^{K} \widehat{\boldsymbol{\theta}}_{l,j}^2\right)^{1/2} \\
\leq &\frac{1}{2}\boldsymbol{\theta}_k^{*\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \boldsymbol{\theta}_k^* - \left(\widehat{\boldsymbol{\Sigma}}_{(0)}\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{b}_{k,0}\right)^{\mathrm{T}} \boldsymbol{\theta}_k^* + \lambda_1 \sum_{j=1}^{p}\left(\sum_{l\neq k} \widehat{\boldsymbol{\theta}}_{l,j}^2 + (\boldsymbol{\theta}_{k,j}^*)^2\right)^{1/2}.
\end{aligned}$$

By rearranging the terms above, we have

$$\begin{aligned}
&\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right) \\
\leq &\left(\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \boldsymbol{b}_{k,0}\right)^{\mathrm{T}} \left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right) + \lambda_1 \left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k\right\|_1.
\end{aligned} \tag{1.3}$$

According to Lemmas 1.2 and 1.3,

$$\left\|\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \boldsymbol{b}_{k,0}\right\|_\infty = \begin{cases} O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + a_n\sqrt{\frac{\log p}{n}}\right), & \text{System I} \\ O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + a_n\sqrt{\frac{\log p}{n}} + \frac{\alpha}{\sqrt{n}} + \frac{1}{n}\right). & \text{System II} \end{cases}$$

Thus $\|\widehat{\boldsymbol{\Sigma}}_{(0)}\boldsymbol{\theta}_k^* - \boldsymbol{b}_{k,0}\|_\infty \leq \lambda_1/2$ holds with high probability for some sufficiently large positive constant η_1 under both System I and II. Then (1.3) indicates that

$$\frac{1}{2}\sum_{k=2}^{K}\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right) \leq \frac{3\lambda_1}{2}\sum_{k=2}^{K}\left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k\right\|_1. \tag{1.4}$$

By the optimality of $(\widehat{\boldsymbol{\theta}}_2, \cdots, \widehat{\boldsymbol{\theta}}_K)$, we can also obtain that

$$
\begin{aligned}
&\frac{1}{2}\sum_{k=2}^{K}\left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right) + \lambda_1 \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)}\|_2 \\
\leq &\sum_{k=2}^{K}\left(\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \boldsymbol{b}_{k,0}\right)^{\mathrm{T}} \left(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\right) + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2.
\end{aligned} \tag{1.5}
$$

Let $\boldsymbol{c}_k = \widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \boldsymbol{b}_{k,0}$ and $\boldsymbol{c}_{(j)} = (\boldsymbol{c}_{2,j}, ..., \boldsymbol{c}_{K,j})^{\mathrm{T}}$, then it follows from (1.5) and $\|\boldsymbol{c}_k\|_\infty \leq \lambda_1/2$ that

$$
\begin{aligned}
\lambda_1 \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)}\|_2 &\leq \sum_{j=1}^{p} \boldsymbol{c}_{(j)}^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*) + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2 \\
&\leq \sum_{j=1}^{p} \|\boldsymbol{c}_{(j)}\|_2 \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2 \\
&\leq \max_j \|\boldsymbol{c}_{(j)}\|_2 \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2 \\
&\leq \sqrt{K-1} \max_{2\leq k\leq K} \|\boldsymbol{c}_k\|_\infty \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2 \\
&\leq \frac{\lambda_1}{2} \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 + \lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\theta}_{(j)}^*\|_2,
\end{aligned}
$$

which implies that $\sum_{j\in S^c} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 \leq 3\sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2$. In conjunction with the fact that

$$
\left(\sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2\right)^2 \leq s \sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2^2,
$$

we have

$$
\begin{aligned}
\sum_{j\in S^c} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_1 \leq \sqrt{K-1} \sum_{j\in S^c} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 &\leq 3\sqrt{K-1} \sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2 \\
&\leq 3\sqrt{s(K-1)} \left(\sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_2^2\right)^{1/2} \\
&\leq 3\sqrt{s(K-1)} \left(\sum_{k=2}^{K} \|\widehat{\boldsymbol{\theta}}_{k,S}^{(1)} - \boldsymbol{\theta}_{k,S}^*\|_2^2\right)^{1/2}.
\end{aligned}
$$

Similarly, we have $\sum_{j\in S} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_1 \leq \sqrt{s(K-1)}(\sum_{k=2}^{K} \|\widehat{\boldsymbol{\theta}}_{k,S}^{(1)} - \boldsymbol{\theta}_{k,S}^*\|_2^2)^{1/2}$. It implies that

$$
\sum_{k=2}^{K} \left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k\right\|_1 = \sum_{j=1}^{p} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}_{(j)}^*\|_1 \leq 4\sqrt{s(K-1)} \left(\sum_{k=2}^{K} \|\widehat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_k^*\|_2^2\right)^{1/2}. \tag{1.6}
$$

By applying Lemma 1.1, we have

$$\frac{1}{2}\sum_{k=2}^{K}\left(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*\right)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{(0)}\left(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*\right)\geq L\sum_{k=2}^{K}\left\|\boldsymbol{\theta}_k^*-\widehat{\boldsymbol{\theta}}_k\right\|_2^2. \tag{1.7}$$

Combining the inequalities (1.4), (1.6) and (1.7), we have

$$\left(\sum_{k=2}^{K}\left\|\boldsymbol{\theta}_k^*-\widehat{\boldsymbol{\theta}}_k\right\|_2^2\right)^{1/2}\leq\frac{6\sqrt{s(K-1)}}{L}\lambda_1,$$

and

$$\sum_{k=2}^{K}\left\|\boldsymbol{\theta}_k^*-\widehat{\boldsymbol{\theta}}_k\right\|_1\leq\frac{24s(K-1)}{L}\lambda_1.$$

It also indicates that

$$\max_{2\leq k\leq K}\|\boldsymbol{\theta}_k^*-\widehat{\boldsymbol{\theta}}_k\|_2=O_{\mathbb{P}}(\sqrt{s}\lambda_1),$$

and

$$\max_{2\leq k\leq K}\|\boldsymbol{\theta}_k^*-\widehat{\boldsymbol{\theta}}_k\|_1=O_{\mathbb{P}}(s\lambda_1).$$

□

1.2 Proof of Theorem 3.2

It suffices to prove Theorem 1.2 in the following.

Theorem 1.2. *Under conditions* (**C**1)-(**C**5)*, with the same choice of* λ_1 *as in Theorem 1.1, we have* $\widehat{S}^{(1)}\subseteq S$ *holds with probability tending to 1 and* $\widehat{\boldsymbol{\theta}}_k^{(1)}$ *satisfies that*

$$\left\|\widehat{\boldsymbol{\theta}}_k^{(1)}-\boldsymbol{\theta}_k^*\right\|_\infty=O_{\mathbb{P}}\left(\left\|\boldsymbol{\Sigma}_{SS}^{-1}\right\|_\infty\lambda_1\right). \tag{1.8}$$

Moreover, suppose that there exists a sufficiently large constant $C>0$ *such that*

$$\theta_{\min}^*\geq C\left\|\boldsymbol{\Sigma}_{SS}^{-1}\right\|_\infty\lambda_1, \tag{1.9}$$

we have $\widehat{S}^{(1)}=S$ *with probability tending to 1.*

Lemma 1.4. *By partitioning* $\boldsymbol{\Sigma}$ *as*

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{SS} & \boldsymbol{\Sigma}_{SS^c}\\ \boldsymbol{\Sigma}_{S^cS} & \boldsymbol{\Sigma}_{S^cS^c}\end{pmatrix},$$

and $\boldsymbol{\mu}_k$ *according to sets* S *and* S^c *for* $k=2,...,K$ *respectively, we have*

$$\boldsymbol{\theta}_{k,S}^*=\boldsymbol{\Sigma}_{SS}^{-1}\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\right)_S, \tag{1.10}$$

and

$$\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\right)_{S^c}=\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}_{SS}^{-1}\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\right)_S. \tag{1.11}$$

Proof of Theorem 1.2. Here we only prove the results in System II and the proof for System I is similar. First we define the oracle sub-problem as

$$\widehat{\boldsymbol{\theta}}_S^o = \arg \min_{\boldsymbol{\theta}_{k,S^c}=\mathbf{0}} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \boldsymbol{\theta}_k - \widetilde{\boldsymbol{b}}_{k,g-1}^{\mathrm{T}} \boldsymbol{\theta}_k \right\} + \lambda_1 \sum_{j \in S} \left\| \boldsymbol{\theta}_{(j)} \right\|_2. \tag{1.12}$$

Once we show $\widehat{\boldsymbol{\theta}}_k^{(1)} = (\widehat{\boldsymbol{\theta}}_{k,S}^o, \mathbf{0})$ is the solution to (7), it is clear that $\widehat{S}^{(1)} \subseteq S$. According to the KKT condition, for any $j \in S$,

$$\begin{pmatrix} \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS} \widehat{\boldsymbol{\theta}}_{2,S}^o - (\widetilde{\boldsymbol{b}}_{2,0})_S \right)_j \\ \vdots \\ \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS} \widehat{\boldsymbol{\theta}}_{K,S}^o - (\widetilde{\boldsymbol{b}}_{K,0})_S \right)_j \end{pmatrix} + \lambda_1 \boldsymbol{Z}_j = 0, \tag{1.13}$$

and for any $j \notin S$,

$$\begin{pmatrix} \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS^c} \widehat{\boldsymbol{\theta}}_{2,S}^o - (\widetilde{\boldsymbol{b}}_{2,0})_{S^c} \right)_j \\ \vdots \\ \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS^c} \widehat{\boldsymbol{\theta}}_{K,S}^o - (\widetilde{\boldsymbol{b}}_{K,0})_{S^c} \right)_j \end{pmatrix} + \lambda_1 \boldsymbol{Z}_j = 0, \tag{1.14}$$

where $\boldsymbol{Z}_j \in \mathbb{R}^{K-1}$ is subgradient of $\|\boldsymbol{\theta}\|_2$ evaluated at $\widehat{\boldsymbol{\theta}}_{(j)}$. It suffices to show that

$$\lambda_1^{-1} \max_{j \in S^c} \left(\sum_{k=2}^{K} \left\{ \left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\theta}}_{k,S}^o \right)_j - (\widetilde{\boldsymbol{b}}_{k,0})_j \right\}^2 \right)^{1/2} < 1, \tag{1.15}$$

holds with probability tending to 1. From equation (1.13), we have $\widehat{\boldsymbol{\theta}}_{k,S}^o = \widehat{\boldsymbol{\Sigma}}_{(0),SS}^{-1}((\widetilde{\boldsymbol{b}}_{k,0})_S - \lambda_1 \widetilde{\boldsymbol{Z}}_{k,S})$ where $\widetilde{\boldsymbol{Z}}_{k,S} = (Z_{j,k} : j \in S) \in \mathbb{R}^s$ and $\sum_{k=2}^{K} (Z_{j,k})^2 = 1$. Note that

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\theta}}_{k,S}^o - (\widetilde{\boldsymbol{b}}_{k,0})_{S^c} \\
=&\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\Sigma}}_{(0),SS}^{-1} \left((\widetilde{\boldsymbol{b}}_{k,0})_S - \lambda_1 \widetilde{\boldsymbol{Z}}_{k,S} \right) - (\widetilde{\boldsymbol{b}}_{k,0})_{S^c} \\
=&\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\Sigma}}_{(0),SS}^{-1} \left\{ \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS} - \boldsymbol{\Sigma}_{SS} \right) \left(\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^* \right) + \boldsymbol{\Sigma}_{SS} \widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - (\widetilde{\boldsymbol{d}}_{k,0})_S + (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - \boldsymbol{\Sigma}_{SS} \boldsymbol{\theta}_{k,S}^* \right\} \\
&- \left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} - \boldsymbol{\Sigma}_{S^cS} \right) \left(\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^* \right) - \boldsymbol{\Sigma}_{SS} \widehat{\boldsymbol{\theta}}_{k,S}^{(0)} + (\widetilde{\boldsymbol{d}}_{k,0})_{S^c} + \boldsymbol{\Sigma}_{S^cS} \boldsymbol{\theta}_{k,S}^* - (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_{S^c} \\
&- \lambda_1 \widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\Sigma}}_{(0),SS}^{-1} \widetilde{\boldsymbol{Z}}_{k,S}.
\end{aligned}$$

We denote

$$I_1 = \widehat{\boldsymbol{\Sigma}}_{(0),S^cS} \widehat{\boldsymbol{\Sigma}}_{(0),SS}^{-1} \left\{ \left(\widehat{\boldsymbol{\Sigma}}_{(0),SS} - \boldsymbol{\Sigma}_{SS} \right) \left(\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^* \right) + \boldsymbol{\Sigma}_{SS} \widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - (\widetilde{\boldsymbol{d}}_{k,0})_S + (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - \boldsymbol{\Sigma}_{SS} \boldsymbol{\theta}_{k,S}^* \right\},$$

and

$$I_2 = \left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS} - \boldsymbol{\Sigma}_{S^cS} \right) \left(\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^* \right) + \boldsymbol{\Sigma}_{SS} \widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - (\widetilde{\boldsymbol{d}}_{k,0})_{S^c} - \boldsymbol{\Sigma}_{S^cS} \boldsymbol{\theta}_{k,S}^* + (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_{S^c}.$$

Observe that

$$\widehat{\mathbf{\Sigma}}_{(0),S^cS}\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1} = \left(\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right)\left(\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right) + \mathbf{\Sigma}_{S^cS}\left(\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right) + \mathbf{\Sigma}_{SS}^{-1}\left(\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right),$$

which implies

$$\begin{aligned}\left\|\widehat{\mathbf{\Sigma}}_{(0),S^cS}\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1}\right\|_\infty \le& \left\|\left(\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right)\left(\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right)\right\|_\infty \\ &+ \left\|\mathbf{\Sigma}_{S^cS}\left(\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right)\right\|_\infty + \left\|\mathbf{\Sigma}_{SS}^{-1}\left(\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right)\right\|_\infty \\ \le& s^{3/2}\left|\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right|_\infty \left\|\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right\|_2 \\ &+ s^{3/2}\left|\mathbf{\Sigma}_{S^cS}\right|_\infty \left\|\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right\|_2 + s^{3/2}\left|\mathbf{\Sigma}_{SS}^{-1}\right|_\infty \left\|\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}\right\|_2.\end{aligned}$$

Using the inequalities (58a) and (58b) in Wainwright [2009], we have

$$\left\|\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{SS}^{-1}\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{s}{n}}\right),$$

and

$$\left\|\widehat{\mathbf{\Sigma}}_{(0),SS} - \mathbf{\Sigma}_{SS}\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{s}{n}}\right).$$

Combining with the fact $|\widehat{\mathbf{\Sigma}}_{(0),S^cS} - \mathbf{\Sigma}_{S^cS}|_\infty = O_{\mathbb{P}}(\sqrt{\log p/n})$, it yields

$$\left\|\widehat{\mathbf{\Sigma}}_{(0),S^cS}\widehat{\mathbf{\Sigma}}_{(0),SS}^{-1} - \mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1}\right\|_\infty = O_{\mathbb{P}}\left(s^{3/2}\sqrt{\frac{\log p + s}{n}}\right).$$

Owing to the fact that $\boldsymbol{\theta}_{k,S} = \mathbf{\Sigma}_{SS}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_S$ in Lemma 1.4, we have

$$(\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - \mathbf{\Sigma}_{SS}\boldsymbol{\theta}_{k,S}^* = (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_S.$$

It yields that

$$\left\|(\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - \mathbf{\Sigma}_{SS}\boldsymbol{\theta}_{k,S}^*\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}}\right).$$

Moreover, note that

$$\left\|\left(\widehat{\mathbf{\Sigma}}_{(0),SS} - \mathbf{\Sigma}_{SS}\right)\left(\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^*\right)\right\|_\infty \le \left|\widehat{\mathbf{\Sigma}}_{(0),SS} - \mathbf{\Sigma}_{SS}\right|_\infty \left\|\widehat{\boldsymbol{\theta}}_{k,S}^{(0)} - \boldsymbol{\theta}_{k,S}^*\right\|_1 = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}a_n\right).$$

Then together with the assumption $\|\mathbf{\Sigma}_{S^cS}\mathbf{\Sigma}_{SS}^{-1}\|_\infty \le \xi$ and Lemma 1.2 we have

$$\|I_1\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + \sqrt{\frac{\log p}{n}}a_n + \frac{\alpha}{\sqrt{n}} + \frac{1}{n}\right).$$

Similarly, we can show that

$$\|I_2\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{N}} + \sqrt{\frac{\log p}{n}} a_n + \frac{\alpha}{\sqrt{n}} + \frac{1}{n}\right).$$

Owing to the choice of λ_1 in Theorem 1.1 and following the analysis above, there exists some positive constant C_1 such that for $k = 2, ..., K$,

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\theta}}^o_{k,S} - \boldsymbol{\theta}^*_{k,S}\right\|_\infty \\
\leq& \lambda_1 \left\|\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\widetilde{\boldsymbol{Z}}_{k,S}\right\|_\infty + \left\|\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\left(\widehat{\boldsymbol{\Sigma}}_{(0),SS} - \boldsymbol{\Sigma}_{SS}\right)\left(\widehat{\boldsymbol{\theta}}^{(0)}_{k,S} - \boldsymbol{\theta}^*_{k,S}\right)\right\|_\infty \\
+& \left\|\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\left(\boldsymbol{\Sigma}_{SS}\widehat{\boldsymbol{\theta}}^{(0)}_{k,S} - (\widetilde{\boldsymbol{d}}_{k,0})_S\right)\right\|_\infty + \left\|\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\left((\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)_S - \boldsymbol{\Sigma}_{SS}\boldsymbol{\theta}^*_{k,S}\right)\right\|_\infty \\
\leq& C_1\left\|\boldsymbol{\Sigma}^{-1}_{SS}\right\|_\infty \lambda_1,
\end{aligned} \tag{1.16}$$

holds with probability tending to 1. Moreover, for $j \in S^c$, we have

$$\begin{aligned}
\left|\left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS}\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\widetilde{\boldsymbol{Z}}_{k,S}\right)_j\right| \leq& \left\|\widehat{\boldsymbol{\Sigma}}_{(0),S^cS}\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS} - \boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\right\|_\infty \left\|\widetilde{\boldsymbol{Z}}_{k,S}\right\|_\infty \\
&+ \left\|\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\right\|_\infty \left\|\widetilde{\boldsymbol{Z}}_{k,S} - \boldsymbol{Z}^*_{k,S}\right\|_\infty + \left|\left(\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\boldsymbol{Z}^*_{k,S}\right)_j\right|,
\end{aligned}$$

and

$$\begin{aligned}
\|\widetilde{\boldsymbol{Z}}_{k,S} - \boldsymbol{Z}^*_{k,S}\|_\infty =& \max_{j\in S}\left|\frac{\widehat{\theta}^o_{kj}}{\|\boldsymbol{\theta}^o_{(j)}\|_2} - \frac{\theta^*_{kj}}{\|\boldsymbol{\theta}^*_{(j)}\|_2}\right| \\
\leq& \max_{j\in S}\frac{\left|\widehat{\theta}^o_{kj} - \theta^*_{kj}\right|}{\|\boldsymbol{\theta}^*_{(j)}\|_2} + \max_{j\in S}|\widehat{\theta}^o_{kj}|\frac{\left|\|\boldsymbol{\theta}^o_{(j)}\|_2 - \|\boldsymbol{\theta}^*_{(j)}\|_2\right|}{\|\boldsymbol{\theta}^o_{(j)}\|_2\|\boldsymbol{\theta}^*_{(j)}\|_2} \\
\leq& \max_{j\in S}\frac{\left|\widehat{\theta}^o_{kj} - \theta^*_{kj}\right|}{\|\boldsymbol{\theta}^*_{(j)}\|_2} + \max_{j\in S}\frac{\|\boldsymbol{\theta}^o_{(j)} - \boldsymbol{\theta}^*_{(j)}\|_2}{\|\boldsymbol{\theta}^*_{(j)}\|_2} \\
\lesssim& 2\max_{2\leq k\leq K}\|\widehat{\boldsymbol{\theta}}^o_{k,S} - \boldsymbol{\theta}^*_{k,S}\|_\infty/\theta^*_{\min}.
\end{aligned}$$

Combining with the Conditions (**C**2) and inequality (1.16), with probability tending to 1 we have

$$\begin{aligned}
&\lambda_1^{-2}\max_{j\in S^c}\sum_{k=2}^{K}\left\{\left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS}\widehat{\boldsymbol{\theta}}^o_{k,S}\right)_j - (\boldsymbol{b}_{k,g-1})_j\right\}^2 \leq \sum_{k=2}^{K}\left|\left(\widehat{\boldsymbol{\Sigma}}_{(0),S^cS}\widehat{\boldsymbol{\Sigma}}^{-1}_{(0),SS}\widetilde{\boldsymbol{Z}}_{k,S}\right)_j\right|^2 + o(1) \\
\leq& \max_{j\in S^c}\sum_{k=2}^{K}\left|\left(\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\boldsymbol{Z}^*_{k,S}\right)_j\right|^2 + C_1^2\left\|\boldsymbol{\Sigma}^{-1}_{SS}\right\|^2_\infty\left\|\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\right\|^2_\infty(K-1)\lambda_1^2/\theta^{*2}_{\min} + o(1) \\
\leq& 1 - \kappa + C_1^2\left\|\boldsymbol{\Sigma}^{-1}_{SS}\right\|^2_\infty\left\|\boldsymbol{\Sigma}_{S^cS}\boldsymbol{\Sigma}^{-1}_{SS}\right\|^2_\infty(K-1)\lambda_1^2/\theta^{*2}_{\min} + o(1) \\
\leq& 1 - \kappa/2,
\end{aligned} \tag{1.17}$$

then we have shown the inequality (1.15) holds. Recall that with inequality (1.16), we have

$$\left\|\widehat{\boldsymbol{\theta}}_k^o - \boldsymbol{\theta}_k^*\right\|_\infty \le C_1 \left\|\boldsymbol{\Sigma}_{SS}^{-1}\right\|_\infty \lambda_1,$$

holds with probability tending to 1. And note that $\widehat{\boldsymbol{\theta}}_k^o$ is a solution to (7) with probability tending to 1, that is $\mathbb{P}(\widehat{\boldsymbol{\theta}}_k^{(1)} = \widehat{\boldsymbol{\theta}}_k^o) \to 1$. It yields

$$\left\|\widehat{\boldsymbol{\theta}}_k^{(1)} - \boldsymbol{\theta}_k^*\right\|_\infty \le C_1 \left\|\boldsymbol{\Sigma}_{SS}^{-1}\right\|_\infty \lambda_1,$$

holds with probability tending to 1. If $\theta_{\min}^* \ge C\|\boldsymbol{\Sigma}_{SS}^{-1}\|_\infty \lambda_1$ for some sufficiently large positive constant C, then $\widehat{S}^{(1)} = S$ holds with probability tending to 1. In fact, the inequality (1.17) still holds if we choose sufficiently large C. Therefore, we have finished the proof of Theorem 1.2. □

2 Proof of Auxiliary Lemmas

2.1 Proof or Lemma 1.1

Proof or Lemma 1.1. With probability tending to 1, there exists some sufficiently large positive constant L such that

$$\begin{aligned}
\sum_{k=1}^{K-1} \boldsymbol{x}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{(0)} \boldsymbol{x}_k &\ge \sum_{k=1}^{K-1} \boldsymbol{x}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{x}_k - \left|\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right|_\infty \sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_1^2 \\
&\ge \sum_{k=1}^{K-1} \boldsymbol{x}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{x}_k - \left|\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right|_\infty \left(\sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_1\right)^2 \\
&\ge \lambda_{\min}(\boldsymbol{\Sigma}) \sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_2^2 - 16s(K-1)\left|\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right|_\infty \sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_2^2 \\
&\ge L \sum_{k=1}^{K-1} \|\boldsymbol{x}_k\|_2^2,
\end{aligned}$$

and the last inequality follows from the fact that $|\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}|_\infty = O_{\mathbb{P}}(\sqrt{\log p/n})$ and $s\sqrt{\log p/n} = o(1)$. □

2.2 Proof of Lemma 1.2

First note that,

$$\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \widehat{\boldsymbol{b}}_{k,0} = \left(\widehat{\boldsymbol{\Sigma}}_{(0)} - \widehat{\boldsymbol{\Sigma}}\right)\left(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*\right) - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k^* + (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1).$$

Due to the definition of $\widehat{\boldsymbol{\Sigma}}$, we have

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^* &= \frac{1}{N}\sum_{d=1}^{K}\sum_{\{i:Y_i=d\}}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_d)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^* \\ &= \frac{1}{N}\sum_{d=1}^{K}\sum_{\{i:Y_i=d\}}(\boldsymbol{X}_i - \boldsymbol{\mu}_d)(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^* + \frac{1}{N}\sum_{d=1}^{K}N_d(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^*.\end{aligned}$$

We note that $\boldsymbol{X}_i^{\top}\boldsymbol{\theta}_k^* \sim \mathcal{N}(\boldsymbol{\mu_k}^{\top}\boldsymbol{\theta}_k^*, (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1))$ for i such that $Y_i = k$ and $k \neq 1$, which yields that $|(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^*| = O_{\mathbb{P}}(\Delta_{\max}/\sqrt{N})$. Let $\boldsymbol{D}_{di} = (\boldsymbol{X}_i - \boldsymbol{\mu}_d)(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^*$, then $\boldsymbol{D}_{di,j}$ is sub-exponential variable with parameter $\sigma_{j,j}\Delta_k$. According to Bernstein's inequality for sub-exponential variable [Vershynin, 2018], we have

$$\left\|\frac{1}{N}\sum_{d=1}^{K}\sum_{\{i:Y_i=d\}}\boldsymbol{D}_{di}\right\|_{\infty} = O_{\mathbb{P}}\left(\Delta_{\max}\sqrt{\frac{\log p}{N}}\right).$$

It follows that

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^*\right\|_{\infty} &\leq \left\|\frac{1}{N}\sum_{d=1}^{K}\sum_{\{i:Y_i=d\}}\boldsymbol{D}_{di}\right\|_{\infty} + \max_d \left\|(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^*\right\|_{\infty} \\ &\leq \left\|\frac{1}{N}\sum_{d=1}^{K}\sum_{\{i:Y_i=d\}}\boldsymbol{D}_{di}\right\|_{\infty} + \max_d \|(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)\|_{\infty}\,|(\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^{\mathrm{T}}\boldsymbol{\theta}_k^*| \\ &\lesssim \Delta_{\max}\sqrt{\frac{\log p}{N}} + \sqrt{\frac{\log p}{N}}\frac{\Delta_{\max}}{\sqrt{N}}\end{aligned}$$

with high probability. It yields that

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \widehat{\boldsymbol{b}}_{k,0}\right\|_{\infty} &\leq \left|\widehat{\boldsymbol{\Sigma}}_{(0)} - \widehat{\boldsymbol{\Sigma}}\right|_{\infty}\left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k^{(0)}\right\|_1 + \left\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k^* - (\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1)\right\|_{\infty} \\ &\leq \left|\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right|_{\infty}\left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k^{(0)}\right\|_1 + \left|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right|_{\infty}\left\|\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k^{(0)}\right\|_1 \\ &\quad + \left\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k^* - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^*\right\|_{\infty} + \|(\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)\|_{\infty} \\ &\lesssim \sqrt{\frac{\log p}{n}}a_n + \sqrt{\frac{\log p}{N}}a_n + \Delta_{\max}\sqrt{\frac{\log p}{N}} + \sqrt{\frac{\log p}{N}} \\ &= O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}a_n + \sqrt{\frac{\log p}{N}}\right),\end{aligned}$$

where the third inequality follows from the basic bound $\|\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}\|_{\infty} = O_{\mathbb{P}}(\sqrt{\log p/N})$.

2.3 Proof of Lemma 1.3

Lemma 2.1 (Berry-Esseen inequality [Petrov, 1975])**.** *Let $X_1, \cdots, X_n$ are i.i.d random variables and suppose*

$$\mathbb{E}X_1 = 0, \quad \mathbb{E}X_1^2 = \sigma^2 > 0, \quad \mathbb{E}\,|X_1|^3 < \infty, \quad \varrho = \frac{\mathbb{E}\,|X_1|^3}{\sigma^3}.$$

Then for some absolute positive constants A

$$\sup_x \left| \mathbb{P}\left(\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_j < x \right) - \Phi(x) \right| \leqq A \frac{\varrho}{\sqrt{n}}.$$

Proof of Lemma 1.3. Denote the Byzantine local machines by $\mathcal{B}$ and $|\mathcal{B}| = \alpha M$. Let $Y_l = \sqrt{n}\left(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\right)_l / \sqrt{\sigma_{ll}}$ then

$$Y_l = \operatorname{med}\left\{Y_{l,0}, Y_{l,1}, \cdots, Y_{l,M}\right\}.$$

where $Y_{l,m} = \sqrt{n}(\widehat{\boldsymbol{\mu}}_1^{(m)} - \boldsymbol{\mu}_1)_l / \sqrt{\sigma_{ll}} \sim N(0,1)$ if $m \notin \mathcal{B}$. Using the uniform bound and the fact that for any $t \in \mathbb{R}$

$$\left| \frac{1}{M+1} \sum_{m=0}^{M} \mathbb{I}(Y_{l,m} \geq t) - \frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} \mathbb{I}(Y_{l,m} \geq t) \right| \leq \alpha,$$

we have

$$\begin{aligned}
&\mathbb{P}\left(\max_l \left| \frac{\sqrt{n}}{\sigma_{ll}} \left(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\right)_l \right| \geq u_n \right) \\
\leq & p \max_l \mathbb{P}\left(\left| \frac{\sqrt{n}}{\sigma_{ll}} \left(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\right)_l \right| \geq u_n \right) \\
= & p \max_l \mathbb{P}\left(\frac{1}{M+1} \sum_{m=0}^{M} \mathbb{I}(Y_{l,m} \geq u_n) \geq \frac{1}{2} \right) + p \max_l \mathbb{P}\left(\frac{1}{M+1} \sum_{m=0}^{M} \mathbb{I}(Y_{l,m} \leq -u_n) \leq \frac{1}{2} \right) \\
= & p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} \mathbb{I}(Y_{l,m} \geq u_n) - \mathbb{P}(Y_{l,m} \geq u_n) \geq \frac{1}{2} - \alpha - (1 - \Phi(u_n)) \right) \\
+ & p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} \mathbb{I}(Y_{l,m} \leq -u_n) - \mathbb{P}(Y_{l,m} \leq -u_n) \leq \frac{1}{2} + \alpha - \Phi(-u_n) \right).
\end{aligned}$$

By Taylor expansion we have

$$\Phi(u_n) = \Phi(0) + \phi(0)u_n + o(u_n).$$

Thus

$$\begin{aligned}
&\mathbb{P}\left(\max_l \left| \frac{\sqrt{n}}{\sigma_{ll}} \left(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\right)_l \right| \geq u_n \right) \\
\leq & p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} \mathbb{I}(Y_{l,m} \geq u_n) - \mathbb{P}(Y_{l,m} \geq u_n) \geq \phi(0)u_n + o(u_n) - \alpha \right) \\
+ & p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} \mathbb{I}(Y_{l,m} \leq -u_n) - \mathbb{P}(Y_{l,m} \leq -u_n) \leq \phi(0)u_n + o(u_n) + \alpha \right).
\end{aligned}$$

Let $u_n = \rho^{'}(\sqrt{\log p/((1-\alpha)M+1)} + \alpha)$ for some sufficiently large positive constant $\rho^{'}$ and using Bernstein's inequality we have

$$\mathbb{P}\left(\max_{l} \frac{\sqrt{n}}{\sigma_{ll}} \left(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\right)_l \geq u_n\right) \leq 2p^{-1}.$$

And this means that

$$\|\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|_\infty \lesssim \sqrt{\frac{\log p}{N}} + \frac{\alpha}{\sqrt{n}},$$

holds with at least probability $1-2p^{-1}$. Similarly, we can prove

$$\|\widetilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_\infty \lesssim \sqrt{\frac{\log p}{N}} + \frac{\alpha}{\sqrt{n}},$$

holds with at least probability $1-2p^{-1}$. For the second inequality, note that

$$\left(\widetilde{\boldsymbol{d}}_{k,0}\right)_l = \operatorname{med}\left\{\frac{1}{n}\sum_{d=1}^{K}\sum_{\{i\in\mathcal{H}_m: Y_i=d\}} (X_{il} - \widehat{\mu}_{dl}^{(m)})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_d^{(m)})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)} :\ m = 1, 2, \cdots, M\right\},$$

where X_{il} is the $l-$th entry of $\boldsymbol{X}_i$, $\widehat{\mu}_{dl}^{(m)}$ is the $l-$th entry of $\widehat{\boldsymbol{\mu}}_d^{(m)}$. By straightforward calculation we can write

$$\begin{aligned}
&\frac{1}{n}\sum_{d=1}^{K}\sum_{\{i\in\mathcal{H}_m: Y_i=d\}} (X_{il} - \widehat{\mu}_{dl}^{(m)})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_d^{(m)})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)} \\
=&\frac{1}{n}\sum_{d=1}^{K}\sum_{\{i\in\mathcal{H}_m: Y_i=d\}} (X_{il} - \mu_{dl})(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)} + \frac{1}{n}\sum_{d=1}^{K} n_k(\widehat{\mu}_{dl}^{(m)} - \mu_{dl})(\widehat{\boldsymbol{\mu}}_d^{(m)} - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)},
\end{aligned}$$

and for $m \notin \mathcal{B}$

$$(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)} \sim \mathcal{N}\left(0, (\widehat{\boldsymbol{\theta}}_k^{(0)})^{\mathrm{T}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_k^{(0)})\right), \quad i \in \mathcal{H}_m$$

where $(\widehat{\boldsymbol{\theta}}_k^{(0)})^{\mathrm{T}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_k^{(0)}) \lesssim \boldsymbol{\theta}_k^{*T}\boldsymbol{\Sigma}\boldsymbol{\theta}_k^* \leq \Delta_{\max}^2$. Conditioning on $\widehat{\boldsymbol{\theta}}_k^{(0)}$, we have

$$\mathbb{E}\left[(X_{il} - \mu_{dl})(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)}\middle|\widehat{\boldsymbol{\theta}}_k^{(0)}\right] = \left(\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)_l,$$

and

$$\begin{aligned}
\widetilde{\sigma}_l^2 &:= \operatorname{Var}\left[(X_{il} - \mu_{dl})(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)}\middle|\widehat{\boldsymbol{\theta}}_k^{(0)}\right] \\
&\leq \left(\mathbb{E}(X_{il} - \mu_{dl})^4\right)^{1/2}\left(\mathbb{E}\left((\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)^4\right)^{1/2} \lesssim 3\sigma_{ll}^2\Delta_k^2,
\end{aligned}$$

for $i \in \mathcal{H}_m$ and $m \notin \mathcal{B}$ and $\widetilde{\sigma}_l^2 < \infty$ according to assumption. Let

$$W_{l,m} = \frac{1}{\sqrt{n}}\sum_{d=1}^{K}\sum_{\{i\in\mathcal{H}_m: Y_i=d\}} (X_{il} - \mu_{dl})(\boldsymbol{X}_i - \boldsymbol{\mu}_d)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_k^{(0)} - \left(\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)_l,$$

and

$$V_{l,m} = \frac{1}{\sqrt{n}} \sum_{d=1}^{K} n_k (\widehat{\mu}_{dl}^{(m)} - \mu_{dl})(\widehat{\boldsymbol{\mu}}_d^{(m)} - \boldsymbol{\mu}_d)^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_k^{(0)},$$

then for $m \notin \mathcal{B}$

$$W_{l,m} \xrightarrow{d} N\left(0, \widetilde{\sigma}_l^2\right) \text{ and } V_{l,m} = O_{\mathbb{P}}\left(\frac{\widetilde{\sigma}_l}{\sqrt{n}}\right).$$

Denote

$$Z_{l,m} = \mathbb{I}(W_{l,m} + V_{l,m} \geq u_n) - \mathbb{P}(W_{l,m} + V_{l,m} \geq u_n),$$

and

$$Z_{l,m'} = \mathbb{I}(W_{l,m} + V_{l,m} \leq -u_n) - \mathbb{P}(W_{l,m} + V_{l,m} \leq -u_n).$$

Owing to the definition of sample median, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_l \sqrt{n}\left|\left(\widetilde{\boldsymbol{d}}_{k,0} - \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)_l\right| \geq u_n\right) \leq p \max_l \mathbb{P}\left(\sqrt{n}\left|\left(\widetilde{\boldsymbol{d}}_{k,0} - \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)_l\right| \geq u_n\right) \\
=& p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} Z_{l,m} \geq \frac{1}{2} - \alpha - \mathbb{P}(W_{l,m} + V_{l,m} \geq u_n)\right) \\
+& p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} Z_{l,m'} \leq \frac{1}{2} + \alpha - \mathbb{P}(W_{l,m} + V_{l,m} \leq -u_n)\right).
\end{aligned}$$

Using the fact $\mathbb{P}(W_{l,m} + V_{l,m} \leq u_n) = \mathbb{P}(W_{l,m}/\widetilde{\sigma}_l + V_{l,m}/\widetilde{\sigma}_l \leq u_n/\theta_l)$ we have

$$\begin{aligned}
&\left|\mathbb{P}\left(\frac{W_{l,m}}{\widetilde{\sigma}_l} + \frac{V_{l,m}}{\widetilde{\sigma}_l} \leq \frac{u_n}{\widetilde{\sigma}_l}\right) - \Phi\left(\frac{u_n}{\widetilde{\sigma}_l}\right)\right| \\
\leq& \left|\mathbb{P}\left(\frac{W_{l,m}}{\widetilde{\sigma}_l} \leq \frac{u_n}{\widetilde{\sigma}_l} - \frac{V_{l,m}}{\widetilde{\sigma}_l}\right) - \Phi\left(\frac{u_n}{\widetilde{\sigma}_l} - \frac{V_{l,m}}{\widetilde{\sigma}_l}\right)\right| + \left|\Phi\left(\frac{u_n}{\widetilde{\sigma}_l} - \frac{V_{l,m}}{\widetilde{\sigma}_l}\right) - \Phi\left(\frac{u_n}{\widetilde{\sigma}_l}\right)\right| \\
\leq& \sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\frac{W_{l,m}}{\widetilde{\sigma}_l} \leq x\right) - \Phi(x)\right| + \left|\Phi\left(\frac{u_n}{\widetilde{\sigma}_l} - \frac{V_{l,m}}{\widetilde{\sigma}_l}\right) - \Phi\left(\frac{u_n}{\widetilde{\sigma}_l}\right)\right| \\
=& \sup_{x \in \mathbb{R}} \left|\mathbb{P}\left(\frac{W_{l,m}}{\widetilde{\sigma}_l} \leq x\right) - \Phi(x)\right| + \phi\left(\frac{u_n}{\widetilde{\sigma}_l}\right)\frac{V_{l,m}}{\widetilde{\sigma}_l} + o(V_{l,m}) \\
\lesssim& \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}},
\end{aligned}$$

where the last inequality follows from Berry-Esseen inequality and the normal density function $\phi(x)$ is bounded. It yields that

$$\begin{aligned}
&\mathbb{P}\left(\max_l \sqrt{n}\left|\left(\widetilde{\boldsymbol{d}}_{k,0} - \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right)_l\right| \geq u_n\right) \\
\leq& p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} Z_{l,m} \geq \phi(0)u_n/\widetilde{\sigma}_l - \alpha + O(\frac{1}{\sqrt{n}}) + o(u_n)\right) \\
+& p \max_l \mathbb{P}\left(\frac{1}{(1-\alpha)M+1} \sum_{m \notin \mathcal{B}} Z_{l,m'} \leq \phi(0)u_n/\widetilde{\sigma}_l + \alpha + O(\frac{1}{\sqrt{n}}) + o(u_n)\right).
\end{aligned}$$

Let $u_n = \rho^{''}(\Delta_{\max}\sqrt{\log p/(M+1)} + 1/\sqrt{n} + \alpha)$ for some sufficiently large positive constant $\rho^{''}$, then by Bernstein's inequality we can prove that

$$\left\|\widetilde{\boldsymbol{d}}_{k,0} - \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right\|_\infty = O_{\mathbb{P}}\left(\Delta_{\max}\sqrt{\frac{\log p}{N}} + \frac{\alpha}{\sqrt{n}} + \frac{1}{n}\right).$$

By the definition of $\widetilde{\boldsymbol{b}}_{k,0}$, we have

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\Sigma}}_{(0)}(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*) - \widetilde{\boldsymbol{b}}_{k,0}\right\|_\infty \\
=&\left\|\left(\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right)\left(\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\theta}_k^*\right) - \widetilde{\boldsymbol{d}}_{k,0} + \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)} - \boldsymbol{\Sigma}\boldsymbol{\theta}_k^* + (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)\right\|_\infty \\
\leq&\left\|\left(\widehat{\boldsymbol{\Sigma}}_{(0)} - \boldsymbol{\Sigma}\right)\left(\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k^{(0)}\right)\right\|_\infty + \left\|\widetilde{\boldsymbol{d}}_{k,0} - \boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_k^{(0)}\right\|_\infty + \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1 - (\widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_1)\|_\infty,
\end{aligned}$$

then the results follow. □

2.4 Proof of Lemma 1.4

Proof of Lemma 1.4. By the definition of the support set S and $\boldsymbol{\Sigma}\boldsymbol{\theta}_k^* = \boldsymbol{\mu}_k - \boldsymbol{\mu}_1$ we have

$$\begin{pmatrix}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_S \\ (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_{S^c}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{SS} & \boldsymbol{\Sigma}_{SS^c} \\ \boldsymbol{\Sigma}_{S^cS} & \boldsymbol{\Sigma}_{S^cS^c}\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}_{k,S}^* \\ \mathbf{0}\end{pmatrix},$$

then the results follow immediately. □

References

V. V. Petrov. *Sums of Independent Random Variables*, volume 82. Springer, New York, 1975.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.