Supplement to "CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality"*

T. Tony Cai, Jing Ma, and Linjun Zhang

University of Pennsylvania

April 22, 2018

Abstract

This note summarizes the supplementary materials to the paper "CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and Its Optimality". Section A shows the proof of Theorem 3.1, which is the upper bound of the estimation error for $\boldsymbol{\beta}^*$. The lower bound of this estimation error, i.e. part (2) in Theorem 3.3, is proved in Section B. Section C states and proves the technical lemmas used in the proofs of the main results given in the paper and Section D provides additional simulation results.

A Proof of Theorem 3.1

A.1 Auxiliary lemmas

We begin by stating a lemma that is used in the proof of Theorem 3.1.

Lemma A.1. *Suppose* $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$ *and* **(C1)**, **(C2)** *hold. If* $\lambda_n^{(t+1)} \geq 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|}{\sqrt{s}}$, *where* C_{con}, κ_0 *are defined in* **(C1)**, **(C2)** *and* $\hat{\boldsymbol{\beta}}^{(t+1)}$ *is solved by*

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} \boldsymbol{\beta}^\top \hat{\Sigma}^{(t+1)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)} \|\boldsymbol{\beta}\|_1 \right\},$$

then

1. *For* $\boldsymbol{u}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*$, *we have* $\boldsymbol{u}^{(t+1)} \in \Gamma(s)$, *that is,*

$$2\|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 \leq 4\|\boldsymbol{u}_S^{(t+1)}\|_1 + 3\sqrt{s}\|\boldsymbol{u}^{(t+1)}\|_2.$$

[*] The research was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

2.

$$\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \le 4M \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*) + 2M\sqrt{s}\lambda_n^{(t+1)}.$$

The proof of Lemma A.1 is given in Section A.3. Recall that $M_n(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$ are defined in (2.4) and (3.5) respectively.

Lemma A.2. *If* $d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \le r\Delta$, $\boldsymbol{\beta} - \boldsymbol{\beta}^* \in \Gamma(s)$*, and* $r < \frac{|c_0 - c_\omega|}{\Delta} \wedge \frac{\sqrt{9M+16c_1} - \sqrt{9M}}{4} \wedge \sqrt{\frac{c_1}{M}} \wedge \frac{C_b}{5\sqrt{s}}$*, then*

$$\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s).$$

A.2 Proof of Theorem 3.1

A.2.1 Preliminaries

Recall that we update $\lambda_n^{(t)}$ by $\lambda_n^{(t)} = \kappa\lambda_n^{(t-1)} + C_\lambda\sqrt{\frac{\log p}{n}}$ and $\lambda_n^{(0)} = C_1 \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} + C_\lambda\sqrt{\frac{\log p}{n}}$ with $C_1 = \frac{1}{4M}$, so we have

$$\lambda_n^{(t)} = \kappa^t \cdot C_1 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} + \frac{1 - \kappa^{t+1}}{1 - \kappa} C_\lambda \sqrt{\frac{\log p}{n}}.$$

Let $\kappa = (1 \vee (8M)) \cdot \kappa_0$, since $0 < \kappa_0 < \frac{1}{2 \vee (16M)}$, we have $\kappa \in (0, 1/2)$. Now let us define

$$C^* = \left[\left(\frac{2\kappa^2 - 4\kappa + 2}{2\kappa^2 - 5\kappa + 2} \cdot (4M + \frac{6M}{1-\kappa})\right) \vee \frac{1-\kappa}{1-2\kappa} \right] C_{con},$$

and

$$C_\lambda = 3C_{con} + \frac{2\kappa_0}{1-\kappa} C^*.$$

We claim

(i) $\kappa \ge \kappa_0$, $C_1\kappa \ge 2\kappa_0$, $M\kappa_0 + 2MC_1\kappa \le \kappa$.

(ii) $\frac{\kappa_0}{1-\kappa}C^* + C_{con} \le C^*$, and $4MC_{con} + \frac{2M}{1-\kappa}C_\lambda \le C^*$.

In fact, the three inequalities in (i) can be seen from the definition of C_1 and κ.

For the first inequality in (ii), it's equivalent to $C_{con} \le \frac{1-\kappa-\kappa_0}{1-\kappa}C^*$. By the fact that $\kappa_0 \le \kappa$ and the definition of C^*, we then have

$$\frac{1-\kappa-\kappa_0}{1-\kappa}C^* \ge \frac{1-2\kappa}{1-\kappa}C^* \ge C_{con}.$$

For the second inequality in (ii), we have

$$\begin{aligned}4MC_{con} + \frac{2M}{1-\kappa}C_\lambda =& 4MC_{con} + \frac{2M}{1-\kappa}(3C_{con} + \frac{2\kappa_0}{1-\kappa}C^*)\\ =&(4M + \frac{6M}{1-\kappa})C_{con} + \frac{4M\kappa_0}{(1-\kappa)^2}C^*\end{aligned}$$

By noticing $4M\kappa_0 = \frac{\kappa_0}{C_1} \leq \frac{\kappa}{2}$ and the definition of C^*, we then have

$$4MC_{con} + \frac{2M}{1-\kappa}C_\lambda \leq \frac{2\kappa^2 - 5\kappa + 2}{2\kappa^2 - 4\kappa + 2}C^* + \frac{1}{2(1-\kappa)^2}C^* = C^*.$$

Further more, since $\sqrt{s\log p/n} = o(r)$, we have for sufficiently large n,

$$r \geq \frac{1+\kappa}{1-\kappa}(C^* + 3\frac{C_{con}}{C_1})\sqrt{\frac{s\log p}{n}}. \tag{A.1}$$

We then divide the proof into two main steps.

A.2.2 Main proofs

We use induction to show that

$$\lambda_n^{(t+1)} \geq 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|}{\sqrt{s}},$$

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\| \leq \kappa^{t+1} \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}},$$

and

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\| \leq r\Delta, \quad \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \in \Gamma(s),$$

where the last two inequalities, by Lemma A.2, implies that when r satisfies the condition in **(C1)**

$$\hat{\boldsymbol{\theta}}^{(t+1)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s).$$

According to the conditions in Theorem 3.1 we have **(C1)**, that is, $d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\| \leq r\Delta$ and $\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^* \in \Gamma(s)$. Then by Lemma A.2, we have $\hat{\boldsymbol{\theta}}^{(0)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$. Then it follows from Lemma 3.1 and Lemma 3.2 that

$$\begin{aligned}d_{2,s}(\hat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{\theta}^*) \leq& d_{2,s}(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^*) + C_{con}\sqrt{\frac{s\log p}{n}}\\ \leq&\kappa_0 \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + C_{con}\sqrt{\frac{s\log p}{n}}\\ \leq&\kappa \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}},\end{aligned}$$

since $\kappa_0 \le \kappa$, $C_{con} \le \frac{\kappa_0}{1-\kappa}C^* + C_{con} \le C^* \le (1+\kappa)C^*$

This also implies

$$\begin{aligned}\lambda_n^{(1)} =& \kappa\lambda_n^{(0)} + C_\lambda\sqrt{\frac{\log p}{n}} = C_1\kappa\frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} + (\kappa C_\lambda + C_\lambda)\sqrt{\frac{\log p}{n}}\\ \ge& 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}},\end{aligned}$$

since $\kappa C_\lambda + C_\lambda \ge 3C_{con}$, $C_1\kappa \ge 2\kappa_0$.

Moreover, by Lemma A.1,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_2 \le& 4M \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{\theta}^*) + 2M\sqrt{s}\lambda_n^{(1)}\\ \le& 4M \cdot (\kappa_0 \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + C_{con}\sqrt{\frac{s\log p}{n}})\\ &+ 2M \cdot C_1\kappa \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\| + 2M(1+\kappa)C_\lambda\sqrt{\frac{s\log p}{n}}\\ =& (4M\kappa_0 + 2MC_1\kappa) \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + (4MC_{con} + 2M(1+\kappa)C_\lambda)\sqrt{\frac{s\log p}{n}}\\ \le& \kappa \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}}.\end{aligned}$$

since $4M\kappa_0 + 2MC_1\kappa \le \kappa$, $4MC_{con} + 2M(1+\kappa)C_\lambda \le (1+\kappa)C^*$.

Therefore, we have

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\| \le \kappa \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^2}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}}.$$

In addition, since $d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\| \le r$,

$$\begin{aligned}d_{2,s}(\hat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\| \le& \kappa \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + (1+\kappa)C^*\sqrt{\frac{s\log p}{n}}\\ \le& \kappa r + (1+\kappa)C^*\sqrt{\frac{s\log p}{n}}\\ \le& r,\end{aligned}$$

since $r \ge \frac{1+\kappa}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}}$ by (A.1).

Therefore the properties hold at the $t =$ 1-st step.

Now let us assume the properties hold at the t-th step. That is,

$$\lambda_n^{(t)} \ge 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(t-1)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\|}{\sqrt{s}},$$

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\| \leq \kappa^t \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^{t+1}}{1-\kappa} C^* \sqrt{\frac{s \log p}{n}},$$

and

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\| \leq r\Delta, \quad \hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* \in \Gamma(s),$$

which by Lemma A.2 implies

$$\hat{\boldsymbol{\theta}}^{(t)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s).$$

Then

$$\begin{aligned}
&3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} \\
\leq &3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{\kappa^t \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^{t+1}}{1-\kappa} C^* \sqrt{\frac{s \log p}{n}}}{\sqrt{s}} \\
= &2\kappa_0\kappa^t \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} + 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{1-\kappa^{t+1}}{1-\kappa} C^* \sqrt{\frac{\log p}{n}} \\
\leq &\kappa^{t+1} \cdot C_1 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|}{\sqrt{s}} + \frac{1-\kappa^{t+2}}{1-\kappa} C_\lambda \sqrt{\frac{\log p}{n}} \\
= &\lambda_n^{(t+1)}
\end{aligned}$$

since $\frac{1-\kappa^{t+2}}{1-\kappa} C_\lambda \geq 3C_{con} + 2\kappa_0 \frac{1-\kappa^{t+1}}{1-\kappa} C^*$, $C_1\kappa \geq 2\kappa_0$.

Further, use Lemmas A.1, 3.1 and Lemma 3.2, we have

$$\begin{aligned}
d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*) \leq &d_{2,s}(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^*) + C_{con}\sqrt{\frac{s \log p}{n}} \\
\leq &\kappa_0 \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|) + C_{con}\sqrt{\frac{s \log p}{n}} \\
\leq &\kappa_0 \cdot (\kappa^t \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^{t+1}}{1-\kappa} C^* \sqrt{\frac{s \log p}{n}}) + C_{con}\sqrt{\frac{s \log p}{n}} \\
\leq &\kappa_0 \cdot \kappa^t \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + (\kappa_0 \frac{1-\kappa^{t+1}}{1-\kappa} C^* + C_{con})\sqrt{\frac{s \log p}{n}} \\
\leq &\kappa^{t+1} \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|) + \frac{1-\kappa^{t+2}}{1-\kappa} C^* \sqrt{\frac{s \log p}{n}}
\end{aligned}$$

since $\kappa_0 \leq \kappa$, $\kappa_0 \frac{1-\kappa^{t+1}}{1-\kappa} C^* + C_{con} \leq \frac{1-\kappa^{t+2}}{1-\kappa} C^*$.

Moreover, by Lemma A.1

$$
\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(t+1)}-\boldsymbol{\beta}^*\|_2 \leq & 4M\cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)},\boldsymbol{\theta}^*)+2M\sqrt{s}\lambda_n^{(t+1)} \\
\leq & 4M\cdot(\kappa_0\cdot\kappa^t\cdot(d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|)+(\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^*+C_{con})\sqrt{\frac{s\log p}{n}}) \\
& +2M\sqrt{s}\cdot(\kappa^{t+1}\cdot C_1\cdot\frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|}{\sqrt{s}}+\frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda\sqrt{\frac{\log p}{n}}) \\
\leq & (4M\kappa_0+2MC_1\kappa)\kappa^t\cdot(d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|) \\
& +(\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^*+4MC_{con}+2M\frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda)\sqrt{\frac{s\log p}{n}} \\
\leq & \kappa^{t+1}\cdot(d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|)+\frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}},
\end{aligned}
$$

since $4M\kappa_0+2MC_1\kappa\leq\kappa$, $\kappa_0\frac{1-\kappa^{t+1}}{1-\kappa}C^*+4MC_{con}+2M\frac{1-\kappa^{t+2}}{1-\kappa}C_\lambda\leq\frac{1-\kappa^{t+2}}{1-\kappa}C^*$

Further,

$$
\begin{aligned}
d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(t+1)}-\boldsymbol{\beta}^*\| \leq & \kappa^{t+1}\cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|+\frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}} \\
\leq & \kappa^{t+1}\cdot r+\frac{1-\kappa^{t+2}}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}} \\
\leq & r,
\end{aligned}
$$

since $r\geq\frac{1+\kappa}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}}$.

Now we complete the induction and have

$$
d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^*\|\leq\kappa^t\cdot(d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|)+\frac{1-\kappa^{t+1}}{1-\kappa}C^*\sqrt{\frac{s\log p}{n}}.
$$

Therefore, when $T_0\gtrsim\frac{\log n+\log d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)}{-\log\kappa}$,

$$
\|\hat{\boldsymbol{\beta}}^{(T_0)}-\boldsymbol{\beta}^*\|_2\lesssim\sqrt{\frac{s\log p}{n}}.
$$

□

A.3 Proof of Lemma A.1

At first, since

$$
\hat{\boldsymbol{\beta}}^{(t+1)}=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\boldsymbol{\beta}^\top\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}-\boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1^{(t+1)}-\hat{\boldsymbol{\mu}}_2^{(t+1)})+\lambda_n^{(t+1)}\|\boldsymbol{\beta}\|_1\right\},
$$

by letting $\hat{\boldsymbol{\delta}}^{(t+1)} = \hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}$, we have

$$\frac{1}{2}\hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\Sigma}^{(t+1)}\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\boldsymbol{\delta}}^{(t+1)} + \lambda_n^{(t+1)}\|\hat{\boldsymbol{\beta}}^{(t+1)}\|_1 \le \frac{1}{2}\boldsymbol{\beta}^{*\top}\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*\top}\hat{\boldsymbol{\delta}}^{(t+1)} + \lambda_n^{(t+1)}\|\boldsymbol{\beta}^*\|_1,$$

which implies

$$\lambda_n^{(t+1)}(\|\hat{\boldsymbol{\beta}}^{(t+1)}\|_1 - \|\boldsymbol{\beta}^*\|_1) \le \frac{1}{2}\boldsymbol{\beta}^{*\top}\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*\top}\hat{\boldsymbol{\delta}}^{(t+1)} - (\frac{1}{2}\hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\Sigma}^{(t+1)}\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\boldsymbol{\delta}}^{(t+1)}), \tag{A.2}$$

Let $\boldsymbol{u}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*$ and recall that $S = \text{supp}(\boldsymbol{\beta})$, then we have

$$\|\hat{\boldsymbol{\beta}}^{(t+1)}\|_1 - \|\boldsymbol{\beta}^*\|_1 \ge \|\boldsymbol{\beta}^* + \boldsymbol{u}_{S^c}^{(t+1)}\|_1 - \|\boldsymbol{u}_S^{(t+1)}\|_1 - \|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 - \|\boldsymbol{u}_S^{(t+1)}\|_1.$$

In addition, by Lemma 3.2, we have with probability $1-o(1)$, $\|\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \Sigma^{(t+1)}\boldsymbol{\beta}^*\|_\infty \le C_{con}\sqrt{\frac{\log p}{n}}$, and it's followed by

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\beta}^{*\top}\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*\top}\hat{\boldsymbol{\delta}}^{(t+1)} - (\frac{1}{2}\hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\Sigma}^{(t+1)}\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\boldsymbol{\delta}}^{(t+1)})\\
=&-\langle\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle - \frac{1}{2}\langle\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*, \hat{\Sigma}^{(t+1)}(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*)\rangle\\
\le&-\langle\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle\\
\le&|\langle\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \Sigma^{(t+1)}\boldsymbol{\beta}^* + \boldsymbol{\delta}^{(t+1)} - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle| + |\langle\Sigma^{(t+1)}\boldsymbol{\beta}^* - \boldsymbol{\delta}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle|\\
\le&|\langle\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \Sigma^{(t+1)}\boldsymbol{\beta}^* + \boldsymbol{\delta}^{(t+1)} - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle| + |\langle(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^* + (\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)}), \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle|\\
\le&C_{con}\sqrt{\frac{\log p}{n}}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_1 + |\langle(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^* + (\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)}), \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\rangle|\\
=&C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + |\langle(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^* + (\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)}), \boldsymbol{u}^{(t+1)}\rangle|\\
=&C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + \|\langle(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^* + (\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)})\| \cdot \|\boldsymbol{u}^{(t+1)}\|\\
=&C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + (\|(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^*\|_2 + \|\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)})\|_2) \cdot \|\boldsymbol{u}^{(t+1)}\|\\
\le&C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + 2d_{2,s}(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^*) \cdot \|\boldsymbol{u}^{(t+1)}\|\\
\le&C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + 2\kappa_0 \cdot (d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|) \cdot \|\boldsymbol{u}^{(t+1)}\|,
\end{aligned}$$

where the last inequality uses Lemma 3.1.

Then we have

$$\lambda_n^{(t+1)}(\|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 - \|\boldsymbol{u}_S^{(t+1)}\|_1) \le C_{con}\sqrt{\frac{\log p}{n}}\|\boldsymbol{u}^{(t+1)}\|_1 + |\langle(\Sigma^{(t+1)} - \Sigma^*)\boldsymbol{\beta}^* + (\boldsymbol{\delta}^* - \boldsymbol{\delta}^{(t+1)}), \boldsymbol{u}^{(t+1)}\rangle|.$$

If $\lambda_n^{(t+1)}$ satisfies

$$\lambda_n^{(t+1)} \geq 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|}{\sqrt{s}},$$

we then have

$$\|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 - \|\boldsymbol{u}_S^{(t+1)}\|_1 \leq \frac{1}{3}\|\boldsymbol{u}^{(t+1)}\|_1 + \sqrt{s}\|\boldsymbol{u}^{(t+1)}\|_2 = \frac{1}{3}(\|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 + \|\boldsymbol{u}_S^{(t+1)}\|_1) + \sqrt{s}\|\boldsymbol{u}^{(t+1)}\|_2,$$

This implies

$$2\|\boldsymbol{u}_{S^c}^{(t+1)}\|_1 \leq 4\|\boldsymbol{u}_S^{(t+1)}\|_1 + 3\sqrt{s}\|\boldsymbol{u}^{(t+1)}\|_2,$$

that is, $\boldsymbol{u}^{(t+1)} \in \Gamma(s)$

Then we proceed to prove (ii). By (A.2), the term on the right hand side equals

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{\beta}^{*\top}\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*\top}\hat{\boldsymbol{\delta}}^{(t+1)} - (\frac{1}{2}\hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\Sigma}^{(t+1)}\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t+1)\top}\hat{\boldsymbol{\delta}}^{(t+1)}) \\
=& - \langle \hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \rangle - \frac{1}{2}\langle \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*, \hat{\Sigma}^{(t+1)}(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*)\rangle
\end{aligned}$$

We then have

$$\lambda_n^{(t+1)}(\|\hat{\boldsymbol{\beta}}^{(t+1)}\|_1 - \|\boldsymbol{\beta}^*\|_1) \leq -\langle \hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \rangle - \frac{1}{2}\langle \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*, \hat{\Sigma}^{(t+1)}(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*)\rangle$$

This implies

$$\begin{aligned}
&|\langle \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*, \hat{\Sigma}^{(t+1)}(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*)\rangle| \\
\leq& 2|\langle \hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^* \rangle| + 2\lambda_n^{(t+1)}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_1 \\
\leq& 2\|\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}\|_{2,s}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 + 2\sqrt{s}\lambda_n^{(t+1)}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \\
\leq& 2\|\hat{\Sigma}^{(t+1)}\boldsymbol{\beta}^* - \Sigma^*\boldsymbol{\beta}^* + \boldsymbol{\delta}^* - \hat{\boldsymbol{\delta}}^{(t+1)}\|_{2,s}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 + 2\sqrt{s}\lambda_n^{(t+1)}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \\
\leq& 4d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*)\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 + 2\sqrt{s}\lambda_n^{(t+1)}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2
\end{aligned}$$

where we apply the definition of $\|\cdot\|_{2,s}$ and Lemma 8.1 to the second inequality.

In addition, since $\|\Omega^*\|_2 \leq M$, we have $|\langle \hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*, \Sigma^*(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*)\rangle| \geq \frac{1}{M}\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2^2$, which implies that

$$\|\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\|_2 \leq 4M \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(t+1)}, \boldsymbol{\theta}^*) + 2M\sqrt{s}\lambda_n^{(t+1)}.$$

A.4 Proof of Lemma A.2

Recall that

$$\begin{aligned} B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s) = \big\{ & \boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}, \Sigma \succ 0, \\ & \omega \in (c_0, 1-c_0), (1-c_1)\Delta^2 < |\delta_1(\boldsymbol{\beta})|, |\delta_2(\boldsymbol{\beta})|, \sigma^2(\boldsymbol{\beta}) < (1+c_1)\Delta^2, \\ & \boldsymbol{\beta} - \boldsymbol{\beta}^* \in \Gamma(s), \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le C_b\Delta, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^*\|_{2,s}, \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^*\|_{2,s} \le C_b\Delta \,\big\}, \end{aligned} \tag{A.3}$$

where $\delta_1(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$, $\delta_2(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_2^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$, and $\sigma(\boldsymbol{\beta}) = \sqrt{\boldsymbol{\beta}^\top\Sigma^*\boldsymbol{\beta}}$.

Since $\omega^* \in (c_\omega, 1-c_\omega)$, when $|\omega - \omega^*| < r\Delta < |c_0 - c_\omega|$, we have $\omega \in (c_0, 1-c_0)$.

$\Delta^2 = \boldsymbol{\beta}^{*\top}\boldsymbol{\delta}^* = \boldsymbol{\beta}^{*\top}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$, then

$$\begin{aligned} & |\Delta^2/2 - \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})| = |\boldsymbol{\beta}^{*\top}(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}) - \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})| \\ \le & |\boldsymbol{\beta}^{*\top}(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2})| - \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2})| + |\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}) - \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})| \\ \le & \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| \cdot \|\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}\| + |\boldsymbol{\beta}^\top(\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} - \frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2})| \\ \le & r\Delta \cdot \frac{\sqrt{M}\Delta}{2} + r\Delta \cdot (\sqrt{M}\Delta + r\Delta) \\ = & (\frac{3r\sqrt{M}}{2} + r^2)\Delta^2. \end{aligned}$$

When $r < \frac{\sqrt{9M+16c_1}-\sqrt{9M}}{4}$, $(\frac{3r\sqrt{M}}{2} + r^2)\Delta^2 \le c_1\Delta^2$.

Moreover, when $r < \sqrt{\frac{c_1}{M}}$,

$$\|\sigma^2(\boldsymbol{\beta}) - \Delta^2\| \le \|\Sigma^*\|\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \le Mr^2\Delta^2 \le c_1\Delta^2.$$

Further, since $\boldsymbol{\beta} - \boldsymbol{\beta}^* \in \Gamma(s)$, and thus there exists some $S \subset [p]$ with $|S| = s$ such that

$$\begin{aligned} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le & \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|_1 + \|\boldsymbol{\beta}_{S^c} - \boldsymbol{\beta}_{S^c}^*\|_1 \\ \le & 3\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|_1 + \frac{3\sqrt{s}}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\ \le & 5\sqrt{s}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le 5\sqrt{s} \cdot r\Delta. \end{aligned}$$

Therefore, when $r < \frac{|c_0 - c_\omega|}{\Delta} \wedge \frac{\sqrt{9M+16c_1}-\sqrt{9M}}{4} \wedge \sqrt{\frac{c_1}{M}} \wedge \frac{C_b}{5\sqrt{s}}$, we have

$$\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s).$$

B Proof of the part (2) in Theorem 3.3

In this section we prove the lower bound for estimating $\boldsymbol{\beta}^*$. We follow the notations used in Section 8.2 in the main paper and consider the parameter space $\Theta_1 = \{\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \epsilon \boldsymbol{u} + \lambda \boldsymbol{e}_1, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \Sigma = \sigma^2 \mathbf{I}_p; \boldsymbol{u} \in \tilde{\mathcal{A}}_s\}$ defined in Sectoin 8.2, $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 2\boldsymbol{\mu}_1$.

Our strategy is to first verify the result in Lemma 3.5 and invoke Fano's lemma. Here we use the ℓ_2 loss, and define $\tilde{L}_{\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. For $\boldsymbol{u} \in \tilde{\mathcal{A}}_s = \{\boldsymbol{u} \in \{0,1\}^p :, \boldsymbol{u}^\top \boldsymbol{e}_1 = 0, \|\boldsymbol{u}\|_0 = s\}$, denote $\boldsymbol{\beta}_u = 2\epsilon \boldsymbol{u} + 2\lambda \boldsymbol{e}_1$, where ϵ and λ are chosen the same as that in Section 8.2. Thus for $\boldsymbol{u} \neq \boldsymbol{v}$,

$$\|\boldsymbol{\beta}_u - \boldsymbol{\beta}_v\|_2^2 = 4\epsilon^2 \cdot \rho_H(\boldsymbol{u}, \boldsymbol{v}) \asymp \frac{s \log p}{n}.$$

By triangle inequality, for any $\boldsymbol{u}, \boldsymbol{v} \in \tilde{\mathcal{A}}_s$,

$$\|\boldsymbol{\beta}_u - \hat{\boldsymbol{\beta}}\|_2 + \|\boldsymbol{\beta}_v - \hat{\boldsymbol{\beta}}\|_2 \geq \|\boldsymbol{\beta}_u - \boldsymbol{\beta}_v\|_2 \asymp \sqrt{\frac{s \log p}{n}}. \tag{B.1}$$

Combining (8.4), (B.1) and Fano's lemma, we obtain the desired lower bound for the estimation error in $\boldsymbol{\beta}^*$.

C Proof of Technical Lemmas

In this section we collect the detailed proofs of Lemmas 3.1, 3.2 and A.1.

C.1 Proof of Lemma 3.1

C.1.1 Goal

Let $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. We are going to show

$$\begin{aligned}
&|\omega(\boldsymbol{\theta}) - \omega^*| \leq \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*);\\
&\|\boldsymbol{\mu}_k(\boldsymbol{\theta}) - \boldsymbol{\mu}_k^*\| \leq \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \quad k = 1, 2;\\
&\|\Sigma(\boldsymbol{\theta}) - \Sigma^*\| \leq \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*).
\end{aligned}$$

C.1.2 Preliminaries

Firstly, let's show the self-consistency: $M(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$. To see this, recall that $M(\boldsymbol{\theta}) = (\omega(\boldsymbol{\theta}), \boldsymbol{\mu}_1(\boldsymbol{\theta}), \boldsymbol{\mu}_2(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ can be calculated analytically as (3.6) and (3.7). Then we have

$$\omega(\boldsymbol{\theta}^*) = \mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)] = \mathbb{E}[\mathbb{P}_{\boldsymbol{\theta}^*}(Y = 2 \mid Z)] = \mathbb{P}_{\boldsymbol{\theta}^*}(Y = 2) = \omega^*,$$

$$
\begin{aligned}
\boldsymbol{\mu}_1(\boldsymbol{\theta}^*) =& \frac{\mathbb{E}[(1-\gamma_{\boldsymbol{\theta}^*}(Z))Z]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}^*}(Z)]} = \frac{\mathbb{E}[\mathbb{P}_{\boldsymbol{\theta}^*}(Y=1 \mid Z)Z]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}^*}(Z)]} = \frac{\mathbb{E}[\mathbb{E}[I(Y=1) \mid Z]Z]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}^*}(Z)]} \\
=& \frac{\mathbb{E}[\mathbb{E}[ZI(Y=1) \mid Z]]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}^*}(Z)]} = \frac{\mathbb{E}[ZI(Y=1)]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}^*}(Z)]} = \boldsymbol{\mu}_1^*.
\end{aligned}
$$

Similarly, we obtain $\boldsymbol{\mu}_2(\boldsymbol{\theta}^*) = \boldsymbol{\mu}_2^*$ and $\Sigma(\boldsymbol{\theta}^*) = \Sigma^*$.

Therefore, our goal changes to

$$
\begin{aligned}
&|\omega(\boldsymbol{\theta}) - \omega(\boldsymbol{\theta}^*)| \le \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*); \\
&\|\boldsymbol{\mu}_k(\boldsymbol{\theta}) - \boldsymbol{\mu}_k(\boldsymbol{\theta}^*)\| \le \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \quad k=1,2; \\
&\|\Sigma(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}^*)\| \le \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*).
\end{aligned}
$$

Then we inroduce the following elementary facts on the functions $f_1(t) := \frac{1}{\{\omega e^t + (1-\omega)e^{-t}\}^2}$, $f_2(t) := \frac{t}{\{\omega e^t + (1-\omega)e^{-t}\}^2}$ and $f_3(t) := \frac{t^2-b^2}{\{\omega e^t + (1-\omega)e^{-t}\}^2}$, which will be used in our following proof:

$$
\begin{aligned}
f_1(t) &\le \frac{1}{4\omega(1-\omega)} \le \frac{1}{4\min\{\omega, 1-\omega\}}, \quad \text{for all } t \in \mathbb{R}, && \text{(C.1)} \\
\sup_{t\in[a,\infty]} f_1(t) &\le \frac{1}{\min\{\omega, 1-\omega\}^2} \exp(-2a), \quad \text{for all } a \ge 0, && \text{(C.2)} \\
|f_2(t)| &\le \frac{|t|e^{-|t|}}{\min\{\omega, 1-\omega\}^2} \le \frac{1}{4\min\{\omega^2, (1-\omega)^2\}}, \quad \text{for all } t \in \mathbb{R}, && \text{(C.3)} \\
\sup_{t\in[a,\infty]} |f_2(t)| &\le \frac{1}{\min\{\omega, 1-\omega\}^2} \exp(-3a/2), \quad \text{for all } a \ge 0, && \text{(C.4)}
\end{aligned}
$$

$$
\begin{aligned}
|f_3(t)| &\le \frac{|t^2-b^2| \cdot e^{-|t|}}{\min\{\omega, 1-\omega\}^2} \le \frac{1+b^2}{\min\{\omega^2, (1-\omega)^2\}}, \quad \text{for all } t \in \mathbb{R}, && \text{(C.5)} \\
\sup_{t\in[a,\infty]} |f_3(t)| &\le \frac{1+b^2}{\min\{\omega, 1-\omega\}^2} \exp(-a), \quad \text{for all } a \ge 0. && \text{(C.6)}
\end{aligned}
$$

C.1.3 Taylor expansion of $\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]$ and $\frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)]}$

For $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$, recall that $M(\boldsymbol{\theta}) = (\omega(\boldsymbol{\theta}), \boldsymbol{\mu}_1(\boldsymbol{\theta}), \boldsymbol{\mu}_2(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$, where

$$
\gamma_{\boldsymbol{\theta}}(Z) = \frac{\omega}{\omega + (1-\omega)\exp\{\boldsymbol{\beta}^\top(Z - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)\}},
$$

$$
\omega(\boldsymbol{\theta}) = \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)], \quad \boldsymbol{\mu}_1(\boldsymbol{\theta}) = \frac{\mathbb{E}[(1-\gamma_{\boldsymbol{\theta}}(Z))Z]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}}(Z)]}, \quad \boldsymbol{\mu}_2(\boldsymbol{\theta}) = \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}.
$$

To show Lemma 3.1, we first verify the following two inequalities:

$$|\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)] - \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]| \le \kappa_\omega(|\omega - \omega^*| \vee \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \quad \text{(C.7)}$$

$$\left\| \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)]} - \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]} \right\|_2 \le \kappa_\gamma(|\omega - \omega^*| \vee \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \quad \text{(C.8)}$$

where the constants κ_ω and κ_γ are to be determined.

It is clear that $\gamma_{\boldsymbol{\theta}}(Z)$ only depends on parameters $\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\beta}$. With a little abuse of notations, we write $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\beta})$ in the rest of the proof. Let $\Delta_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$, $\boldsymbol{\theta}_u = \boldsymbol{\theta}^* + u\Delta_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}(Z) = \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}$. Then

$$\begin{aligned}
&\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z) - \gamma_{\boldsymbol{\theta}^*}(Z)] = \mathbb{E}\Big[\int_0^1 \langle \frac{d\gamma_{\boldsymbol{\theta}}(Z)}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\theta}}\rangle\, du\Big] \\
=&\mathbb{E}\Big[\int_0^1 \langle \frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_\omega\rangle + \langle \frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\beta}}\rangle + \sum_{k=1}^2 \langle \frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\mu}_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\mu}_k}\rangle\, du\Big], \quad \text{(C.9)}
\end{aligned}$$

and

$$\begin{aligned}
g_{\boldsymbol{\theta}}(Z) - g_{\boldsymbol{\theta}^*}(Z) &= \int_0^1 (\frac{dg_{\boldsymbol{\theta}}(Z)}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u})\Delta_{\boldsymbol{\theta}}\, du \\
= \int_0^1 (\frac{\partial g_{\boldsymbol{\theta}}(Z)}{\partial\omega}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u})\Delta_\omega &+ (\frac{\partial g_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u})\Delta_{\boldsymbol{\beta}} + \sum_{k=1}^2 (\frac{\partial g_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\mu}_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u})\Delta_{\boldsymbol{\mu}_k}\, du.
\end{aligned}$$

We need further calculation to study $\partial g_{\boldsymbol{\theta}}(Z)$. By chaining rule, we have

$$\begin{aligned}
\partial g_{\boldsymbol{\theta}}(Z) =&\frac{1}{(\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)])^2}(\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)] \cdot \mathbb{E}[Z(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top] - \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]\mathbb{E}[\partial\gamma_{\boldsymbol{\theta}}(Z)]^\top) \\
=&\frac{1}{\omega(\boldsymbol{\theta})^2}(\omega(\boldsymbol{\theta})\mathbb{E}[Z(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top] - \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]\mathbb{E}[\partial\gamma_{\boldsymbol{\theta}}(Z)]^\top) \\
=&\frac{1}{\omega(\boldsymbol{\theta})^2}(\omega(\boldsymbol{\theta})\mathbb{E}[(Z - \boldsymbol{\mu}_2)(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top] - \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z - \boldsymbol{\mu}_2)]\mathbb{E}[\partial\gamma_{\boldsymbol{\theta}}(Z)]^\top), \quad \text{(C.10)}
\end{aligned}$$

where the last equality is due to the fact that

$$\omega(\boldsymbol{\theta})\mathbb{E}[\boldsymbol{\mu}_2(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top] = \omega(\boldsymbol{\theta})\boldsymbol{\mu}_2\mathbb{E}[(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top] = \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]\boldsymbol{\mu}_2\mathbb{E}[(\partial\gamma_{\boldsymbol{\theta}}(Z))^\top].$$

Therefore, $\partial g_{\boldsymbol{\theta}}(Z)$ is a function of $\partial\gamma_{\boldsymbol{\theta}}(Z)$. To show (C.7) and (C.8), we look at the partial derivatives of $\gamma_{\boldsymbol{\theta}}(Z)$ with respect to each parameter in $\boldsymbol{\theta}$. Some calculations yield that

$$
\begin{aligned}
\frac{\partial}{\partial \omega}\gamma_{\boldsymbol{\theta}}(Z) &= \frac{\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}}{\big[\omega+(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}\big]^2},\\
\frac{\partial}{\partial \boldsymbol{\beta}}\gamma_{\boldsymbol{\theta}}(Z) &= -\frac{\omega(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}}{\big[\omega+(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}\big]^2}\Big\{Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big\}\\
&= -\omega(1-\omega)\frac{\partial}{\partial \omega}\gamma_{\boldsymbol{\theta}}(Z)\Big\{Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big\}, \qquad\qquad \text{(C.11)}\\
\frac{\partial}{\partial \boldsymbol{\mu}_k}\gamma_{\boldsymbol{\theta}}(Z) &= \frac{\omega(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}}{\big[\omega+(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}\big]^2}\frac{\boldsymbol{\beta}}{2}\\
&= \omega(1-\omega)\frac{\partial}{\partial \omega}\gamma_{\boldsymbol{\theta}}(Z)\cdot\frac{\boldsymbol{\beta}}{2}, \qquad\qquad k=1,2.
\end{aligned}
$$

Due to the self-consistency shown above, $\boldsymbol{\theta}^*$ is a stationary point such that $M(\boldsymbol{\theta}^*)=\boldsymbol{\theta}^*$. This implies that the terms on the right hand side of (C.11) should be very small in expectation when $\boldsymbol{\theta}$ is close to $\boldsymbol{\theta}^*$.

It is easy to see that all three terms in (C.11) depend on the random variable $\boldsymbol{\beta}^\top\{Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2\}$. In the following, we first show that $\boldsymbol{\beta}^\top\{Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2\}$ can be equivalently written as a one-dimensional normal random variable, and use this fact to obtain probabilistic bounds of the expectations of the three terms in (C.11).

Let $\tilde{Z}=(\Omega^*)^{1/2}\{Z-(\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*)/2\}$. Then

$$
\tilde{Z} \sim (1-\omega^*)N_p\big((\Omega^*)^{1/2}\frac{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*}{2},\mathbf{I}_p\big)+\omega^* N_p\big(-(\Omega^*)^{1/2}\frac{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*}{2},\mathbf{I}_p\big) := \Psi + Z_N,
$$

where $\Psi \sim (1-\omega^*)\cdot(\Omega^*)^{1/2}(\frac{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*}{2})+\omega^*\cdot(-\Omega^*)^{1/2}(\frac{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*}{2})$, and $Z_N \sim N_p(\mathbf{0},\mathbf{I}_p)$. The random vector Z can be further rewritten in terms of $\tilde{Z}$ as $Z=(\Sigma^*)^{1/2}\tilde{Z}+(\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*)/2$. For convenience, we introduce the following notations. Let

$$
\Delta_{\boldsymbol{\beta}}=\boldsymbol{\beta}-\boldsymbol{\beta}^*, \quad \Delta_{\boldsymbol{\mu}}=(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^*-\boldsymbol{\mu}_1^*)/2,
$$

$$
\delta_{\boldsymbol{\beta}}=\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}\Psi-\boldsymbol{\beta}^\top\Delta_{\boldsymbol{\mu}}, \text{that is, } \mathbb{P}(\delta_{\boldsymbol{\beta}}=\delta_1(\boldsymbol{\beta}))=1-\omega^*=1-\mathbb{P}(\delta_{\boldsymbol{\beta}}=\delta_2(\boldsymbol{\beta})),
$$

$$
\delta_1(\boldsymbol{\beta})=\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}), \delta_2(\boldsymbol{\beta})=\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}), \sigma(\boldsymbol{\beta})=\sqrt{\boldsymbol{\beta}^\top\Sigma^*\boldsymbol{\beta}}.
$$

Then

$$
\begin{aligned}
\boldsymbol{\beta}^\top\Big(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big) &= \boldsymbol{\beta}^\top(\Sigma^*)^{1/2}\tilde{Z}-\boldsymbol{\beta}^\top\Delta_{\boldsymbol{\mu}}\\
&\overset{d}{=} \boldsymbol{\beta}^\top(\Sigma^*)^{1/2}(\Psi+Z_N)-\boldsymbol{\beta}^\top\Delta_{\boldsymbol{\mu}}\\
&= \delta_{\boldsymbol{\beta}}+\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}Z_N\\
&\overset{d}{=} \delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1},
\end{aligned}
$$

where Z_{N_1} is a standard normal random variable.

C.1.4 Contraction for the mixing proportion

Define the event

$$\mathcal{E}_1 = \{|\sigma(\boldsymbol{\beta})Z_{N_1}| < \frac{1-c_1}{2}\Delta^2\}.$$

Since on $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$, $\sigma(\boldsymbol{\beta}) < c_2\Delta$, using an elementary bound on the tails of Gaussian distribution we obtain

$$\mathbb{P}(\mathcal{E}_1^c) \le 2\exp\left\{-\frac{(1-c_1)^2\Delta^2}{8(1+c_1)}\right\}.$$

Since $\mathbb{P}(\delta_{\boldsymbol{\beta}} = \delta_1(\boldsymbol{\beta})) = 1-\omega^* = 1-\mathbb{P}(\delta_{\boldsymbol{\beta}} = \delta_2(\boldsymbol{\beta}))$, $|\delta_{\boldsymbol{\beta}}| \ge c_1\Delta^2$ on $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$. Therefore on the event $\mathcal{E}_1$, we also have $|\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1}| \ge |\delta_{\boldsymbol{\beta}}| - |\sigma(\boldsymbol{\beta})Z_{N_1}| \ge (1-c_1)\Delta^2/2$. It follows that

$$\begin{aligned}
\mathbb{E}\Big[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\Big] =& \mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega + (1-\omega)\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2}\right] \\
=& \mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega + (1-\omega)\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2} \mid \mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1) + \\
& \mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega + (1-\omega)\exp(\delta_{\boldsymbol{\beta}} + \sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2} \mid \mathcal{E}_1^c\right]\mathbb{P}(\mathcal{E}_1^c) \\
\le& \frac{1}{\min\{\omega^2, (1-\omega)^2\}}\exp\left(-\frac{1-c_1}{2}\Delta^2\right) + \frac{1}{4\min\{\omega, 1-\omega\}}\exp\left(-\frac{(1-c_1)^2\Delta^2}{8(1+c_1)}\right) \\
\le& \frac{2}{c_0^2}\exp\Big(-\big(\frac{1-c_1}{2} \wedge \frac{(1-c_1)^2}{8(1+c_1)}\big)\Delta^2\Big) := c_3\exp(-c_4\Delta^2),
\end{aligned} \tag{C.12}$$

where the first inequality comes from the facts in (C.1) and (C.2), and the constants $c_3 := \frac{2}{c_0^2}$, $c_4 = \frac{1-c_1}{2} \wedge \frac{(1-c_1)^2}{8(1+c_1)}$ also depend on c_0, c_1, c_2, defined in $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$.

To bound $\left|\langle \frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\beta}}\rangle\right|$, we notice that

$$\begin{aligned}
\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}} =& \frac{\omega(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}}{\big[\omega + (1-\omega)\exp\{\boldsymbol{\beta}^\top(Z - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\}\big]^2}\Big(Z - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big) \\
\overset{d}{=}& \omega(1-\omega)\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\Big\{(\Sigma^*)^{1/2}(\Psi + Z_N) - \Delta_{\boldsymbol{\mu}}\Big\}.
\end{aligned}$$

Then

$$\begin{aligned}\Big|\mathbb{E}\Big[\langle\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u},\Delta_{\boldsymbol{\beta}}\rangle\Big]\Big| \le & \Big|\langle\mathbb{E}\Big[\omega(1-\omega)\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\cdot(\Sigma^*)^{1/2}Z_N\Big],\Delta_{\boldsymbol{\beta}}\rangle\Big| \\ &+\Big\|\mathbb{E}\Big[\omega(1-\omega)\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\cdot(\Sigma^*)^{1/2}\Psi\Big]\Big\|_2\cdot\|\Delta_{\boldsymbol{\beta}}\|_2 \\ &+\Big\|\mathbb{E}\Big[\omega(1-\omega)\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\cdot\Delta_{\boldsymbol{\mu}}\Big]\Big\|_2\cdot\|\Delta_{\boldsymbol{\beta}}\|_2. \qquad \text{(C.13)}\end{aligned}$$

We bound the last two terms in (C.13) first. By (C.12), the second term in (C.13) can be bounded from above by

$$\begin{aligned}&\omega(1-\omega)\cdot c_3\exp(-c_4\Delta^2)\cdot\Big\|(\Sigma^*)^{1/2}\Psi\Big\|_2\cdot\|\Delta_{\boldsymbol{\beta}}\|_2\\ \le&\omega(1-\omega)\cdot c_3\exp(-c_4\Delta^2)\cdot\frac{1}{2}\sqrt{M}\Delta\cdot\|\Delta_{\boldsymbol{\beta}}\|_2\\ \le&\frac{c_3}{8}\exp(-c_4\Delta^2)\cdot\sqrt{M}\Delta\cdot\|\Delta_{\boldsymbol{\beta}}\|_2, \qquad \text{(C.14)}\end{aligned}$$

and the third term by

$$\begin{aligned}&\omega(1-\omega)\cdot c_3\exp(-c_4\Delta^2)\cdot\|\Delta_{\boldsymbol{\mu}}\|_2\cdot\|\Delta_{\boldsymbol{\beta}}\|_2\\ \le&\frac{C_bc_3}{4}\exp(-c_4\Delta^2)\cdot\Delta\cdot\|\Delta_{\boldsymbol{\beta}}\|_2. \qquad \text{(C.15)}\end{aligned}$$

To bound the first term in (C.13), let $\boldsymbol{\alpha}^\top=\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}$, and H be an orthogonal matrix whose first row is $\boldsymbol{\alpha}^T/\|\boldsymbol{\alpha}\|_2$. Then we have

$$H\boldsymbol{\alpha}=\|\boldsymbol{\alpha}\|_2\boldsymbol{e}_1=\sigma(\boldsymbol{\beta})\boldsymbol{e}_1,$$

where $\boldsymbol{e}_1$ is the basis vector in the Euclidean space whose first entry is 1 and zero elsewhere. Note that $\mathbb{E}[\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}(\Sigma^*)^{1/2}Z_N]=(\Sigma^*)^{1/2}H^\top\mathbb{E}[\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}HZ_N]$, and

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}HZ_N\right]=&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}Z_N\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}Z_N\}\big]^2}HZ_N\right]\\ =&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\alpha}^\top H^\top HZ_N\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\alpha}^\top H^\top HZ_N\}\big]^2}HZ_N\right]\\ \overset{Y=HZ_N}{=}&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1\}\big]^2}Y\right]\\ =&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}\big]^2}Z_{N_1}\boldsymbol{e}_1\right],\end{aligned}$$

where Y_1 is the first coordinate of $Y\sim N_p(\mathbf{0},\mathbf{I}_p)$, and is a standard normal random variable.

Thus $Y_1 \overset{d}{=} Z_{N_1}$, and the last equality uses the fact that Z_{N1} and Z_{Ni} are independent for any $1 < i \le p$ and $\mathbb{E}[Z_{Ni}] = 0$.

Multiplying $(\Sigma^*)^{1/2}H^\top$ and $\mathbb{E}[\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}HZ_N]$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}(\Sigma^*)^{1/2}Z_N\right] =&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}\big]^2}Z_{N_1}\right](\Sigma^*)^{1/2}H^\top\boldsymbol{e}_1\\
=&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}\big]^2}Z_{N_1}\right](\Sigma^*)^{1/2}\frac{\boldsymbol{\alpha}}{\sigma(\boldsymbol{\beta})}\\
=&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}\}\big]^2}\sigma(\boldsymbol{\beta})Z_{N_1}\right]\cdot\Sigma^*\boldsymbol{\beta}\cdot\frac{1}{\sigma^2(\boldsymbol{\beta})}.
\end{aligned}$$

Recall the event $\mathcal{E}_1 = \{|\sigma(\boldsymbol{\beta})Z_{N_1}| < c_1\Delta^2/2\}$. Using the facts in (C.1), (C.2), (C.3), (C.4) and $(1-c_1)\Delta^2 < \sigma^2(\boldsymbol{\beta}), \delta(\boldsymbol{\beta}) < (1+c_1)\Delta^2$, we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega+(1-\omega)\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2}\sigma(\boldsymbol{\beta})Z_{N_1}\right]\right|\\
=&\left|\mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega+(1-\omega)\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2}\sigma(\boldsymbol{\beta})Z_{N_1}\mid\mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\right.\\
&\left.\mathbb{E}\left[\frac{\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})}{\big\{\omega+(1-\omega)\exp(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})\big\}^2}\sigma(\boldsymbol{\beta})Z_{N_1}\mid\mathcal{E}_1^c\right]\mathbb{P}(\mathcal{E}_1^c)\right|\\
\le&\frac{1}{2\min\{\omega^2,(1-\omega)^2\}}\exp\left(-\frac{3(1-c_1)}{8}\Delta^2\right)+\frac{1}{4\min\{\omega^2,(1-\omega)^2\}}\exp\left(-\frac{(1-c_1)^2\Delta^2}{8(1+c_1)}\right)\\
\le&\frac{2}{c_0^2}\exp(-(\frac{3(1-c_1)}{8}\wedge\frac{(1-c_1)^2}{8(1+c_1)})\Delta^2)\le c_3\exp(-c_4\Delta^2).
\end{aligned}$$

Combining the pieces, the first term in (C.13) is bounded by

$$\begin{aligned}
&\left|\langle\mathbb{E}\left[\omega(1-\omega)\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\omega}\cdot(\Sigma^*)^{1/2}Z_N\right],\Delta_{\boldsymbol{\beta}}\rangle\right|\\
\le&\frac{\omega(1-\omega)}{\sigma^2(\boldsymbol{\beta})}c_3\exp(-c_4\Delta^2)\,|\langle\Sigma^*\boldsymbol{\beta},\Delta_{\boldsymbol{\beta}}\rangle\,|\\
\le&\frac{\sqrt{M}}{4\sqrt{1-c_1}\Delta}c_3\exp(-c_4\Delta^2)\cdot\|\Delta_{\boldsymbol{\beta}}\|_2,
\end{aligned}\tag{C.16}$$

where the second inequality uses the fact that $\|\Sigma^*\boldsymbol{\beta}\| \le \sqrt{M}\|\Sigma^{1/2*}\boldsymbol{\beta}\| = \sqrt{M}\sigma(\boldsymbol{\beta})$ and $\sigma(\boldsymbol{\beta}) > \sqrt{1-c_1}\Delta$.

Therefore by (C.14), (C.15) and (C.16), we obtain

$$\left|\mathbb{E}\Big[\langle\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u},\Delta_{\boldsymbol{\beta}}\rangle\Big]\right|\le c_{\boldsymbol{\beta}1}\|\Delta_{\boldsymbol{\beta}}\|_2\tag{C.17}$$

where $c_{\boldsymbol{\beta}1} = (\frac{\sqrt{M}+C_b}{4}\Delta + \frac{\sqrt{M}}{4\sqrt{1-c_1}\Delta}) \cdot c_3 \exp(-c_4\Delta^2)$.

Finally,

$$\begin{aligned}
|\langle \frac{\partial}{\partial \boldsymbol{\mu}_1}\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)], \Delta_{\boldsymbol{\mu}_1}\rangle| = &\left|\mathbb{E}\left[\frac{\omega(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)\}}{\left[\omega+(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)\}\right]^2}\right]\langle\frac{\boldsymbol{\beta}}{2}, \Delta_{\boldsymbol{\mu}_1}\rangle\right| \\
\le&\omega(1-\omega)\cdot c_3 \cdot \exp(-c_4\Delta^2)\cdot(|\langle\frac{\boldsymbol{\beta}-\boldsymbol{\beta}^*}{2}, \Delta_{\boldsymbol{\mu}_1}\rangle| + |\langle\frac{\boldsymbol{\beta}^*}{2}, \Delta_{\boldsymbol{\mu}_1}\rangle|) \\
\le&\omega(1-\omega)\cdot c_3\cdot\exp(-c_4\Delta^2)\cdot\frac{C_b+\sqrt{M}}{2}\Delta\|\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_1\|_{2,s} \\
\le&\frac{C_b+\sqrt{M}}{8}\Delta\cdot c_3\exp(-c_4\Delta^2)\|\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_1\|_{2,s},
\end{aligned}$$

By symmetry, this is also true for $\boldsymbol{\mu}_2$, and therefore we have

$$\left|\mathbb{E}\Big[\langle\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\mu}_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\mu}_k}\rangle\Big]\right| \le \frac{C_b+\sqrt{M}}{8}\Delta\cdot c_3\exp(-c_4\Delta^2)\|\boldsymbol{\mu}_k^*-\boldsymbol{\mu}_k\|_{2,s}, \quad k=1,2. \tag{C.18}$$

In summary, plugging (C.12), (C.17), (C.18) into (C.9), and letting $\kappa_\omega = c_3\exp(-c_4\Delta^2)\cdot[(\frac{\sqrt{M}+C_b}{2}\Delta + \frac{\sqrt{M}}{4\sqrt{1-c_1}\Delta}) + 1]$, we have

$$|\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)] - \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]| \le \kappa_\omega(|\omega-\omega^*| \vee \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{2,s}).$$

which concludes the first inequality (C.7).

C.1.5 Contraction on the mean

Then let us proceed to show the inequality (C.8). First recall that $Z = (\Sigma^*)^{1/2}Z_N + \{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi\}$, where $Z_N \sim N_p(\mathbf{0}, \mathbf{I}_p)$ and $\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi \sim (1-\omega)^*\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*$. Then by (C.10), (C.12) and (C.16), to show the contraction $\|\frac{\partial g_{\boldsymbol{\theta}}(Z)}{\partial\omega}\Delta\omega\|_2 \le c_{\omega 2}|\Delta_\omega|$, we only need to bound

$$\begin{aligned}
&2\left\|\frac{\partial}{\partial\omega}\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)]\right\|_2 \\
=&2\left\|\mathbb{E}\Big[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\Big]\right\|_2 + 2\left\|\mathbb{E}\Big[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\Big]\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi - \boldsymbol{\mu}_2\Big\}\right\|_2 \\
\le&2\left\|\mathbb{E}\Big[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\Big]\right\|_2 + 2c_3\exp(-c_4\Delta^2)(\sqrt{M}+C_b)\Delta \\
\le&2\frac{\sqrt{M}}{\sqrt{1-c_1}\Delta}c_3\exp(-c_4\Delta^2) + 2c_3\exp(-c_4\Delta^2)(\sqrt{M}+C_b)\Delta \\
\le&2(\frac{\sqrt{M}}{\sqrt{1-c_1}\Delta}c_3 + c_3(\sqrt{M}+C_b)\Delta)\exp(-c_4\Delta^2) := c_6\exp(-c_4\Delta^2).
\end{aligned} \tag{C.19}$$

Next we are going to show $\|\mathbb{E}[\langle \frac{\partial \gamma_{\boldsymbol{\theta}}(Z)}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u}, \Delta_{\boldsymbol{\beta}}\rangle (Z-\boldsymbol{\mu}_2)]\|_2 \le c_{\boldsymbol{\beta}2}\|\Delta_{\boldsymbol{\beta}}\|_2$ for some $c_{\boldsymbol{\beta}2} > 0$. By (C.10), it suffices to bound $\|\mathbb{E}[\frac{\partial}{\partial \boldsymbol{\beta}}\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)]\|_2$.

Since $Z = (\Sigma^*)^{1/2}\tilde{Z} + (\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*)/2$ with $\tilde{Z} = \Psi + Z_N$, we have $Z = (\Sigma^*)^{1/2}(\Psi + Z_N) + (\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*)/2 = (\Sigma^*)^{1/2}Z_N + (\Sigma^*)^{1/2}\Psi + (\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*)/2$, and it follows

$$
\begin{aligned}
(Z-\boldsymbol{\mu}_2)(Z - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})^\top =& (\Sigma^*)^{1/2} Z_N {Z_N}^\top (\Sigma^*)^{1/2} + \Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi - \boldsymbol{\mu}_2\Big\}{Z_N}^\top(\Sigma^*)^{1/2} \\
&+ (\Sigma^*)^{1/2}Z_N\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big\}^\top \\
&+ \Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi - \boldsymbol{\mu}_2\Big\}\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\Big\}^\top.
\end{aligned}
$$

By definition,

$$
\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} + (\Sigma^*)^{1/2}\Psi \sim (1-\omega^*)\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*.
$$

Therefore by (C.11)

$$
\begin{aligned}
&\frac{1}{\omega(1-\omega)}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)] = \mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)Z^\top\right] - \mathbb{E}[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)]\cdot(\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})^\top \\
=&\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N{Z_N}^\top(\Sigma^*)^{1/2}\right] + \mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} + (\Sigma^*)^{1/2}\Psi\Big\}^\top\right] \\
&+\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} - \boldsymbol{\mu}_2 + (\Sigma^*)^{1/2}\Psi\Big\}{Z_N}^\top(\Sigma^*)^{1/2}\right] \\
&+\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\right]\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} - \boldsymbol{\mu}_2 + (\Sigma^*)^{1/2}\Psi\Big\}\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} + (\Sigma^*)^{1/2}\Psi\Big\}^\top.
\end{aligned}
\tag{C.20}
$$

The second to forth terms are easier to bound. For the second term in (C.20), its $\|\cdot\|_2$ norm can be bounded from above by

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} + (\Sigma^*)^{1/2}\Psi\Big\}^\top\right]\right\|_2 \\
\le&\left\|\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\right]\right\|_2 \cdot \left\|\Big\{\Delta_{\boldsymbol{\mu}} + (\Sigma^*)^{1/2}\Psi\Big\}^\top\right\|_2 \\
\le& c_6\exp(-c_4\Delta^2)\cdot(\sqrt{M} + C_b)\Delta,
\end{aligned}
$$

where the last inequality is due to (C.19). The $\|\cdot\|_2$ norm of the third term in (C.20) can be similarly bounded by

$$\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}-\boldsymbol{\mu}_2+(\Sigma^*)^{1/2}\Psi\Big\}Z_N{}^\top(\Sigma^*)^{1/2}\right]\right\|_2\\
\leq&\left\|\mathbb{E}\left[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)(\Sigma^*)^{1/2}Z_N\right]\right\|_2\cdot\left\|\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}-\boldsymbol{\mu}_2+(\Sigma^*)^{1/2}\Psi\Big\}^\top\right\|_2\\
\leq& c_6\exp(-c_4\Delta^2)\cdot(\sqrt{M}+C_b)\Delta,
\end{aligned}$$

Further, the forth term can be bounded by

$$\begin{aligned}
&\left\|\mathbb{E}\Big[\frac{\partial}{\partial\omega}\gamma_{\boldsymbol{\theta}}(Z)\Big]\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}-\boldsymbol{\mu}_2+(\Sigma^*)^{1/2}\Psi\Big\}\Big\{\frac{\boldsymbol{\mu}_1^*+\boldsymbol{\mu}_2^*}{2}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}+(\Sigma^*)^{1/2}\Psi\Big\}^\top\right\|_2\\
\leq& 2c_3\exp(-c_4\Delta^2)\cdot(M+C_b^2)\Delta^2
\end{aligned}$$

Lastly we bound the first term in (C.20). Recall that $\boldsymbol{\alpha}^\top=\boldsymbol{\beta}^\top(\Sigma^*)^{1/2}$, H is an orthogonal matrix whose first row is $\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|_2$, and $H\boldsymbol{\alpha}=\|\boldsymbol{\alpha}\|_2\boldsymbol{e}_1$. Then

$$\begin{aligned}
&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\alpha}^\top H^\top HZ_N\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\boldsymbol{\alpha}^\top H^\top HZ_N\}\big]^2}HZ_NZ_N{}^\top H^\top\right]\\
\overset{Y=HZ_N}{=}&\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1\}}{\big[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1\}\big]^2}YY^\top\right]\\
=&\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1}\big\}^2}(Y_1^2-1)\right]\boldsymbol{e}_1\boldsymbol{e}_1{}^\top+\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\|\boldsymbol{\alpha}\|_2Y_1}\big\}^2}\right]\mathbf{I}_p\\
\overset{d}{=}&\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\big\}^2}(Z_{N_1}^2-1)\right]\boldsymbol{e}_1\boldsymbol{e}_1{}^\top+\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\big\}^2}\right]\mathbf{I}_p\\
:=&v_1\boldsymbol{e}_1\boldsymbol{e}_1{}^\top+v_2\mathbf{I}_p,
\end{aligned}$$

where we have defined

$$v_1=\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\big\}^2}(Z_{N_1}^2-1)\right],\quad v_2=\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\big\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\big\}^2}\right].$$

Recall the event $\mathcal{E}_1=\{|\sigma(\boldsymbol{\beta})Z_{N_1}|<\frac{1-c_1}{2}\Delta^2\}$. Using the facts in (C.1), (C.2), (C.5) and

(C.6), we obtain

$$
\begin{aligned}
v_1 =& \left|\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\left\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\right\}^2}(Z_{N_1}^2-1)\right]\right| \\
=& \left|\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\left\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\right\}^2}(Z_{N_1}^2-1)\mid \mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\right. \\
& \left.\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\left\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\right\}^2}(Z_{N_1}^2-1)\mid \mathcal{E}_1^c\right]\mathbb{P}(\mathcal{E}_1^c)\right| \\
=& \left|\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\left\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\right\}^2}(\frac{(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})^2-2\delta_{\boldsymbol{\beta}}(\sigma(\boldsymbol{\beta})Z_{N_1}+\delta_{\boldsymbol{\beta}})-2\delta_{\boldsymbol{\beta}}^2-\sigma^2(\boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})})\mid \mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\right. \\
& \left.\mathbb{E}\left[\frac{e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}}{\left\{\omega+(1-\omega)e^{\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1}}\right\}^2}(\frac{(\delta_{\boldsymbol{\beta}}+\sigma(\boldsymbol{\beta})Z_{N_1})^2-2\delta_{\boldsymbol{\beta}}(\sigma(\boldsymbol{\beta})Z_{N_1}+\delta_{\boldsymbol{\beta}})-2\delta_{\boldsymbol{\beta}}^2-\sigma^2(\boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})})\mid \mathcal{E}_1^c\right]\mathbb{P}(\mathcal{E}_1^c)\right| \\
\leq& \frac{1+2(1+c_1)\Delta^2+2(1+c_1)^2\Delta^4}{(1-c_1)\Delta^2}\cdot\frac{2}{\min\{\omega^2,(1-\omega)^2\}}\exp\Big(-\frac{c_1}{4}\Delta^2\Big) \\
& +\frac{1+2(1+c_1)\Delta^2+2(1+c_1)^2\Delta^4}{(1-c_1)\Delta^2\min\{\omega^2,(1-\omega)^2\}}\exp\left(-\frac{(1-c_1)^2\Delta^2}{8(1+c_1)}\right) \\
\leq& \frac{1+2(1+c_1)\Delta^2+2(1+c_1)^2\Delta^4}{(1-c_1)\Delta^2}\cdot\frac{2}{c_0^2}\exp\left\{-\Big(\frac{c_1}{4}\wedge\frac{(1-c_1)^2}{8(1+c_1)}\Big)\Delta^2\right\}:=c_7\exp(-c_4\Delta^2).
\end{aligned}
$$

Additionally, v_2 can be similarly bounded by

$$
v_2\leq c_3\exp(-c_4\Delta^2).
$$

By multiplying $\Sigma^{*1/2}$ on both left and right hands, we then get

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\frac{\exp\{\delta_{\boldsymbol{\beta}}+(\boldsymbol{\beta}^*+\Delta_{\boldsymbol{\beta}})^\top(\Sigma^*)^{1/2}Z_N\}}{\left[\omega+(1-\omega)\exp\{\delta_{\boldsymbol{\beta}}+(\boldsymbol{\beta}^*+\Delta_{\boldsymbol{\beta}})^\top(\Sigma^*)^{1/2}Z_N\}\right]^2}(\Sigma^*)^{1/2}Z_N{Z_N}^\top(\Sigma^*)^{1/2}\right]\right\|_2 \\
=&\|(\Sigma^*)^{1/2}H^\top(v_1\boldsymbol{e}_1{\boldsymbol{e}_1}^\top+v_2\mathbf{I}_p)H(\Sigma^*)^{1/2}\|_2 \\
=&\|\frac{v_1}{\|\boldsymbol{\alpha}\|_2^2}(\Sigma^*)^{1/2}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top(\Sigma^*)^{1/2}+v_2\Sigma^*\|_2 \\
\leq&\frac{v_1}{\|\boldsymbol{\alpha}\|_2^2}M\Delta^2+v_2M \\
\leq&\frac{v_1}{\sigma(\boldsymbol{\beta})^2}M\Delta^2+v_2M \\
\leq&\frac{4Mc_7}{(1-c_1)}\exp(-c_4\Delta^2)+c_3M\exp(-c_4\Delta^2).
\end{aligned}
$$

Combining the pieces yields

$$
\|\mathbb{E}[\langle\frac{\partial\gamma_{\boldsymbol{\theta}}(Z)}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_u},\Delta_{\boldsymbol{\beta}}\rangle Z]\|_2\leq c_{\boldsymbol{\beta}2}\|\Delta_{\boldsymbol{\beta}}\|_{2,s},
$$

where $c_{\boldsymbol{\beta}2} = \exp(-c_4\Delta^2)\left[\left(M + (M + C_b^2)\Delta^2\right)c_3 + (1 + 2(\sqrt{M} + C_b)\Delta)c_6 + \frac{4M}{1-c_1}c_7\right]$.

Finally, by (C.19)

$$\begin{aligned}\left\|\frac{\partial}{\partial\boldsymbol{\mu}_k}\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)]\Delta_{\boldsymbol{\mu}_k}\right\|_2 =& \left\|\mathbb{E}\left[\frac{\omega(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)\}}{\left[\omega+(1-\omega)\exp\{\boldsymbol{\beta}^\top(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)\}\right]^2}(Z-\boldsymbol{\mu}_2)\right]\right\|_2 \frac{\langle\boldsymbol{\beta},\Delta_{\boldsymbol{\mu}_k}\rangle}{2}\\ \le&\omega(1-\omega)\left\|\frac{\partial}{\partial\omega}\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2)]\right\|_2\frac{\|\boldsymbol{\beta}^*\|_2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2}{2}\|\Delta_{\boldsymbol{\mu}_k}\|_{2,s}\\ \le&2(C_b+\sqrt{M})\Delta c_6\exp(-c_4\Delta^2)\|\Delta_{\boldsymbol{\mu}_k}\|_{2,s},\end{aligned}$$

where the first inequality follows the fact that $\boldsymbol{\beta}^*, \boldsymbol{\beta}-\boldsymbol{\beta}^* \in \Gamma(s)$.

Let $\kappa_{\boldsymbol{\mu}} = c_3\exp(-c_4\Delta^2) + \exp(-c_4\Delta^2)\left[\left(M + (M + C_b^2)\Delta^2\right)c_3 + (1 + 2(\sqrt{M} + C_b)\Delta)c_6 + \frac{4M}{1-c_1}c_7\right] + 2(C_b+\sqrt{M})\Delta c_6\exp(-c_4\Delta^2)$. Then (C.8) holds:

$$\|\frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}^*}(Z)]} - \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}\|_2 \le \kappa_{\boldsymbol{\mu}}(|\omega-\omega^*| \vee \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2).$$

Similarly, we also have

$$\|\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1{}^*\|_2 \le \kappa_{\boldsymbol{\mu}}\cdot(|\omega-\omega^*| \vee \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2),$$

which implies

$$\|\boldsymbol{\mu}_k(\boldsymbol{\theta}) - \boldsymbol{\mu}_k^*\|_2 \le \kappa_{\boldsymbol{\mu}}(|\omega-\omega^*| \vee \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^*\|_{2,s} \vee \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2), \quad k=1,2. \tag{C.21}$$

C.1.6 Contraction on the covariance matrix

Let $\boldsymbol{\mu}^* = (1-\omega^*)\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*$, we have

$$\begin{aligned}\Sigma(\boldsymbol{\theta}) - \Sigma^* =& [1-\omega(\boldsymbol{\theta})]\boldsymbol{\mu}_1(\boldsymbol{\theta})[\boldsymbol{\mu}_1(\boldsymbol{\theta})]^\top - [1-\omega^*]\boldsymbol{\mu}_1^*(\boldsymbol{\mu}_1^*)^\top + \omega(\boldsymbol{\theta})\boldsymbol{\mu}_2(\boldsymbol{\theta})[\boldsymbol{\mu}_2(\boldsymbol{\theta})]^\top - \omega^*\boldsymbol{\mu}_2^*(\boldsymbol{\mu}_2^*)^\top\\ =& [1-\omega(\boldsymbol{\theta})](\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)[\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*]^\top - [1-\omega^*](\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)^\top\\ &+ \omega(\boldsymbol{\theta})(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)[\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*]^\top - \omega^*(\boldsymbol{\mu}_2-\boldsymbol{\mu}^*)^*(\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*)^\top,\end{aligned}$$

where the second equality uses the fact $(1-\omega(\boldsymbol{\theta}))\boldsymbol{\mu}_1(\boldsymbol{\theta}) + \omega(\boldsymbol{\theta})\boldsymbol{\mu}_2(\boldsymbol{\theta}) = \mathbb{E}[(1-\gamma_\theta(Z))Z] + \mathbb{E}[\gamma_\theta(Z)Z] = \mathbb{E}[Z] = (1-\omega^*)\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*$.

By triangle inequality,

$$\begin{aligned}\|\Sigma(\boldsymbol{\theta}) - \Sigma^*\|_2 \le& \underbrace{\|[1-\omega(\boldsymbol{\theta})](\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top - [1-\omega^*](\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)^\top\|_2}_{(i)}\\ &+ \underbrace{\|\omega(\boldsymbol{\theta})(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)[\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*]^\top - \omega^*(\boldsymbol{\mu}_2-\boldsymbol{\mu}^*)^*(\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*)^\top\|_2}_{(ii)}.\end{aligned}$$

The first term (i) can be bounded from above by

$$\begin{aligned}
(i) \leq & |\omega^* - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
& + |1 - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)^\top - (\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
\leq & |\omega^* - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
& + |1 - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)^\top - (\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
& + |1 - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top - (\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
\leq & |\omega^* - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
& + |1 - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1^*)^\top\|_2 \\
& + |1 - \omega(\boldsymbol{\theta})| \cdot \|(\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1^*)(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*)^\top\|_2 \\
\leq & |\omega^* - \omega(\boldsymbol{\theta})| \cdot \Delta^2 + \|\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1^*\|_2 \cdot \|\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\|_2 \\
\leq & |\omega^* - \omega(\boldsymbol{\theta})| \cdot \Delta^2 + \|\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1^*\|_2 \cdot (\|\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}_1^*\|_2 + \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*\|_2) \\
\leq & \kappa_\omega \Delta^2 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \kappa_{\boldsymbol{\mu}} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)(\kappa_{\boldsymbol{\mu}} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \sqrt{M}\Delta) \\
\leq & \kappa_\omega \Delta^2 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \kappa_{\boldsymbol{\mu}} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)(C_b\Delta + \sqrt{M}\Delta) \\
= & (\kappa_\omega \Delta^2 + \kappa_{\boldsymbol{\mu}} \cdot (C_b\Delta + \sqrt{M}\Delta)) d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*).
\end{aligned}$$

Term (ii) can be bounded similarly. It follows immediately that

$$\begin{aligned}
\|(\Sigma(\boldsymbol{\theta}) - \Sigma^*)\boldsymbol{\beta}^*\|_2 \leq & \|\Sigma(\boldsymbol{\theta}) - \Sigma^*\|_2 \cdot \sqrt{M}\Delta & \text{(C.22)} \\
\leq & \kappa_\Sigma(|\omega - \omega^*| \vee \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^*\|_2 \vee \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^*\|_{2,s} \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{2,s}), & \text{(C.23)}
\end{aligned}$$

with $\kappa_\Sigma := \kappa_\omega \sqrt{M}\Delta^3 + \kappa_{\boldsymbol{\mu}} \cdot (\sqrt{M}C_b\Delta^2 + M\Delta^3)$.

Recall that $\kappa_\omega = c_3 \exp(-c_4\Delta^2) \cdot [(\frac{\sqrt{M}+C_b}{2}\Delta + \frac{\sqrt{M}}{4\sqrt{1-c_1\Delta}}) + 1]$,

$$\begin{aligned}
\kappa_{\boldsymbol{\mu}} &= \exp(-c_4\Delta^2)\left[c_3 + \left(M + (M + C_b^2)\Delta^2\right)c_3 + (1 + 2(\sqrt{M} + C_b)\Delta)c_6 + \frac{4M}{1-c_1}c_7 + 2(C_b + \sqrt{M})\Delta c_6\right] \\
&= \exp(-c_4\Delta^2)\left[(c_3 + Mc_3 + \frac{4M}{1-c_1}c_7 + c_6) + 4(\sqrt{M} + C_b)c_6\Delta + (M + C_b^2)c_3\Delta^2\right].
\end{aligned}$$

Assuming $\Delta > M^{-1/6}$, we then have

$$\begin{aligned}
\kappa_0 :=& \kappa_\omega \vee \kappa_{\boldsymbol{\mu}} \vee \kappa_\Sigma \leq \kappa_\omega \sqrt{M}\Delta^3 + \kappa_{\boldsymbol{\mu}} \cdot (\sqrt{M}C_b\Delta^2 + M\Delta^3 + 1) \\
\leq & \exp(-c_4\Delta^2) \cdot \Big\{ [(\frac{M + \sqrt{M}C_b}{2} c_3 \Delta^4 + \frac{Mc_3}{4\sqrt{1-c_1}}\Delta^2) + \sqrt{M}\Delta^3] \\
& + \left[(c_3 + Mc_3 + \frac{4M}{1-c_1}c_7 + c_6) + 4(\sqrt{M} + C_b)c_6\Delta + (M + C_b^2)c_3\Delta^2 \right] \cdot (\sqrt{M}C_b\Delta^2 + M\Delta^3 + 1) \Big\} \\
= & \exp(-c_4\Delta^2) \cdot \mathrm{poly}(\Delta; c_0, c_1, M, C_b) \\
:= & f_\kappa(\Delta; c_0, c_1, M, C_b),
\end{aligned}$$

where as denoted before, $c_3 = \frac{2}{c_0^2}$, $c_4 = \frac{1-c_1}{2} \wedge \frac{(1-c_1)^2}{8(1+c_1)}$, $c_6 = 2(\frac{\sqrt{M}}{\sqrt{1-c_1}\Delta}c_3 + c_3(\sqrt{M} + C_b)\Delta)$, $c_7 = \frac{1+2(1+c_1)\Delta^2+2(1+c_1)^2\Delta^4}{(1-c_1)\Delta^2} \cdot \frac{2}{c_0^2}$.

Let $C(c_0, c_1, M, C_b)$ denotes the number that satisfy

$$f_\kappa(\Delta; c_0, c_1, M, C_b) < \frac{1}{2 \vee (16M)} \tag{C.24}$$

when $\Delta > C(c_0, c_1, M, C_b)$.

Therefore, when $\Delta > M^{-1/6} \vee C(c_0, c_1, M, C_b)$, there exists $\kappa_0 \in \frac{2}{12MC_1+3}$, such that

$$d_{2,s}(M(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

This concludes the proof.

C.2 Proof of Lemma 3.2

We divide the proof into several parts, where each part shows derivation of the concentration inequalities for the estimate $\hat{\omega}$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ and $\hat{\Sigma}$ separately. Key components in showing our concentration results are the following lemma on a variant of Talagrand comparison inequality and another lemma on the covering number of $\Gamma(s)$.

C.2.1 A variant of Talagrand comparison inequality

First, recall a function $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz with constant L if and only if $|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, where $\mathcal{X}$ is some Polish space.

Lemma C.1. *Let $\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(n)}$ be n independent realizations of the random variable $\boldsymbol{z} \in \mathcal{X}$ and $\mathcal{F}$ be a function class defined on $\mathcal{X}$. Suppose $\epsilon_1, ..., \epsilon_n$ are i.i.d. Rademacher random variables. Consider the Lipschitz functions $\psi_i(\cdot)$ $(i = 1, ..., n)$ with Lipschitz constant L that satisfy $\psi(0) = 0$. Then for any increasing convex function $\phi(\cdot)$ and a fixed $g \in \mathcal{F}$, we* *have*

$$\mathbb{E}\left[\phi\left(\Big|\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot\psi_i[f(\boldsymbol{z}^{(i)})]\cdot g(\boldsymbol{z}^{(i)})\Big|\right)\right]\le\mathbb{E}\left[\phi\left(2\Big|L\cdot\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot f(\boldsymbol{z}^{(i)})\cdot g(\boldsymbol{z}^{(i)})\Big|\right)\right]$$

Proof. First we fix sample $S=(\boldsymbol{z}^{(1)},...,\boldsymbol{z}^{(n)})$, and write

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\big\{\phi[|\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot\psi_i[f(\boldsymbol{z}^{(i)})]\cdot g(\boldsymbol{z}^{(i)})|]\big\}=&\mathbb{E}_{\epsilon_1,...,\epsilon_{n-1}}\mathbb{E}_{\epsilon_n}\big\{\phi[|\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\epsilon_i\cdot\psi_i[f(\boldsymbol{z}^{(i)})]\cdot g(\boldsymbol{z}^{(i)})|]\big\}\\=&\mathbb{E}_{\epsilon_1,...,\epsilon_{n-1}}\mathbb{E}_{\epsilon_n}\big\{\phi[|\sup_{f\in\mathcal{F}}A_{n-1}(f)+\epsilon_n\cdot\psi_n[f(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})|]\big\}\\=&\frac{1}{2}\mathbb{E}_{\epsilon_1,...,\epsilon_{n-1}}\mathbb{E}_{\epsilon_n}\big\{\phi[|\sup_{f\in\mathcal{F}}A_{n-1}(f)+\psi_n[f(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})|]\big\}\\&+\frac{1}{2}\mathbb{E}_{\epsilon_1,...,\epsilon_{n-1}}\mathbb{E}_{\epsilon_n}\big\{\phi[|\sup_{f\in\mathcal{F}}A_{n-1}(f)-\psi_n[f(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})|]\big\},\end{aligned}$$

where $A_{n-1}(f)=\sum_{i=1}^{n-1}\epsilon_i\cdot\psi_i[f(\boldsymbol{z}^{(i)})]\cdot g(\boldsymbol{z}^{(i)})$.

Suppose the suprema on the right hand side of the equation above are achieved at (f_1,f_2). Then we claim

$$\begin{aligned}&\mathbb{E}_{\epsilon_n}\big\{\phi[\sup_{f\in\mathcal{F}}A_{n-1}(f)+\epsilon_n\cdot\psi_n[f(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})]\big\}\\=&\frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_1)+\psi_n[f_1(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})]\big\}+\frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_2)-\psi_n[f_2(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)})]\big\}\\\le&\frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_1)+sLf_1(\boldsymbol{z}^{(n)})\cdot g(\boldsymbol{z}^{(n)})]\big\}+\frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_2)-sLf_2(\boldsymbol{z}^{(n)})\cdot g(\boldsymbol{z}^{(n)})]\big\},\end{aligned}\tag{C.25}$$

where $s=\operatorname{sgn}(\{f_1(\boldsymbol{z}^{(n)})-f_2(\boldsymbol{z}^{(n)})\}g(\boldsymbol{z}^{(n)}))$.

In the following we proceed to prove (C.25). Without loss of generality, assume $g(\boldsymbol{z}^{(n)})\ge 0$. We consider the following two cases: 1). $f_1(\boldsymbol{z}^{(n)})f_2(\boldsymbol{z}^{(n)})\ge 0$, and 2). $f_1(\boldsymbol{z}^{(n)})f_2(\boldsymbol{z}^{(n)})<0$.

In the first case $f_1(\boldsymbol{z}^{(n)})f_2(\boldsymbol{z}^{(n)})\ge 0$, we assume, without loss of generality, $f_1(\boldsymbol{z}^{(n)})\ge f_2(\boldsymbol{z}^{(n)})\ge 0$. (The proof for the cases $f_2(\boldsymbol{z}^{(n)})\ge f_1(\boldsymbol{z}^{(n)})\ge 0$ and $f_1(\boldsymbol{z}^{(n)}),f_2(\boldsymbol{z}^{(n)})\le 0$ are similar.)

By the Lipschitz property, we have

$$L\{f_1(\boldsymbol{z}^{(n)})-f_2(\boldsymbol{z}^{(n)})\}\cdot g(\boldsymbol{z}^{(n)})\ge\psi_n[f_1(\boldsymbol{z}^{(n)})]-\psi_n[f_2(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)}).$$

Let

$$a=A_{n-1}(f_1)+\psi_n[f_1(\boldsymbol{z}^{(n)})]\cdot g(\boldsymbol{z}^{(n)}),\quad b=A_{n-1}(f_1)+sLf_1(\boldsymbol{z}^{(n)})\cdot g(\boldsymbol{z}^{(n)}),$$

and

$$c = A_{n-1}(f_2) - \psi_n[f_2(\boldsymbol{z}^{(n)})] \cdot g(\boldsymbol{z}^{(n)}), \quad d = A_{n-1}(f_2) - sLf_2(\boldsymbol{z}^{(n)}) \cdot g(\boldsymbol{z}^{(n)}).$$

Then the Lipschitz property implies that $b - a \geq c - d$.

In addition, by the increasing property of ϕ, and the facts that f_1 is the maximizer and $\psi_n(0) = 0$,

$$\begin{aligned} a = A_{n-1}(f_1) + \psi_n[f_1(\boldsymbol{z}^{(n)})] \cdot g(\boldsymbol{z}^{(n)}) &\geq A_{n-1}(f_2) + \psi_n[f_2(\boldsymbol{z}^{(n)})] \cdot g(\boldsymbol{z}^{(n)}) \\ \geq A_{n-1}(f_2) - sLf_2(\boldsymbol{z}^{(n)}) \cdot g(\boldsymbol{z}^{(n)}) = d. \end{aligned}$$

Therefore by the convexity of ϕ, the increment $\phi(b) - \phi(a)$ is lager than $\phi(c) - \phi(d)$, and we have the inequality (C.25).

For the second case where $f_1(\boldsymbol{z}^{(n)})f_2(\boldsymbol{z}^{(n)}) < 0$, we assume, without loss of generality, $f_1(\boldsymbol{z}^{(n)}) > 0 > f_2(\boldsymbol{z}^{(n)})$ and $s = 1$. The the monotonicity of ϕ directly yields

$$\phi(a) + \phi(c) \leq \phi(b) + \phi(d),$$

which is equivalent to (C.25).

Further, note ϵ_i's are Rademacher random variables,

$$\begin{aligned} &\frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_1) + sLf_1(\boldsymbol{z}^{(n)}) \cdot g(\boldsymbol{z}^{(n)})]\big\} + \frac{1}{2}\mathbb{E}_{\epsilon_n}\big\{\phi[A_{n-1}(f_2) - sLf_2(\boldsymbol{z}^{(n)}) \cdot g(\boldsymbol{z}^{(n)})]\big\} \\ \leq &\mathbb{E}_{\epsilon_n}\big\{\sup_{f \in \mathcal{F}} \phi[A_{n-1}(f) + L\epsilon_n[f(\boldsymbol{z}^{(n)})] \cdot g(\boldsymbol{z}^{(n)})]\big\}. \end{aligned}$$

Following the same procedure, we can replace all $\psi_i[f(\boldsymbol{z}^{(i)})]$ by $Lf(\boldsymbol{z}^{(i)})$. This completes the proof. □

C.2.2 Covering number of $\Gamma(s)$

Recall that $\Gamma(s) = \{\boldsymbol{u} \in \mathbb{R}^p : 2\|\boldsymbol{u}_{S^c}\|_1 \leq 4\|\boldsymbol{u}_S\|_1 + 3\sqrt{s}\|\boldsymbol{u}\|_2$, for some $S \subset [p]$ with $|S| = s\}$. Let $\{\boldsymbol{u}^1, ..., \boldsymbol{u}^{M_{net}}\}$ denote a 1/2-net of the $\Gamma(s) \cap \mathbb{S}^{p-1}$, that is, for any $\boldsymbol{v} \in \Gamma(s) \cap \mathbb{S}^{p-1}$, there is some index $j \in [M_{net}]$ such that $\|\boldsymbol{v} - \boldsymbol{u}^j\|_2 \leq 1/2$. In this section we provide the following lemmas to bound M_{net}, the packing number of $\Gamma(s)$, which will be used in the later proof.

Lemma C.2. *There exists some constant $C_{net} > 0$, such that*

$$\log M_{net} \leq d \log(\frac{5ep}{d}) \leq C_{net} \cdot s \log(\frac{p}{s}).$$

Proof. We need two additional lemmas, and Lemma C.4 is proved in Section C.2.6.

Lemma C.3 (Milman and Schechtman (1986))**.** *Given $m \geq 1$ and $\epsilon > 0$. There exists an ϵ-net $\Pi \subset B_2^m$ of B_2^m with respect to the Euclidean metric such that $B_2^m \subset (1-\epsilon)^{-1}conv(\Pi)$ and $|\Pi| \leq (1+2/\epsilon)^m$.*

Lemma C.4. *Let $1 > \delta > 0$. Let $0 < s < p$ and $k_0 > 0$. Define $d = 5185s$, then*

$$\Gamma(s) \cap \mathbb{S}^{p-1} \subset 2 \cdot \mathrm{conv}(\cup_{\mathrm{|J| \leq d}} \mathrm{E_J} \cap \mathbb{S}^{\mathrm{p-1}}), \tag{C.26}$$

where $\mathrm{conv}(\cdot)$ *denotes the convex hull and $E_J = span(\boldsymbol{e}_j : j \in J)$.*

These two lemmas imply that

$$M_{net} \leq (1 + 2/(\frac{1}{2}))^d \cdot \binom{p}{d} \leq (\frac{5ep}{d})^d = \exp(d \log(\frac{5ep}{d})),$$

which implies that

$$\log M_{net} \leq d \log(\frac{5ep}{d}) \leq C_{net} \cdot s \log(\frac{p}{s}),$$

for some $C_{net} > 0$, when $s = o(p)$. □

C.2.3 Concentration of the mixing proportion

Since $\Delta = \sqrt{\boldsymbol{\beta}^{*\top} \Omega^* \boldsymbol{\beta}^*} \leq \sqrt{M} M_b$ is bounded. Let's denote $C_\Delta = \sqrt{M} M_b$ in the following proof.

Recall that

$$\begin{aligned}
\hat{\omega}(\boldsymbol{\theta}) =& \frac{1}{n} \sum_{i=1}^{n} \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\omega}{\omega + (1-\omega) \exp\{\boldsymbol{\delta}^\top \Omega(\boldsymbol{z}^{(i)} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)\}} \\
=& \frac{1}{n} \sum_{i=1}^{n} \frac{\omega}{\omega + (1-\omega) \exp\{\boldsymbol{\delta}^\top \Omega(\boldsymbol{z}^{(i)} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)\}} \\
:=& \frac{1}{n} \sum_{i=1}^{n} \frac{\omega}{\omega + (1-\omega) \exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}},
\end{aligned}$$

whose corresponding population estimate is $\omega(\boldsymbol{\theta}) = \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]$ and

$$C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}) = \boldsymbol{\delta}^\top \Omega \Big\{ \boldsymbol{z}^{(i)} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \Big\}.$$

We show next that with probability at least $1 - o(1)$,

$$\sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} |\hat{\omega}(\boldsymbol{\theta}) - \omega(\boldsymbol{\theta})| \lesssim \sqrt{\frac{s \log p}{n}}.$$

Let

$$\boldsymbol{z}_\omega = \sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} |\hat{\omega}(\boldsymbol{\theta}) - \omega(\boldsymbol{\theta})|.$$

Letting $\{\epsilon_i\}_{i=1}^n$ denote an *i.i.d.* sequence of Rademacher variables, for any $\lambda > 0$, we have

$$\mathbb{E}[e^{\lambda \boldsymbol{z}_\omega}] \le \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}})\Big|\right)\right].$$

It is easy to check that $\psi(x) = \omega/(\omega + (1-\omega)e^x) - \omega$ is Lipschitz with constant $\frac{1-\omega}{\omega} \le \frac{1-c_0}{c_0}$ and $\psi(0) = 0$. Consequently, by the Ledoux-Talagrand contraction for Rademacher processes in Lemma C.1 with $g(\cdot) = 1$, we have

$$\begin{aligned}
&\mathbb{E}[e^{\lambda \boldsymbol{z}_\omega}] \le \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}})\Big|\right)\right]\\
\le&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}}-\omega)\Big|\right)\right]\\
\cdot&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\omega\Big|\right)\right]\\
\le&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\cdot\frac{1-c_0}{c_0}\cdot C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\Big|\right)\right]\cdot\exp(\frac{\lambda^2}{n})\\
=&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\cdot\frac{1-c_0}{c_0}\cdot\Big\{\boldsymbol{\beta}^\top\boldsymbol{z}^{(i)}-\frac{1}{2}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\Big\}\Big|\right)\right]\cdot\exp(\frac{\lambda^2}{n})\\
\le&\underbrace{\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\boldsymbol{\beta}^\top(\boldsymbol{z}^{(i)}-(1-\omega^*)\boldsymbol{\mu}_1^*-\omega^*\boldsymbol{\mu}_2^*)\Big|\right)\right]}_{(i)}\\
&\cdot 2\underbrace{\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\exp\left(\frac{\lambda^2}{n}\Big|\frac{1-c_0}{c_0}\boldsymbol{\beta}^\top(\omega^*\boldsymbol{\mu}_1^*+(1-\omega^*)\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\Big|^2\right)}_{(ii)}\cdot\exp(\frac{\lambda^2}{n}),
\end{aligned}$$

where the last inequality uses the property of the sub-gaussian norm of bounded random variables.

We bound term (ii) first, Recall that $\delta_1(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$, and $\delta_2(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_2^* -$

$\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$), we have

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} |\omega^*\delta_1(\boldsymbol{\beta}) + (1-\omega^*)\delta_2(\boldsymbol{\beta})| \\ =&\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} |\boldsymbol{\beta}^\top(\omega^*\boldsymbol{\mu}_1^* + (1-\omega^*)\boldsymbol{\mu}_2^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})| \le c_1\Delta^2\end{aligned}$$

Consequently,

$$(ii) \le \exp(\frac{\lambda^2}{n}\cdot(\frac{1-c_0}{c_0})^2 c_1^2 C_\Delta^4).$$

Next, we proceed to bound term (i).

Let $\boldsymbol{z}_N^{(i)} = \boldsymbol{z}^{(i)} - ((1-\omega^*)\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*)$, then

$$\begin{aligned}(i) =&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\boldsymbol{\beta}^\top(\boldsymbol{z}^{(i)} - (1-\omega^*)\boldsymbol{\mu}_1^* - \omega^*\boldsymbol{\mu}_2^*)\Big|\right)\right] \\ =&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\boldsymbol{\beta}^\top\boldsymbol{z}_N^{(i)}\Big|\right)\right] \\ =&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\cdot\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\|\boldsymbol{\beta}\|_1\sup_{j\in[p]}|\sum_{i=1}^n \epsilon_i\boldsymbol{z}_{N_j}^{(i)}|\right)\right] \\ \le&\sum_{j=1}^p\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\cdot(M_b+C_b\Delta)|\sum_{i=1}^n \epsilon_i\boldsymbol{z}_{N_j}^{(i)}|\right)\right] \\ \le&\exp\left\{\frac{\lambda^2(M+C_b\Delta)^2}{n}(\frac{1-c_0}{c_0})^2\cdot C\Delta^2 + \log p\right\} \\ \le&\exp\left\{\frac{\lambda^2(M+C_b\Delta)^2}{n}(\frac{1-c_0}{c_0})^2\cdot C\Delta^2 + \log p\right\},\end{aligned}$$

Then we obtain

$$\begin{aligned}&\mathbb{E}[e^{\lambda\boldsymbol{z}_\omega}] \le \exp\left\{\frac{\lambda^2(M_b+C_b\Delta)^2}{n}(\frac{1-c_0}{c_0})^2\cdot C\Delta^2 + \log p\right\}\cdot\exp(\frac{\lambda^2}{n}\cdot(\frac{1-c_0}{c_0})^2 c_1^2 C_\Delta^4)\cdot\exp(\frac{\lambda^2}{n}) \\ :=& \exp(\frac{\lambda^2}{n}C_{con}^{(\omega)} + \log(p))\end{aligned}$$

Using the Chernoff approach, letting $\lambda = \sqrt{n\log p/C_{con}^{(\omega)}}$ and $t = 3\sqrt{C_{con}^{(w)}}\cdot\sqrt{\log p/n}$, the above bound on the moment generating function (MGF) implies that,

$$\begin{aligned}
&\mathbb{P}(\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} |\hat{\omega}(\boldsymbol{\theta})-\omega(\boldsymbol{\theta})|>t)=\mathbb{P}(\boldsymbol{z}_\omega>t)\le e^{-\lambda t}\mathbb{E}[e^{\lambda \boldsymbol{z}_\omega}]\\
\le&\exp(-\lambda t)\cdot\exp\left(\frac{\lambda^2}{n}C_{con}^{(\omega)}+\log p\right)\le e^{-3\log p+2\log p}\\
=&e^{-\log p}=o(1). \qquad \text{(C.27)}
\end{aligned}$$

Therefore, as long as $n\ge C_1\log p$ for a sufficiently large constant C_1, we have with probability at least $1-o(1)$,

$$\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} |\hat{\omega}(\boldsymbol{\theta})-\omega(\boldsymbol{\theta})|=\boldsymbol{z}_\omega\lesssim\sqrt{\frac{\log p}{n}}.$$

C.2.4 Concentration of the mean under $\|\cdot\|_{2,s}$ norm

Without loss of generality, we provide here the derivation for $\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})$, as the derivation for $\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})$ is similar. Recall that

$$\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})=\left\{\sum_{j=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\right\}^{-1}\left\{\sum_{i=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\boldsymbol{z}^{(i)}\right\},$$

whose corresponding population estimate is

$$\boldsymbol{\mu}_2(\boldsymbol{\theta})=\frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}.$$

Firstly, we observe that

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}_2(\boldsymbol{\theta})&=\left\{\sum_{j=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\right\}^{-1}\left\{\sum_{i=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\boldsymbol{z}^{(i)}\right\}-\frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}\\
&=\left\{\sum_{j=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\right\}^{-1}\left\{\sum_{i=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\right\}-\frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]}
\end{aligned}$$

Let

$$W^{(\boldsymbol{\mu})}=\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\|\frac{1}{n}\sum_{i=1}^n\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})-\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})]\|_{2,s},\quad j=1,2,...,p,$$

and

$$W_{\boldsymbol{u}}^{(\boldsymbol{\mu})}=\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\langle\frac{1}{n}\sum_{i=1}^{n}\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})-\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})],\boldsymbol{u}\rangle,\quad j=1,2,...,p,$$

where $\boldsymbol{u}\in\Gamma(s)\cap\mathbb{S}^{p-1}$, where $\mathbb{S}^{p-1}=\{\boldsymbol{u}\in\mathbb{R}^p:\|\boldsymbol{u}\|_2=1\}$.

Let $\{\boldsymbol{u}^1,...,\boldsymbol{u}^{M_{net}}\}$ denote a 1/2-net of the $\Gamma(s)\cap\mathbb{S}^{p-1}$. This means that for any $\boldsymbol{v}\in\Gamma(s)\cap\mathbb{S}^{p-1}$, there is some index $j\in[M_{net}]$ such that $\|\boldsymbol{v}-\boldsymbol{u}^j\|_2\le 1/2$. Similar to the derivation before, we have

$$W_{\boldsymbol{v}}^{(\boldsymbol{\mu})}\le W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}+|W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}-W_{\boldsymbol{v}}^{(\boldsymbol{\mu})}|\le\max_{j\in[M_{net}]}W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}+W^{(\boldsymbol{\mu})}\cdot\|\boldsymbol{v}-\boldsymbol{u}_j\|\le\max_{j\in[M_{net}]}W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}+\frac{1}{2}W^{(\boldsymbol{\mu})}.$$

It's followed by

$$W^{(\boldsymbol{\mu})}=\sup_{\boldsymbol{v}\in\Gamma(s)\cap\mathbb{S}^{p-1}}W_{\boldsymbol{v}}^{(\boldsymbol{\mu})}\le\max_{j\in[M_{net}]}W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}+\frac{1}{2}W^{(\boldsymbol{\mu})},$$

and then implies that

$$W^{(\boldsymbol{\mu})}\le 2\max_{j\in[M_{net}]}W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})}.$$

Consequently it suffices to bound $W_{\boldsymbol{u}}^{(\boldsymbol{\mu})}$ for a fixed $\boldsymbol{u}$. Let $\{\epsilon_i\}_{i=1}^n$ denote an *i.i.d.* sequence of Rademacher variables, for any $\lambda>0$, we have

$$\begin{aligned}\mathbb{E}[e^{\lambda W_{\boldsymbol{u}}^{(\boldsymbol{\mu})}}]\le&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i\frac{\omega\langle\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}}\Big|\right)\right]\\ \le&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}}-\omega)\langle\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\ &\cdot\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i\cdot\omega\cdot\langle\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\ \le&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}}-\omega)\langle\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\ &\cdot\exp\left(\frac{\lambda^2}{n}\tilde{C}_0+C_{net}\cdot\log(p/s)\right),\end{aligned}$$

where $C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})=\boldsymbol{\beta}^\top(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$ and the last inequality follows the same idea of the concentration of the proportion.

Let's recall $\boldsymbol{z}_N^{(i)}=\boldsymbol{z}^{(i)}-(1-\omega^*)\boldsymbol{\mu}_1^*-\omega^*\boldsymbol{\mu}_2^*$ be the centered random variable $\boldsymbol{z}^{(i)}$. Since $\psi(x)=\omega/(\omega+(1-\omega)e^x)-\omega$ is Lipschitz and $\psi(0)=0$, by Lemma C.1 and let

$\boldsymbol{\mu}^* := (1-\omega^*)\boldsymbol{\mu}_1^* + \omega^*\boldsymbol{\mu}_2^*$, we then have

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\big(\frac{\omega}{\omega+(1-\omega)\exp\{C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\}}-\omega\big)\langle \boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\
\leq&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i C_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\langle \boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\
=&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\Big\{\boldsymbol{\beta}^\top\boldsymbol{z}^{(i)}-\frac{1}{2}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\Big\}\langle \boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big|\right)\right]\\
\leq&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\beta},\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]\\
&\cdot\underbrace{\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\beta},\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]}_{(i)}\\
&\cdot\underbrace{\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\beta},\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]}_{(ii)}\\
&\cdot\underbrace{\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\beta}\in\Gamma(s)+\boldsymbol{\beta}^*}\Big|\frac{1-c_0}{c_0}\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{\beta}\rangle\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]}_{(iii)}\\
\leq&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n \epsilon_i\langle\boldsymbol{\beta},\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big|\right)\right]\cdot\exp\left\{\frac{\lambda^2}{n}\tilde{C}_1+s\log p\right\},
\end{aligned}
\tag{C.28}
$$

for some constant $\tilde{C}_1>0$. To see this, we need to bound (i), (ii), (iii).

We first have

$$
\Big|\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\langle\boldsymbol{\beta},(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\rangle\Big|\leq c_1\Delta^2.
$$

and thus

$$
\begin{aligned}
(i)=&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\beta},-(1-\omega^*)\boldsymbol{\mu}_1^*-\omega^*\boldsymbol{\mu}_2^*\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]\\
=&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\frac{1-c_0}{c_0}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\langle\boldsymbol{\beta},-(1-\omega^*)\boldsymbol{\mu}_1^*-\omega^*\boldsymbol{\mu}_2^*\rangle\cdot\Big|\sum_{i=1}^n \epsilon_i\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big|\right)\right]\\
\leq&\exp\left\{\frac{\lambda^2}{n}\tilde{C}_{11}\right\},
\end{aligned}
$$

where $\tilde{C}_{11} = (\frac{1-c_0}{c_0})^2 \cdot c_1^2 C_\Delta^4$.

The bound for the second term (ii) follows the same covering idea

$$
\begin{aligned}
(ii) =& \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n \epsilon_i\Big\{\langle\boldsymbol{\beta},\boldsymbol{z}_N^{(i)}\rangle\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]\\
\leq& \mathbb{E}\left[\exp\left(\frac{\lambda}{n}|\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\frac{1-c_0}{c_0}\cdot\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^n\epsilon_i\boldsymbol{\beta}^\top\boldsymbol{z}_N^{(i)}\Big|\right)\right]\\
\leq& \mathbb{E}\left[\exp\left(\frac{\lambda}{n}|\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\frac{1-c_0}{c_0}\cdot\Big|\sum_{i=1}^n\epsilon_i\boldsymbol{\beta}^{*\top}\boldsymbol{z}_N^{(i)}\Big|\right)\right]\\
&\cdot\mathbb{E}\left[\exp\left(\frac{\lambda}{n}|\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\frac{1-c_0}{c_0}\cdot\sup_{\boldsymbol{u}\in\Gamma(s)}\Big|\sum_{i=1}^n\epsilon_i\boldsymbol{u}^\top\boldsymbol{z}_N^{(i)}\Big|\right)\right]
\end{aligned}
$$

Let $\{\tilde{u}^1,...,\tilde{u}^{M_{net}}\}$ be the $\frac{1}{2}$-net of $\Gamma(s)\cap\mathbb{S}^{p-1}$, then similar to (**??**) and use Lemma C.2, we get

$$
\begin{aligned}
(ii) \leq& \exp\left(\frac{\lambda^2}{n}|\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\frac{1-c_0}{c_0}\cdot\Big|\sum_{i=1}^n\epsilon_i\boldsymbol{\beta}^{*\top}\boldsymbol{z}_N^{(i)}\Big|\right)\\
&\cdot\mathbb{E}\left[\exp\left(\frac{\lambda}{n}|\langle(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\frac{1-c_0}{c_0}\cdot 2C_bC_\Delta\sup_{j\in[M_{net}]}\Big|\sum_{i=1}^n\epsilon_i\tilde{u}^{j\top}\boldsymbol{z}_N^{(i)}\Big|\right)\right]\\
\leq& \exp\left\{\frac{\lambda^2}{n}(\frac{1-c_0}{c_0})^2C_b^2MC_\Delta^4\right\}\cdot\sum_{j\in[M_{net}]}\mathbb{E}\left[\exp\left(\frac{2\lambda}{n}\cdot\frac{1-c_0}{c_0}C_b^2C_\Delta^2\sup_{j\in[p]}|\sum_{i=1}^n\epsilon_i\tilde{u}^{j\top}\boldsymbol{z}_N^{(i)}|\right)\right]\\
\leq& \exp\left\{\frac{4\lambda^2}{n}(\frac{1-c_0}{c_0})^2\cdot(C_b^2M+C_b^4)C_\Delta^4+\log M_{net}\right\}\\
:=& \exp\left\{\frac{\lambda^2}{n}\tilde{C}_{12}+C_{net}\cdot s\log(p/s)\right\},
\end{aligned}
$$

where $\tilde{C}_{12} = 4(\frac{1-c_0}{c_0})^2\cdot(C_b^2M+C_b^4)C_\Delta^4$.

Term (iii) is bounded by

$$
\begin{aligned}
(iii) =& \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\frac{1-c_0}{c_0}\cdot\sum_{i=1}^n\epsilon_i\Big\{\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{\beta}\rangle\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{u}\rangle\Big\}\Big|\right)\right]\\
\leq& \exp\left\{\frac{\lambda^2}{n}(\frac{1-c_0}{c_0})^2c_1^2MC_b^2C_\Delta^6\right\}\\
=& \exp\left\{\frac{\lambda^2}{n}\tilde{C}_{13}\right\},
\end{aligned}
$$

where $\tilde{C}_{13} = (\frac{1-c_0}{c_0})^2(C_b+\sqrt{M})\cdot MC_b^2C_\Delta^6$ and we use the fact that $|\langle\boldsymbol{\mu}^*-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2},\boldsymbol{\beta}\rangle|\leq c_1C_\Delta^2$, and $\|\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_1\|_{2,s},\|\boldsymbol{\mu}_2^*-\boldsymbol{\mu}_2\|_{2,s}\leq C_b\Delta\leq\sqrt{M}C_\Delta$.

Combining the bounds for these three terms $(i), (ii)$ and (iii), we obtain (C.28) with $\tilde{C}_1 = \tilde{C}_{11} + \tilde{C}_{12} + \tilde{C}_{13}$.

Then we proceed to bound

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i\langle\boldsymbol{\beta},\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big|\right)\right]\\
\leq&\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\Big|\sum_{i=1}^{n}\epsilon_i\langle\boldsymbol{\beta}-\boldsymbol{\beta}^*,\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big|\right)\right]\\
\cdot&\left[\exp\left(\frac{\lambda}{n}\cdot\frac{1-c_0}{c_0}\Big|\sum_{i=1}^{n}\epsilon_i\langle\boldsymbol{\beta}^*,\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\Big|\right)\right]
\end{aligned}$$

When $\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$, $\boldsymbol{\beta} - \boldsymbol{\beta}^* \in \Gamma(s)$, which implies $\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Gamma(s)$. Now let us define

$$\tilde{W}_{\boldsymbol{u}} = \sup_{\tilde{\boldsymbol{u}}\in\Gamma(s)\cap\mathbb{S}^{p-1}}\langle\tilde{\boldsymbol{u}},\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{z}_c^{(i)}\boldsymbol{z}_c^{(i)\top}\boldsymbol{u}\rangle,$$

and

$$\tilde{W}_{\tilde{\boldsymbol{u}},\boldsymbol{u}} = \langle\tilde{\boldsymbol{u}},\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\boldsymbol{z}_c^{(i)}\boldsymbol{z}_c^{(i)\top}\boldsymbol{u}\rangle.$$

In the meantime, let $\{\tilde{\boldsymbol{u}}^1, ..., \tilde{\boldsymbol{u}}^{M_{net}}\}$ be the $\frac{1}{2}$-net of $\Gamma(s) \cap \mathbb{S}^{p-1}$. Then we have

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\boldsymbol{\beta}-\boldsymbol{\beta}^*,\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle \leq&(\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|)\cdot\tilde{W}_{\boldsymbol{u}}\\
\leq&C_bC_\Delta\cdot\tilde{W}_{\boldsymbol{u}}\\
\leq&2C_bC_\Delta\cdot\max_{j\in[M_{net}]}\tilde{W}_{\tilde{\boldsymbol{u}}^j,\boldsymbol{u}},
\end{aligned}$$

where the last inequality uses the same covering argument as before.

Then let's proceed to bound $\langle\tilde{\boldsymbol{u}}, \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_c^{(i)}\boldsymbol{z}_c^{(i)\top}\boldsymbol{u}\rangle$ for fixed $\tilde{\boldsymbol{u}}, \boldsymbol{u}$. Following the Lemma D.2 in Wang et al. (2015), we can similarly calculate the sub-exponential norm ($\|\cdot\|_{\psi_1}$) of $|\langle\tilde{\boldsymbol{u}}, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle|$, which yields, for some constant $C_\psi > 0$,

$$\begin{aligned}
\|\frac{1}{n}\langle\tilde{\boldsymbol{u}},\boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\|_{\psi_1}\leq&\frac{1}{n}C_\psi\cdot\max\{\|\langle\boldsymbol{z}_N^{(i)},\tilde{\boldsymbol{u}}\rangle\|_{\psi_2}^2,\|\langle\boldsymbol{z}_N^{(i)},\boldsymbol{u}\rangle\|_{\psi_2}^2\}\\
\leq&\frac{1}{n}C_\psi\cdot\max\{\langle\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*,\tilde{\boldsymbol{u}}\rangle,\tilde{\boldsymbol{u}}^\top\Sigma^*\tilde{\boldsymbol{u}},\|\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*\|_{2,s},\|\Sigma^*\|_2\}\\
\leq&\frac{C_\psi\cdot[\sqrt{M}C_\Delta+M]}{n}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}\|\frac{1}{n}\langle\boldsymbol{\beta}^*, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle\|_{\psi_1} \leq& \frac{1}{n}C_\psi \cdot \max\{\|\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{\beta}^*\rangle\|_{\psi_2}^2, \|\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle\|_{\psi_2}^2\} \\ \leq& \frac{1}{n}C_\psi \cdot \max\{\langle\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*, \boldsymbol{\beta}^*\rangle, \boldsymbol{\beta}^{*\top}\Sigma^*\boldsymbol{\beta}^*, \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{2,s}, \|\Sigma^*\|_2\} \\ \leq& \frac{C_\psi \cdot [C_\Delta^2 + \sqrt{M}C_\Delta + M]}{n}.\end{aligned}$$

Consequently, for sufficiently small λ

$$\begin{aligned}&\mathbb{E}\left[\exp\Big(\frac{\lambda}{n} \cdot \frac{1-c_0}{c_0} \sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} |\sum_{i=1}^n \epsilon_i \langle\boldsymbol{\beta}, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle||\Big)\right] \\ \leq& \mathbb{E}\left[\exp\Big(\frac{\lambda}{n} \cdot \frac{1-c_0}{c_0} \cdot 2C_bC_\Delta \cdot \sup_{j \in [M_{net}]} |\sum_{i=1}^n \epsilon_i \langle\tilde{\boldsymbol{u}}^j, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle||\Big)\right] \\ &\cdot \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \cdot \frac{1-c_0}{c_0}\Big|\sum_{i=1}^n \epsilon_i \langle\boldsymbol{\beta}^*, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}\rangle\Big|\right)\right] \\ \leq& \exp\Big\{\frac{\lambda^2}{n} \cdot (\frac{1-c_0}{c_0})^2 \cdot (2C_bC_\Delta)^2 \cdot C_\psi^2 \cdot [(\sqrt{M}C_\Delta + M)^2 + (C_\Delta^2 + \sqrt{M}C_\Delta + M)^2] + \log M_{net}\Big\} \\ =& \exp\Big\{\frac{\lambda^2}{n} \cdot \tilde{C}_2 + \log M_{net}\Big\}, \qquad \text{(C.29)}\end{aligned}$$

where $\tilde{C}_2 = (\frac{1-c_0}{c_0})^2 \cdot (2C_bC_\Delta)^2 \cdot C_\psi^2 \cdot [(\sqrt{M}C_\Delta + M)^2 + (C_\Delta^2 + \sqrt{M}C_\Delta + M)^2]$.

Putting the pieces together yields

$$\begin{aligned}&\mathbb{E}[\exp(\lambda W^{(\boldsymbol{\mu})})] \leq \mathbb{E}[\exp(2\lambda \max_{j \in [M_{net}]} W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})})] \leq \sum_{j \in [M_{net}]} \mathbb{E}[\exp(2\lambda W_{\boldsymbol{u}_j}^{(\boldsymbol{\mu})})] \\ \leq& \sum_{j \in [M_{net}]} \mathbb{E}\left[\exp\left(\frac{2\lambda}{n} \cdot \frac{1-c_0}{c_0} \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Gamma(s)} \Big|\sum_{i=1}^n \epsilon_i \langle\boldsymbol{\beta}, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}_j\rangle\Big|\right)\right] \cdot \exp\left\{\frac{4\lambda^2}{n}(\tilde{C}_0 + \tilde{C}_1) + 2C_{net} \cdot s\log(p/s)\right\} \\ \leq& \sum_{j \in [M_{net}]} \mathbb{E}\left[\exp\left(\frac{2\lambda}{n} \cdot \frac{1-c_0}{c_0} \sup_{\boldsymbol{\beta} \in \boldsymbol{\beta}^* + \Gamma(s)} \Big|\sum_{i=1}^n \epsilon_i \langle\boldsymbol{\beta}, \boldsymbol{z}_N^{(i)}\rangle\langle\boldsymbol{z}_N^{(i)}, \boldsymbol{u}_j\rangle\Big|\right)\right] \cdot \exp\left\{\frac{4\lambda^2}{n}\tilde{(}\tilde{C}_0 + \tilde{C}_1) + 2C_{net} \cdot s\log(p/s)\right\} \\ \leq& 2M_{net}\exp\Big\{\frac{4\lambda^2}{n} \cdot \tilde{C}_2 + \log M_{net}\Big\} \cdot \exp\Big\{\frac{4\lambda^2}{n}(\tilde{C}_0 + \tilde{C}_1) + 2C_{net} \cdot s\log(p/s)\Big\} \\ \leq& \exp\Big\{\frac{4\lambda^2}{n} \cdot (\tilde{C}_0 + \tilde{C}_1 + \tilde{C}_2) + 4C_{net} \cdot s\log(p/s)\Big\}.\end{aligned}$$

Using the Chernoff approach, we then have with probability $1 - o(1)$,

$$W^{(\boldsymbol{\mu})} \lesssim \sqrt{\frac{s\log p}{n}}.$$

Combining with the concentration of $\hat{\omega}(\boldsymbol{\theta})$, we thus have with probability at least $1-2p^{-1}$,

$$
\begin{aligned}
\|\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}_2(\boldsymbol{\theta})\|_{2,s} =& \Big\|\Big\{\sum_{j=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\Big\}^{-1}\Big\{\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\Big\} \\
& -\{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]\}^{-1}\,\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})]\Big\|_{2,s} \\
\le& \Big\{\frac{1}{n}\sum_{j=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\Big\}^{-1}\|\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})]-\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})\|_{2,s}+ \\
& \left|\{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]\}^{-1}-\Big\{\frac{1}{n}\sum_{j=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(j)})\Big\}^{-1}\right|\cdot\|\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)(Z-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})]\|_{2,s} \\
\lesssim& \sqrt{\frac{s\log p}{n}}.
\end{aligned}
$$

Similarly, with probability at least $1-o(1)$,

$$
\sup_{\boldsymbol{\theta}\in B_{con}(\boldsymbol{\theta}^*;c_0,c_1,C_b,s)} \|\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})-\mu_1(\boldsymbol{\theta})\|_{2,s}\lesssim\sqrt{\frac{s\log p}{n}}.
$$

C.2.5 Concentration of the covariance matrix

In this section we prove the concentration of the covariance matrix in terms of $\|\{\hat{\Sigma}(\boldsymbol{\theta})-\Sigma(\boldsymbol{\theta})\}\boldsymbol{\beta}^*\|_{2,s}$.

Recall that $\boldsymbol{\mu}^*=(1-\omega^*)\boldsymbol{\mu}_1^*+\omega^*\boldsymbol{\mu}_2^*$, we then have

$$
\begin{aligned}
\hat{\Sigma}(\boldsymbol{\theta}) =& \frac{1}{n}\sum_{i=1}^n\Big\{(1-\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)}))(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta}))(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta}))^\top+\gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta}))(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta}))^\top\Big\} \\
=& \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^{(i)}[\boldsymbol{z}^{(i)}]^\top-\Big\{1-\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\Big\}\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})^\top-\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})^\top \\
=& \frac{1}{n}\sum_{i=1}^n(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}^*)[\boldsymbol{z}^{(i)}-\boldsymbol{\mu}^*]^\top-\Big\{1-\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})\Big\}(\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top \\
& -\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta}}(\boldsymbol{z}^{(i)})(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top
\end{aligned}
$$

where the last equality uses the fact that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^{(i)}=(1-\hat{\omega}(\boldsymbol{\theta}))\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})+\hat{\omega}(\boldsymbol{\theta})\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})$.

Similarly, the corresponding population estimate is

$$\begin{aligned}\Sigma(\boldsymbol{\theta}) =&\mathbb{E}\left[(1-\gamma_{\boldsymbol{\theta}}(Z))(Z-\boldsymbol{\mu}_1(\boldsymbol{\theta}))(Z-\boldsymbol{\mu}_1(\boldsymbol{\theta})^\top+\gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2(\boldsymbol{\theta})(Z-\boldsymbol{\mu}_2(\boldsymbol{\theta}))^\top\right]\\ =&\mathbb{E}[ZZ^\top]-\mathbb{E}[1-\gamma_{\boldsymbol{\theta}}(Z)]\boldsymbol{\mu}_1(\boldsymbol{\theta})\boldsymbol{\mu}_1(\boldsymbol{\theta}))^\top-\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]\boldsymbol{\mu}_2(\boldsymbol{\theta})\boldsymbol{\mu}_2(\boldsymbol{\theta})^\top\\ =&\mathbb{E}[(Z-\boldsymbol{\mu}^*)(Z-\boldsymbol{\mu}^*)^\top]-\mathbb{E}[1-\gamma_{\boldsymbol{\theta}}(Z)](\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\\ &-\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)](\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\end{aligned}$$

where the last equality uses the fact that $\mathbb{E}[Z]=(1-\omega(\boldsymbol{\theta}))\boldsymbol{\mu}_1(\boldsymbol{\theta})+\omega(\boldsymbol{\theta})\boldsymbol{\mu}_2(\boldsymbol{\theta})$.

Therefore by triangle inequality

$$\begin{aligned}\|(\hat{\Sigma}(\boldsymbol{\theta})-\Sigma(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_{2,s}\le&\underbrace{\|\{\frac{1}{n}\sum_{i=1}^n(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}^*)(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}^*)^\top-\mathbb{E}[(Z-\boldsymbol{\mu}^*)(Z-\boldsymbol{\mu}^*)^\top]\}\boldsymbol{\beta}^*\|_{2,s}}_{(i)}\\ +&\underbrace{\|\left[\{1-\hat{\omega}(\boldsymbol{\theta})\}(\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top-(1-\omega^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\right]\boldsymbol{\beta}^*\|_{2,s}}_{(ii)}\\ +&\underbrace{\|\left[\hat{\omega}(\boldsymbol{\theta})(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top-\omega^*(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\right]\boldsymbol{\beta}^*\|_{2,s}}_{(iii)},\end{aligned}\tag{C.30}$$

Term (i) can be bounded similarly to (C.29). We have with probability $1-o(1)$,

$$\|\{\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^{(i)}(\boldsymbol{z}^{(i)})^\top-\mathbb{E}[ZZ^\top]\}\boldsymbol{\beta}^*\|_{2,s}\lesssim\sqrt{\frac{s\log p}{n}}$$

Further, for term (iii) in (C.30), by the concentration results for $\hat{\omega}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})$, we have with probability at least $1-o(1)$,

$$\begin{aligned}(iii)=&\|\left[\hat{\omega}(\boldsymbol{\theta})(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top-\omega^*(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\right]\boldsymbol{\beta}^*\|_{2,s}\\ \le&\|(\hat{\omega}(\boldsymbol{\theta})-\omega(\boldsymbol{\theta}))(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\boldsymbol{\beta}^*\|_{2,s}\\ &+\|\hat{\omega}(\boldsymbol{\theta})\left[(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top-(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\right]\boldsymbol{\beta}^*\|_{2,s}\\ \le&|\hat{\omega}(\boldsymbol{\theta})-\omega(\boldsymbol{\theta})|\cdot\|(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*\|_{2,s}\cdot\|\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)\|_{2,s}\cdot\|\boldsymbol{\beta}^*\|_2\\ &+\|\hat{\omega}(\boldsymbol{\theta})\left[(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}_2(\boldsymbol{\theta}))^\top\right]\boldsymbol{\beta}^*\|_{2,s}\\ &+\|\hat{\omega}(\boldsymbol{\theta})\left[(\hat{\boldsymbol{\mu}}_2(\boldsymbol{\theta})-\boldsymbol{\mu}_2(\boldsymbol{\theta}))(\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*)^\top\right]\boldsymbol{\beta}^*\|_{2,s}\\ \lesssim&\sqrt{\frac{s\log p}{n}}.\end{aligned}$$

Similarly, with probability at least $1-o(1)$,

$$(ii)\lesssim\sqrt{\frac{s\log p}{n}}.$$

Therefore, with probability at least $1 - o(1)$,

$$\sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} \|(\hat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_{2,s} \lesssim \sqrt{\frac{s \log p}{n}}.$$

Similarly, we have with probability at least $1 - o(1)$,

$$\sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} \|(\hat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_{\infty} \lesssim \sqrt{\frac{\log p}{n}}.$$

C.2.6 Proof of Lemma C.4

Proof. Lemma C.4 is an analog of Lemma 13 in Rudelson and Zhou (2012) with a few modifications. For completeness, we present the proof here.

We first state two lemmas.

Lemma C.5 (Lemma 11 in Rudelson and Zhou (2012))**.** *Let* $\boldsymbol{u}_1, ..., \boldsymbol{u}_M \in \mathbb{R}^q$*. Let* $\boldsymbol{y} \in \operatorname{conv}(\boldsymbol{u}_1, ..., \boldsymbol{u}_M)$*. There exists a set* $L \subset [M]$*, such that*

$$|L| \leq m = \frac{4 \max_{j \in [M]} \|\boldsymbol{u}_j\|_2^2}{e\epsilon^2}$$

and a vector $\boldsymbol{y}' \in \operatorname{conv}(\boldsymbol{u}_j : j \in L\}$*, such that*

$$\|\boldsymbol{y}' - \boldsymbol{y}\|_2 \leq \epsilon.$$

Lemma C.6 (Lemma 21 in Rudelson and Zhou (2012))**.** *Let* $\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{x} \in \mathbb{R}^q$ *be vectors such that i).* $\|\boldsymbol{\theta}\|_2 = 1$*, ii).*$\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \neq 0$*, (iii)* $\boldsymbol{u}$ *is not parallel to* $\boldsymbol{x}$*. Define* $\phi : \mathbb{R} \to \mathbb{R}$ *by:*

$$\phi(\lambda) = \frac{\langle \boldsymbol{x} + \lambda \boldsymbol{u}, \boldsymbol{\theta} \rangle}{\|\boldsymbol{x} + \lambda \boldsymbol{u}\|_2}.$$

Assume $\phi(\lambda)$ *has a local maximum at* 0*, then*

$$\frac{\langle \boldsymbol{x} + \boldsymbol{u}, \boldsymbol{\theta} \rangle}{\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle} \geq 1 - \frac{\|\boldsymbol{u}\|_2}{\|\boldsymbol{x}\|_2}.$$

Then let's proceed to prove Lemma C.4. Without loss of generality, assume that $d < p$, otherwise the lemma is trivially true. For each vector $\boldsymbol{u} \in \mathbb{R}^p$, let T_0 denote the locations of the s largest coefficients of $\boldsymbol{u}$ in absolute values. Decompose a vector $\boldsymbol{u} \in \Gamma(s) \cap \mathbb{S}^{p-1}$ as

$$\boldsymbol{u} = \boldsymbol{u}_{T_0} + \boldsymbol{u}_{T_0^c} \in \boldsymbol{u}_{T_0} + (2\|\boldsymbol{u}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{u}\|_2) \cdot \operatorname{absconv}(\boldsymbol{e}_j : j \in T_0^c),$$

where absconv$(\cdot)$ denotes the absolutely convex set. Since

$$\begin{aligned}\|\boldsymbol{u}_{T_0^c}\|_2^2 \leq & \|\boldsymbol{u}_{T_0^c}\|_1 \|\boldsymbol{u}_{T_0^c}\|_\infty \leq (2\|\boldsymbol{u}_{T_0}\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{u}\|_2) \cdot \frac{\|\boldsymbol{u}_{T_0}\|_1}{s} \\ \leq & (2\|\boldsymbol{u}_{T_0}\|_2 + 3/2 \cdot \|\boldsymbol{u}\|_2) \cdot \|\boldsymbol{u}_{T_0}\|_2,\end{aligned}$$

then we have

$$1 = \|\boldsymbol{u}\|_2^2 = \|\boldsymbol{u}_{T_0}\|_2^2 + \|\boldsymbol{u}_{T_0^c}\|_2^2 \leq 3\|\boldsymbol{u}_{T_0}\|_2^2 + 3/2 \cdot \|\boldsymbol{u}\|_2 \cdot \|\boldsymbol{u}_{T_0}\|_2,$$

which implies $\|\boldsymbol{u}_{T_0}\|_2 \geq \frac{1}{3}$.

Let's define

$$V = \{\boldsymbol{u}_{T_0} + (2\|\boldsymbol{u}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{u}\|_2) \cdot \text{absconv}(\boldsymbol{e}_j : j \in T_0^c) : \boldsymbol{u} \in \Gamma(s) \cap \mathbb{S}^{p-1}\}.$$

We then have $\Gamma(s) \cap \mathbb{S}^{p-1} \subset V \subset \Gamma(s)$ and V is compact. Therefore, V contains a base of $\Gamma(s)$, that is, for any $\boldsymbol{y} \in \Gamma(s)\backslash\{0\}$, there exists $\lambda > 0$ such that $\lambda y \in V$.

For any nonzero $\boldsymbol{v} \in \mathbb{R}^p$, let's define

$$F(\boldsymbol{v}) = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}.$$

The function F is continuous on $\Gamma(s)\backslash\{0\}$, and in particular, on V. Hence,

$$\Gamma(s) \cap \mathbb{S}^{p-1} = F(\Gamma(s)\backslash\{0\}) = F(V).$$

By duality, inclusion (C.26) can be derived from the fact that the supremum of any linear functional over the left side of (C.26) does not exceed the supremum over the right side of it. By the equality above, it is enough to show that for any $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$, there exists $\boldsymbol{z}' \in \mathbb{R}^p\backslash\{0\}$ such that $\text{supp}(\boldsymbol{z}') \leq d$ and $F(\boldsymbol{z}')$ is well defined, which satisfies

$$\max_{\boldsymbol{v} \in V} \langle F(\boldsymbol{v}), \boldsymbol{\theta} \rangle \leq 2\langle F(\boldsymbol{z}'), \boldsymbol{\theta} \rangle. \tag{C.31}$$

For a given $\boldsymbol{\theta}$, we construct a d-sparse vector $\boldsymbol{z}'$ which satisfies (C.31). Let

$$\boldsymbol{z} = \arg\max_{\boldsymbol{v} \in V} \langle F(\boldsymbol{v}), \boldsymbol{\theta} \rangle.$$

By definition of V, there exists $I \subset [p]$ such that $|I| = s$, and for some $\eta_j \in \{-1, 1\}$,

$$\boldsymbol{z} = \boldsymbol{z}_I + (2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2) \cdot \sum_{j \in I^c} \alpha_j \eta_j \boldsymbol{e}_j,$$

where $\alpha_j \in [0,1], \sum_{j \in I^c} \alpha_j \leq 1$, and

$$1 \geq \|\boldsymbol{z}_I\|_2 \geq \frac{1}{3}. \tag{C.32}$$

Note if $\alpha_i = 1$ for some $i \in I^c$, then $\boldsymbol{z}$ is a sparse vector itself, and we can set $\boldsymbol{z}' = \boldsymbol{z}$ in order for (C.31) to hold. We proceed by assuming $\alpha_i \in [0,1)$ for all $i \in I^c$ in (C.31) from now on, in which case, we construct a required sparse vector $\boldsymbol{z}'$ via Lemma C.5. To satisfy the assumptions of this lemma, denote $\boldsymbol{e}_{p+1} = \boldsymbol{0}$, $\eta_{p+1} = 1$ and set

$$\alpha_{p+1} = 1 - \sum_{j \in I^c} \alpha_j, \text{ hence } \alpha_{p+1} \in [0,1].$$

Let $\boldsymbol{y} = \boldsymbol{z}_{I^c}$, $\mathcal{M} = \{j \in I^c \cup \{p+1\} : \alpha_j > 0\}$, and $\epsilon > 0$ specified later. Applying Lemma C.5 with vector $\boldsymbol{u}_j = (2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2) \cdot \eta_j \boldsymbol{e}_j$ for $j \in \mathcal{M}$, construct a set $J' \subset \mathcal{M}$ satisfying

$$|J'| \leq m := \frac{4 \max_{j \in I^c}(2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2)^2 \|\boldsymbol{e}_j\|_2^2}{\epsilon^2} \leq \frac{64s}{\epsilon^2}, \tag{C.33}$$

and a vector

$$\boldsymbol{y}' = (2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2) \sum_{j \in J'} \beta_j \eta_j \boldsymbol{e}_j,$$

where for $J' \subset \mathcal{M}, \beta_j \in [0,1]$ and $\sum_{j \in J'} \beta_j = 1$, such that $\|\boldsymbol{y}' - \boldsymbol{y}\|_2 \leq \epsilon$.

Set $\boldsymbol{z}' = \boldsymbol{z}_{I^c} + \boldsymbol{y}'$. By construction, $\boldsymbol{z}' \in E_J$, where $J = (I \cup J') \cap [p]$ and $|J| \leq |I| + |J'| \leq s + m$. Furthermore, we have

$$\|\boldsymbol{z} - \boldsymbol{z}'\| = \|\boldsymbol{y} - \boldsymbol{y}'\| \leq \epsilon.$$

For $\{\beta_j : j \in J'\}$ as above, we extend it to $\{\beta_j : j \in I^c \cup \{p+1\}\}$ setting $\beta_j = 0$ if $j \in I^c \cup \{p+1\} \backslash J'$ and write

$$\boldsymbol{z}' = \boldsymbol{z}_I + (2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2) \sum_{j \in I^c \cup \{p+1\}} \beta_j \eta_j \boldsymbol{e}_j,$$

where $\beta_j \in [0,1]$ and $\sum_{j \in I^c \cup \{p+1\}} \beta_j = 1$.

If $\boldsymbol{z}' = \boldsymbol{z}$, the result holds. Otherwise, for some λ to be specified, consider the vector

$$\boldsymbol{z} + \lambda(\boldsymbol{z}' - \boldsymbol{z}) = \boldsymbol{z}_I + (2\|\boldsymbol{z}_S\|_1 + 3/2 \cdot \sqrt{s}\|\boldsymbol{z}\|_2) \sum_{j \in I^c \cup \{p+1\}} [(1-\lambda)\alpha_j + \lambda\beta_j] \eta_j \boldsymbol{e}_j,$$

We have $\sum_{j \in I^c \cup \{p+1\}} [(1-\lambda)\alpha_j + \lambda\beta_j] = 1$ and $(1-\lambda)\alpha_j + \lambda\beta_j \in [0,1]$. Therefore $\sum_{j \in I^c} [(1-\lambda)\alpha_j + \lambda\beta_j] \leq 1$ and $\boldsymbol{z} + \lambda(\boldsymbol{z}' - \boldsymbol{z}) \in V$.

Now let's consider a function ϕ,

$$\phi(\lambda) = \langle F(\boldsymbol{z} + \lambda(\boldsymbol{z}' - \boldsymbol{z})), \boldsymbol{\theta} \rangle = \frac{\langle \boldsymbol{z} + \lambda(\boldsymbol{z}' - \boldsymbol{z}), \boldsymbol{\theta} \rangle}{\|\boldsymbol{z} + \lambda(\boldsymbol{z}' - \boldsymbol{z})\|}.$$

Since $\boldsymbol{z}$ maximizes $\langle F(\boldsymbol{v}), \boldsymbol{\theta} \rangle$ for all $\boldsymbol{v} \in V$, $\phi(\lambda)$ attains the local minimum at $\lambda = 0$. Then by Lemma C.6, we have

$$\frac{\langle \boldsymbol{z}', \boldsymbol{\theta} \rangle}{\langle \boldsymbol{z}, \boldsymbol{\theta} \rangle} = \frac{\langle \boldsymbol{z} + (\boldsymbol{z}' - \boldsymbol{z}), \boldsymbol{\theta} \rangle}{\langle \boldsymbol{z}, \boldsymbol{\theta} \rangle} \geq 1 - \frac{\|\boldsymbol{z}' - \boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2} = \frac{\|\boldsymbol{z}\|_2 - \|\boldsymbol{z}' - \boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2}.$$

Consequently,

$$\begin{aligned}
\frac{\langle F(\boldsymbol{z}), \boldsymbol{\theta} \rangle}{\langle F(\boldsymbol{z}'), \boldsymbol{\theta} \rangle} =& \frac{\langle \boldsymbol{z}'/\|\boldsymbol{z}'\|, \boldsymbol{\theta} \rangle}{\langle \boldsymbol{z}/\|\boldsymbol{z}\|, \boldsymbol{\theta} \rangle} = \frac{\|\boldsymbol{z}\|_2}{\|\boldsymbol{z}'\|_2} \times \frac{\langle \boldsymbol{z}', \boldsymbol{\theta} \rangle}{\langle \boldsymbol{z}, \boldsymbol{\theta} \rangle} \\
\geq& \frac{\|\boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2 + \|\boldsymbol{z}' - \boldsymbol{z}\|_2} \times \frac{\|\boldsymbol{z}\|_2 - \|\boldsymbol{z}' - \boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2} \\
=& \frac{\|\boldsymbol{z}\|_2 - \|\boldsymbol{z}' - \boldsymbol{z}\|_2}{\|\boldsymbol{z}\|_2 + \|\boldsymbol{z}' - \boldsymbol{z}\|_2} \\
\geq& \frac{\|\boldsymbol{z}\|_2 - \epsilon}{\|\boldsymbol{z}\|_2 + \epsilon} = 1 - \frac{2\epsilon}{\|\boldsymbol{z}\|_2 + \epsilon}.
\end{aligned}$$

By the dentition of $\boldsymbol{z}$ and (C.32), we have $\|\boldsymbol{z}\| \geq \|\boldsymbol{z}_I\| \geq \frac{1}{3}$.

Now we set $\epsilon = \frac{1}{9}$, then we have

$$\frac{\langle F(\boldsymbol{z}), \boldsymbol{\theta} \rangle}{\langle F(\boldsymbol{z}'), \boldsymbol{\theta} \rangle} \geq \frac{1}{2},$$

and thus (C.31) holds. Finally, by (C.33), we have

$$m \leq 5184s.$$

Therefore (C.26) holds for $d = s + m = 5185s$. □

C.3 Proof of Lemma 8.5

Let $\boldsymbol{e} = \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$, then

$$\begin{aligned}(\boldsymbol{\gamma}^*)^\top \hat{\boldsymbol{\gamma}} - \|\boldsymbol{\gamma}^*\|_2 \cdot \|\hat{\boldsymbol{\gamma}}\|_2 =& (\boldsymbol{\gamma}^*)^\top (\boldsymbol{\gamma}^* + \boldsymbol{e}) - \|\boldsymbol{\gamma}^*\|_2 \cdot \|\boldsymbol{\gamma}^* + \boldsymbol{e}\|_2 \\ =& \|\boldsymbol{\gamma}^*\|_2^2 + (\boldsymbol{\gamma}^*)^\top \boldsymbol{e} - \|\boldsymbol{\gamma}^*\|_2 \sqrt{\|\boldsymbol{\gamma}^*\|_2^2 + 2(\boldsymbol{\gamma}^*)^\top \boldsymbol{e} + \|\boldsymbol{e}\|_2^2} \\ =& \|\boldsymbol{\gamma}^*\|_2^2 + (\boldsymbol{\gamma}^*)^\top \boldsymbol{e} - \|\boldsymbol{\gamma}^*\|_2^2 \sqrt{1 + \frac{2(\boldsymbol{\gamma}^*)^\top \boldsymbol{e} + \|\boldsymbol{e}\|_2^2}{\|\boldsymbol{\gamma}^*\|_2^2}} \\ \sim& \|\boldsymbol{\gamma}^*\|_2^2 + (\boldsymbol{\gamma}^*)^\top \boldsymbol{e} - \|\boldsymbol{\gamma}^*\|_2^2 \left(1 + \frac{1}{2} \frac{2(\boldsymbol{\gamma}^*)^\top \boldsymbol{e} + \|\boldsymbol{e}\|_2^2}{\|\boldsymbol{\gamma}^*\|_2^2}\right) \\ =& \frac{\|\boldsymbol{e}\|_2^2}{2} \lesssim \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2.\end{aligned}$$

□

C.4 Proof of Lemma 3.6

We reiterate Lemma 3.6 as follows.

Lemma C.7. *Let $\hat{\boldsymbol{\theta}}^{(0)}$ be the estimator constructed by the Hardt-Price algorithm. Under the conditions of Theorem 3.1, if $s(\frac{\log p}{n})^{1/12} = o(1)$, then for sufficiently large n, with probability $1 - o(1)$, $\hat{\boldsymbol{\theta}}^{(0)}$ satisfies* **(IC)** *and thus* **(C1)** *holds.*

Proof. Since $s(\frac{\log p}{n})^{1/12} = o(1)$, by Proposition 3.1, we have

$$\max\left(\|\hat{\boldsymbol{\mu}}_1^{(0)} - \boldsymbol{\mu}_{\pi(1)}^*\|_\infty^2, \|\hat{\boldsymbol{\mu}}_2^{(0)} - \boldsymbol{\mu}_{\pi(2)}^*\|_\infty^2, |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty\right) = o_P(\frac{1}{s^2}).$$

Therefore

$$\|\hat{\boldsymbol{\mu}}_k^{(0)} - \boldsymbol{\mu}_{\pi(k)}^*\|_\infty = o_P(\frac{1}{s}) \ (k = 1, 2), \quad |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty \le o_P(\frac{1}{s}),$$

which implies

$$\|\hat{\boldsymbol{\mu}}_k^{(0)} - \boldsymbol{\mu}_{\pi(k)}^*\|_{2,s} \le o_P(\frac{1}{\sqrt{s}}) \ (k = 1, 2), \quad \|(\hat{\Sigma}^{(0)} - \Sigma^*)\boldsymbol{\beta}^*\|_{2,s} \le o_P(\frac{1}{\sqrt{s}}),$$

and therefore $d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) = o(\frac{1}{\sqrt{s}}) \le r\Delta$, that is, there exists r satisfying $r < \frac{|c_0 - c_\omega|}{\Delta} \wedge \frac{\sqrt{(K+2)^2 M + 16c_1} - \sqrt{(K+2)^2 M}}{4} \wedge \sqrt{\frac{c_1}{K^2 M}} \wedge \frac{C_b}{5K\sqrt{s}}$ with $K = 8MC_1$, such that

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \le r\Delta.$$

In addition, since

$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\boldsymbol{\beta}^\top \hat{\Sigma}^{(0)}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1^{(0)} - \hat{\boldsymbol{\mu}}_2^{(0)}) + \lambda_n^{(0)}\|\boldsymbol{\beta}\|_1 \right\},$$

with $\lambda_n^{(0)} = C_d + C_\lambda\sqrt{\log p/n}$. When C_d, C_λ are chosen such that $\lambda_n^{(0)} \asymp 3C_{con}\sqrt{\frac{\log p}{n}} + 2\kappa_0 \cdot \frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)\vee\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\|}{\sqrt{s}}$, by Lemma A.1, we have

$$\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^* \in \Gamma(s),$$

and

$$\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_2 = o_P(\frac{1}{\sqrt{s}}).$$

□

C.5 Proof of auxiliary Lemmas used in the proof of Theorem 3.3

In this section we consider the model $\frac{1}{2}N(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N(\boldsymbol{\mu}_2, \Sigma)$, and for simplicity, we denote $\boldsymbol{\theta}$ by $\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ in the sequel, and provide a lower bound for the excess mis-clustering error. For ease of presentation, let's assume π is an identity map for both $L_{\boldsymbol{\theta}}(\cdot)$ and $R(\cdot)$ in the this section.

C.5.1 Proof of Lemma 3.4

Let $\eta_{\boldsymbol{\theta}}(Z) = \frac{\phi_1(Z)}{\phi_1(Z)+\phi_2(Z)}$, where ϕ_k's are the density function of $N_p(\boldsymbol{\mu}_k, \Sigma_k)$, $k = 1, 2$. It is easy to check that

$$\mathbb{P}_{\boldsymbol{\theta}}(|\eta_{\boldsymbol{\theta}}(Z) - \frac{1}{2}| \le t) \le \frac{t}{m},$$

for some $m \neq 0$ to be determined later, where $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability with respect to the distribution of Z with parameter $\boldsymbol{\theta}$. To see this,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}}(|\eta_{\boldsymbol{\theta}}(Z) - \frac{1}{2}| \le t) &= \mathbb{P}_{\boldsymbol{\theta}}\Big(|\frac{e^{-(Z-\boldsymbol{\mu}_1)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_1)/2}}{e^{-(Z-\boldsymbol{\mu}_1)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_1)/2} + e^{-(Z-\boldsymbol{\mu}_2)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_2)/2}} - \frac{1}{2}| \le t\Big) \\
&= \mathbb{P}_{\boldsymbol{\theta}}\Big(|\frac{e^{-(Z-\boldsymbol{\mu}_1)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_1)/2} - e^{-(Z-\boldsymbol{\mu}_2)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_2)/2}}{e^{-(Z-\boldsymbol{\mu}_1)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_1)/2} + e^{-(Z-\boldsymbol{\mu}_2)^\top\Sigma^{-1}(Z-\boldsymbol{\mu}_2)/2}}| \le 2t\Big) \\
&= \mathbb{P}_{\boldsymbol{\theta}}\Big(|\frac{1 - e^{-(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)^\top\boldsymbol{\beta}}}{1 + e^{-(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)^\top\boldsymbol{\beta}}}| \le 2t\Big) \\
&= \mathbb{P}_{\boldsymbol{\theta}}\Big(\frac{1-2t}{1+2t} \le e^{-(Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)^\top\boldsymbol{\beta}} \le \frac{1+2t}{1-2t}\Big) \\
&\sim \mathbb{P}_{\boldsymbol{\theta}}(-t \le (Z - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^\top\boldsymbol{\beta} \le t).
\end{aligned}$$

Since $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \ge c_L$, the density function of $(Z - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^\top\boldsymbol{\beta}$ is bounded. Then there exists some $m > 0$, such that $\mathbb{P}_{\boldsymbol{\theta}}(-t \leq (Z-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2)^\top \boldsymbol{\beta} \leq t) \leq \frac{t}{m}$.

Let $S_G = \{Z : G(Z) = 1\}$, $S_{G_{\boldsymbol{\theta}}} = \{Z : G_{\boldsymbol{\theta}}(Z) = 1\} = \{Z : \eta_{\boldsymbol{\theta}}(Z) \geq \frac{1}{2}\}$ and $\mathcal{A} = \{Z : |\eta_{\boldsymbol{\theta}}(Z) - \frac{1}{2}| > t\}$. We claim that

$$\mathbb{P}_{\boldsymbol{\theta}}(G \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}} \neq Y) = \int_{S_G \Delta S_{G_{\boldsymbol{\theta}}}} |2\eta_{\boldsymbol{\theta}}(z) - 1| \, \mathbb{P}_{\boldsymbol{\theta}}(dz).$$

To see this,

$$\mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) = \frac{1}{2}\int_{S_G^c} \phi_1 \, dz + \frac{1}{2}\int_{S_G} \phi_2 \, dz,$$

$$\mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y) = \frac{1}{2}\int_{S_{G_{\boldsymbol{\theta}}}^c} \phi_1 \, dz + \frac{1}{2}\int_{S_{G_{\boldsymbol{\theta}}}} \phi_2 \, dz.$$

Since $S_{G_{\boldsymbol{\theta}}} = \{z : \phi_2(z) \geq \phi_1(z)\}$, we then have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y) &= \int_{S_G \Delta S_{G_{\boldsymbol{\theta}}}} |\frac{1}{2}\phi_1 - \frac{1}{2}\phi_2| \, dz \\ &= \int_{S_G \Delta S_{G_{\boldsymbol{\theta}}}} |2\eta(z) - 1| \mathbb{P}_{\boldsymbol{\theta}}(dz),\end{aligned}$$

which implies that

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}(Y \neq \hat{G}(Z)) - \mathbb{P}_{\boldsymbol{\theta}}(Y \neq G_{\boldsymbol{\theta}}(Z)) &\geq 2t\mathbb{P}_{\boldsymbol{\theta}}(S_{\hat{G}} \Delta S_{G_{\boldsymbol{\theta}}} \cap \mathcal{A}) \\ &\geq 2t\big\{\mathbb{P}_{\boldsymbol{\theta}}(S_{\hat{G}} \Delta S_{G_{\boldsymbol{\theta}}}) - \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{A}^c)\big\} \\ &\geq 2t\big\{\mathbb{P}_{\boldsymbol{\theta}}(S_{\hat{G}} \Delta S_{G_{\boldsymbol{\theta}}}) - \frac{t}{m}\big\}.\end{aligned}$$

Maximizing the last expression with respect to t, and using the fact that

$$\mathbb{P}_{\boldsymbol{\theta}}(S_{\hat{G}} \Delta S_{G_{\boldsymbol{\theta}}}) = \mathbb{P}_{\boldsymbol{\theta}}(\hat{G}(Z) \neq G_{\boldsymbol{\theta}}(Z)),$$

we get if $\mathbb{P}_{\boldsymbol{\theta}}(Y \neq \hat{G}(Z)) - \mathbb{P}_{\boldsymbol{\theta}}(Y \neq G_{\boldsymbol{\theta}}(Z)) \leq 1/m$,

$$\frac{1}{2m}\mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq \hat{G}(Z))^2 \leq \mathbb{P}_{\boldsymbol{\theta}}(Y \neq \hat{G}(Z)) - \mathbb{P}_{\boldsymbol{\theta}}(Y \neq G_{\boldsymbol{\theta}}(Z)).$$

C.5.2 Proof of Lemma 8.4

Lemma 3.5, Lemma 8.4 and Lemma 8.5 largely depend on the results in Azizyan et al. (2013), where Lemma 3.5 is a direct result and Lemma 8.4 is an advanced version. For completeness, we provide the full proof of Lemma 8.4 below.

Let $\cos\psi = \frac{|\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|}{\|\boldsymbol{\mu}\|_2 \|\tilde{\boldsymbol{\mu}}\|_2}$. After affine transformation, two lines on $\mathbb{R}^p$ can always lie in the same plane, say, the x–y plane. That is, by some affine transformations, we can transform $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$ to two lines such that the last $p-2$ coordinates of $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\mu}}$ are 0. Since KL divergence is invariant to affine transformations, we have

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) = \iint p_{\boldsymbol{\theta}}(x,y) \log \frac{p_{\boldsymbol{\theta}}(x,y)}{p_{\tilde{\boldsymbol{\theta}}}(x,y)} \, dx \, dy,$$

where

$$\begin{aligned} p_{\boldsymbol{\theta}}(x,y) &= (1-\omega)\phi(x+\xi_x)\phi(y+\xi_y) + \omega\phi(x-\xi_x)\phi(y-\xi_y), \\ p_{\tilde{\boldsymbol{\theta}}}(x,y) &= (1-\omega)\phi(x+\xi_x)\phi(y-\xi_y) + \omega\phi(x-\xi_x)\phi(y+\xi_y), \end{aligned}$$

with $\xi_x = \xi \cos \frac{\psi}{2}, \xi_y = \xi \sin \frac{\psi}{2}, \xi = \frac{\|\boldsymbol{\mu}\|_2}{\sigma}$ and ϕ being the density function of $N(0,1)$. Then

$$\begin{aligned} \frac{p_{\boldsymbol{\theta}}(x,y)}{p_{\tilde{\boldsymbol{\theta}}}(x,y)} =& \frac{(1-\omega)\phi(x+\xi_x)\phi(y+\xi_y) + \omega\phi(x-\xi_x)\phi(y-\xi_y)}{(1-\omega)\phi(x+\xi_x)\phi(y-\xi_y) + \omega\phi(x-\xi_x)\phi(y+\xi_y)} \\ =& \frac{(1-\omega)\exp(-x\xi_x - y\xi_y) + \omega\exp(x\xi_x + y\xi_y)}{(1-\omega)\exp(-x\xi_x + y\xi_y) + \omega\exp(x\xi_x - y\xi_y)} \end{aligned}$$

Further,

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) &= \iint (1-\omega)\phi(x+\xi_x)\phi(y+\xi_y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y} + \omega e^{x\xi_x + y\xi_y}}{(1-\omega)e^{-x\xi_x + y\xi_y} + \omega e^{x\xi_x - y\xi_y}} \, dx \, dy \\
&\quad + \iint \omega\phi(x-\xi_x)\phi(y-\xi_y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y} + \omega e^{x\xi_x + y\xi_y}}{(1-\omega)e^{-x\xi_x + y\xi_y} + \omega e^{x\xi_x - y\xi_y}} \, dx \, dy \\
&= \iint (1-\omega)\phi(-x+\xi_x)\phi(-y+\xi_y) \log \frac{(1-\omega)e^{x\xi_x + y\xi_y} + \omega e^{-x\xi_x - y\xi_y}}{(1-\omega)e^{x\xi_x - y\xi_y} + \omega e^{-x\xi_x + y\xi_y}} \, dx \, dy \\
&\quad + \iint \omega\phi(x-\xi_x)\phi(y-\xi_y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y} + \omega e^{x\xi_x + y\xi_y}}{(1-\omega)e^{-x\xi_x + y\xi_y} + \omega e^{x\xi_x - y\xi_y}} \, dx \, dy \\
&= \iint (1-\omega)\phi(x-\xi_x)\phi(y-\xi_y) \log \frac{(1-\omega)e^{x\xi_x + y\xi_y} + \omega e^{-x\xi_x - y\xi_y}}{(1-\omega)e^{x\xi_x - y\xi_y} + \omega e^{-x\xi_x + y\xi_y}} \, dx \, dy \\
&\quad + \iint \omega\phi(x-\xi_x)\phi(y-\xi_y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y} + \omega e^{x\xi_x + y\xi_y}}{(1-\omega)e^{-x\xi_x + y\xi_y} + \omega e^{x\xi_x - y\xi_y}} \, dx \, dy \\
&= \iint (1-\omega)\phi(x)\phi(y) \log \frac{(1-\omega)e^{x\xi_x + \xi_x^2 + y\xi_y + \xi_y^2} + \omega e^{-x\xi_x - y\xi_y - \xi_x^2 - \xi_y^2}}{(1-\omega)e^{x\xi_x - y\xi_y + \xi_x^2 - \xi_y^2} + \omega e^{-x\xi_x + y\xi_y - \xi_x^2 + \xi_y^2}} \, dx \, dy \\
&\quad + \iint \omega\phi(x)\phi(y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y - \xi_x^2 - \xi_y^2} + \omega e^{x\xi_x + y\xi_y + \xi_x^2 + \xi_y^2}}{(1-\omega)e^{-x\xi_x + y\xi_y - \xi_x^2 + \xi_y^2} + \omega e^{x\xi_x - y\xi_y + \xi_x^2 - \xi_y^2}} \, dx \, dy \\
&= \iint (1-\omega)\phi(x)\phi(y) \log \frac{(1-\omega)e^{x\xi_x + \xi_x^2 + y\xi_y + \xi_y^2} + \omega e^{-x\xi_x - y\xi_y - \xi_x^2 - \xi_y^2}}{(1-\omega)e^{x\xi_x + y\xi_y + \xi_x^2 - \xi_y^2} + \omega e^{-x\xi_x - y\xi_y - \xi_x^2 + \xi_y^2}} \, dx \, dy \\
&\quad + \iint \omega\phi(x)\phi(y) \log \frac{(1-\omega)e^{-x\xi_x - y\xi_y - \xi_x^2 - \xi_y^2} + \omega e^{x\xi_x + y\xi_y + \xi_x^2 + \xi_y^2}}{(1-\omega)e^{-x\xi_x - y\xi_y - \xi_x^2 + \xi_y^2} + \omega e^{x\xi_x + y\xi_y + \xi_x^2 - \xi_y^2}} \, dx \, dy \\
&= \int (1-\omega)\phi(z) \log \frac{(1-\omega)e^{z\sqrt{\xi_x^2+\xi_y^2} + \xi_x^2 + \xi_y^2} + \omega e^{-z\sqrt{\xi_x^2+\xi_y^2} - \xi_x^2 - \xi_y^2}}{(1-\omega)e^{z\sqrt{\xi_x^2+\xi_y^2} + \xi_x^2 - \xi_y^2} + \omega e^{-z\sqrt{\xi_x^2+\xi_y^2} - \xi_x^2 + \xi_y^2}} \, dz \\
&\quad + \int \omega\phi(z) \log \frac{(1-\omega)e^{-z\sqrt{\xi_x^2+\xi_y^2} - \xi_x^2 - \xi_y^2} + \omega e^{z\sqrt{\xi_x^2+\xi_y^2} + \xi_x^2 + \xi_y^2}}{(1-\omega)e^{-z\sqrt{\xi_x^2+\xi_y^2} - \xi_x^2 + \xi_y^2} + \omega e^{z\sqrt{\xi_x^2+\xi_y^2} + \xi_x^2 - \xi_y^2}} \, dz.
\end{aligned}$$

Let $h_1(x) = (1-\omega)\exp(-x) + \omega \exp(x)$. Then $h_1'(x) = -(1-\omega)\exp(-x) + \omega\exp(x)$ and $h_1''(x) = (1-\omega)\exp(-x) + \omega\exp(x)$. Let $g_1(x) = \log h_1(x)$. Then

$$g_1'' = \frac{h_1'' h_1 - (h_1')^2}{h_1^2} = \frac{h_1^2 - (h_1')^2}{h_1^2} \in (0, 1).$$

Similarly, let $h_2(x) = (1-\omega)\exp(x) + \omega\exp(-x)$ and $g_2(x) = \log h_2(x)$. We also have

$g_2'' \in (0,1)$. By the mean-value theorem,

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) \le & 2\xi_y^2 \int \phi(z)\big((1-\omega)g_1'(\xi z + \xi^2) + \omega g_2'(\xi z + \xi^2)\big)\, dz \\
= & 2\xi_y^2 \int \phi(z)\big((1-\omega)g_1'(\xi z + \xi^2)\big)\, dz + 2\xi_y^2 \int \phi(z)\big(\omega g_2'(\xi z + \xi^2)\big)\, dz \\
= & 2\xi_y^2 \int (1-\omega)\phi(z)\big(g_1'(\xi z + \xi^2) - g_1'(\xi z - \frac{\log \tau}{2})\big)\, dz + \\
& 2\xi_y^2 \int \omega\phi(z)\big(g_2'(\xi z + \xi^2) - g_2'(\xi z - \frac{\log \tau}{2})\big)\, dz \\
\le & 2\xi_y^2 (\xi^2 + \frac{\log \tau}{2}) \\
= & 2\xi^2 (\xi^2 + \frac{\log \tau}{2}) \sin^2 \frac{\psi}{2} \\
= & \xi^2 (\xi^2 + \frac{\log \tau}{2})(1 - \cos \psi) \\
= & (\|\boldsymbol{\mu}\|_2^2 + \frac{\log \tau}{2})(\|\boldsymbol{\mu}\|_2^2 - |\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|).
\end{aligned}
$$

D Additional Simulations

To evaluate how CHIME performs in the case of unequal mixing proportion between the two components, we used the same set of inverse covariance matrices generated in the simulation examples of the main paper, but change the mixing proportion such that $\omega^* = 0.3$. The sparsity level of the discriminant vector s remains to be the same, i.e. $s = 10$. The estimated mis-clustering errors based on 200 test samples from 100 replications are presented in Table A1. In general, the errors from Oracle and the supervised LPD rule decrease when the mixing proportion varies from 0.5 to 0.3, due to better calibration of the sample means and weighted sample covariance matrix. However, the unequal mixing proportion has made it more challenging for clustering, as seen from the increased mis-clustering errors for a majority of the settings where an unsupervised clustering method is used. For CHIME, although the settings where $p = 100, 200$ in Model 1 and $p = 500, 800$ in Model 3 show deteriorated performances due to poor initializations, the corresponding errors are still smaller than those obtained with other clustering methods. More importantly, when good initializations are available, including the settings where $p = 500, 800$ in Model 1, $p = 200, 500, 800$ in Model 2, and $p = 200$ in Model 3, CHIME achieves comparable performances to the supervised LPD rule.

In the last set of simulation examples, we consider modifying the discriminant vector such that the first $s = 20$ entries in $\boldsymbol{\beta}^*$ are nonzero and keep the mixing proportion $\omega^* = 0.5$. Again we used the three models described in Section 5 to generate the inverse covariance matrices. To ensure sufficiently strong signal, $\boldsymbol{\beta}^* = (1, \dots, 1, 0, \dots, 0)^\top$ in Model 1 and 2, and $\boldsymbol{\beta}^* = 3 \cdot (1, \dots, 1, 0, \dots, 0)^\top$ in Model 3. Given Ω^* and $\boldsymbol{\beta}^*$, the mean vectors are

$\boldsymbol{\mu}_1^* = (0, \ldots, 0)^\top$, and $\boldsymbol{\mu}_2^* = \boldsymbol{\mu}_1^* - (\Omega^*)^{-1}\boldsymbol{\beta}^*$. The average mis-clustering errors based on 200 test samples from 100 replications are shown in Table A2.

Overall, we observe a similar pattern as in previous comparisons, where CHIME consistently outperforms other clustering methods and is competitive at several scenarios even when compared with the supervised LPD rule (such as for all p in Model 1, $p = 500$ in Model 2 and $p = 100, 200$ in Model 3). As the inverse covariance matrix in Model 1 is generated from a random graph, several clustering methods (PCCM, SHP and SKM) return poor performances when p varies. While SHP and SKM both behave poorly in Model 2, they outperform KM and PCCM in Model 2, due to stronger signals in the discriminant vector $\boldsymbol{\beta}^*$. Remember due to the special structure of the precision matrix in Model 3, the mean vector $\boldsymbol{\mu}_2^*$ is exactly sparse with $s + 1$ nonzero entries. It is thus not surprising to see that SKM performs well in separating the two components.

Further, clustering methods such as CHIME and PCCM both require iterative algorithms. The running times of different clustering methods are not directly comparable as they were run on different platforms. Nonetheless we can compare their per-iteration computational complexity. At each iteration, the main difference between CHIME and PCCM is that CHIME uses a linear program to solve for $\boldsymbol{\beta}^* \in \mathbb{R}^p$, whereas PCCM uses penalized maximum likelihood to solve for the inverse covariance matrix $\Omega^* \in \mathbb{R}^{p \times p}$. While the per-iteration complexity for CHIME is $O(p^2)$ including cost for computing the empirical covariance matrix, that of PCCM is $O(p^3)$. This also explains why PCCM runs very slowly in practice. In addition, CHIME, PCCM and SHP all require selecting the tuning parameters. Although the CHIME algorithm asks for two constants C_1 and C_2, the main focus should be on C_2 as the performance of CHIME is not sensitive to the choice of C_1. In contrast, PCCM requires choosing two parameters, one for controlling the sparsity of the means and the other for controlling the sparsity of the inverse covariance matrices. Our experience shows that tuning with PCCM generally needs to down over a two-dimensional grid, which can be computationally challenging for large p. The main solver in SHP is the LPD rule for solving the discriminant vector $\boldsymbol{\beta}^*$. Therefore one can choose the tuning parameter in SHP following the practice used in LPD.

Last but not least, we compare CHIME and IF-PCA in a simple model where the covariance matrix is diagonal. Specifically, the mean vector $\boldsymbol{\mu}_1^*$ is set to $\boldsymbol{\mu}_1^* = \mathbf{0}$, and $\boldsymbol{\beta}^* = (1, \ldots, 1, 0, \ldots, 0)^\top$ where the first $s = 10$ entries are 1. Σ is generated as a diagonal matrix with diagonal elements equal to absolute values of *i.i.d.* normal random variables with mean 2 and variance 1. The sample size is $n = 200$ and the probability of being in either of the two classes is set to be equal, i.e. $\omega^* = 0.5$. The simulation results are summarized in Table A3. CHIME still significantly outperforms IF-PCA in this case, although CHIME does not use the fact that the covariance matrix is diagonal, while IF-PCA does. The main reason is that IF-PCA is specifically designed for "weak and rare" signal where the strength of signal is assumed to be vanishing.

Table A3: Simulation results for CHIME and IF-PCA when the covariance matrix is diagonal

p	100	200	500	800
IF-PCA	64.70(21.90)	75.80(19.22)	96.00(2.62)	92.40(7.73)
CHIME	11.31(3.61)	11.00(9.75)	10.66(3.98)	9.73(3.17)
LPD	5.44(2.46)	5.95(2.52)	7.02(2.61)	7.62(2.70)
Oracle	3.96(2.25)	4.03(2.11)	3.78(1.92)	4.2(1.82)

References

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *NIPS*, pages 2139–2147, 2013.

Vitali D Milman and Gideon Schechtman. Asymptotic theory of finite dimensional normed spaces. 1986.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *NIPS*, pages 2521–2529, 2015.

Table A1: Average mis-clustering error (s.e.) based on $n = 200$ test samples from 100 replications under three different models: $\omega^* = 0.3$ and $s = 10$

	p	100	200	500	800
Model 1	KM	50.04(7.77)	66.83(9.87)	30.87(8.73)	21.84(6.21)
	PCCM	50.28(6.22)	63.24(8.38)	30.25(7.68)	22.99(4.89)
	SHP	59.22(13.62)	66.29(10.36)	61.43(12.34)	53.99(16.34)
	SKM	51.54(8.18)	71.49(10.07)	26.60(9.24)	20.77(5.90)
	IF-PCA	92.24(6.10)	92.24(5.60)	92.59(6.02)	92.86(5.65)
	CHIME	14.73(9.81)	27.21(13.21)	4.52(2.80)	3.91(1.72)
	LPD	6.07(2.63)	5.36(1.92)	3.13(1.72)	2.56(1.51)
	Oracle	5.42(2.48)	4.47(1.74)	2.38(1.43)	1.61(1.29)
Model 2	KM	43.26(7.37)	21.62(6.15)	18.56(6.05)	0.56(0.81)
	PCCM	46.13(6.50)	23.57(5.82)	19.43(4.64)	0.57(0.74)
	SHP	53.83(16.40)	30.75(19.47)	23.73(16.88)	13.02(7.97)
	SKM	45.38(8.13)	25.74(6.66)	18.51(5.90)	0.65(0.86)
	IF-PCA	74.69(18.60)	84.65(14.88)	90.20(7.73)	91.32(7.66)
	CHIME	7.89(4.32)	2.42(1.43)	1.15(1.09)	0.02(0.11)
	LPD	4.16(2.00)	1.75(1.34)	0.98(1.02)	0.01(0.10)
	Oracle	3.86(2.00)	1.31(1.23)	0.79(0.86)	0.01(0.10)
Model 3	KM	38.94(24.04)	61.36(23.11)	79.87(11.72)	81.19(12.32)
	PCCM	46.15(25.93)	68.99(13.43)	76.46(5.13)	76.06(5.55)
	SHP	29.62(16.79)	40.00(22.44)	47.44(20.70)	50.03(76.06)
	SKM	14.15(4.31)	13.94(3.92)	18.83(19.75)	28.10(30.48)
	IF-PCA	86.91(12.14)	88.96(9.89)	91.21(7.10)	92.12(6.48)
	CHIME	7.52(3.67)	10.02(3.42)	18.21(14.21)	22.71(23.29)
	LPD	4.78(2.14)	7.03(2.53)	8.18(2.85)	8.94(2.93)
	Oracle	1.45(1.26)	1.31(1.03)	1.47(1.24)	1.46(1.12)

Table A2: Average mis-clustering error (s.e.) based on $n = 200$ test samples from 100 replications under three different models: $\omega^* = 0.5$ and $s = 20$

	p	100	200	500	800
Model 1	KM	7.86(5.14)	21.17(11.59)	12.99(13.70)	6.15(2.60)
	PCCM	4.51(1.64)	13.09(5.91)	7.38(3.27)	5.54(2.41)
	SHP	17.19(15.56)	47.90(19.96)	54.64(20.35)	13.18(4.67)
	SKM	5.88(2.67)	23.53(14.19)	7.38(3.38)	7.07(2.78)
	IF-PCA	90.15(7.41)	93.50(4.93)	94.83(4.73)	94.14(4.03)
	CHIME	0.57(0.82)	0.81(0.84)	0.59(0.94)	0.84(0.98)
	LPD	0.51(0.70)	0.54(0.76)	0.47(0.67)	0.56(0.66)
	Oracle	0.39(0.58)	0.33(0.57)	0.24(0.45)	0.13(0.34)
Model 2	KM	40.50(5.37)	40.06(5.90)	38.74(6.18)	47.67(6.54)
	PCCM	40.93(5.25)	39.72(5.85)	38.47(6.11)	47.28(5.98)
	SHP	47.08(15.35)	48.82(19.04)	40.07(17.05)	55.88(16.94)
	SKM	39.73(5.22)	38.71(5.82)	37.24(6.26)	47.65(6.81)
	IF-PCA	48.04(27.38)	58.87(28.50)	76.92(21.35)	85.35(14.48)
	CHIME	11.92(8.04)	13.26(10.23)	0.72(0.93)	2.31(8.30)
	LPD	2.96(1.83)	1.96(1.51)	0.24(0.47)	0.27(0.62)
	Oracle	1.61(1.47)	0.85(0.95)	0.14(0.38)	0.14(0.40)
Model 3	KM	7.38(2.71)	24.60(31.09)	79.87(20.54)	88.36(11.44)
	PCCM	26.65(34.34)	62.28(33.94)	84.00(5.20)	85.39(3.50)
	SHP	6.31(4.31)	5.97(3.39)	6.80(3.88)	7.41(5.31)
	SKM	6.62(2.58)	6.38(2.29)	6.25(2.28)	6.23(2.44)
	IF-PCA	74.47(18.96)	83.47(15.72)	88.72(10.89)	89.81(9.51)
	CHIME	2.52(1.90)	4.32(2.45)	6.12(3.12)	7.02(3.19)
	LPD	1.13(1.02)	3.25(1.72)	3.93(2.01)	3.98(2.02)
	Oracle	0.07(0.26)	0.10(0.30)	0.10(0.30)	0.11(0.37)