Appendix A Proofs for Population EM

Throughout the proof, we will use C, c, c', c_{any} without explicit mention whenever we need universal constants to bound any terms.

A.1 Key Lemmas for Population EM Analysis

Before getting into detailed proofs, we state some essential lemmas modified from Yi et al. (2016); Balakrishnan et al. (2017).

Lemma A.1 *Let $X \sim \mathcal{N}(0, I_d)$. For any fixed vector $v \in \mathbb{R}^d$, and a set of vectors $u_1, ..., u_k \in \mathbb{R}^d$ such that $\|u_j\| \geq \|v\|$ for all j, we define*

$$\mathcal{E} := \{|\langle X, u_j \rangle| \geq |\langle X, v \rangle|, \ \forall j = 1, ..., k\}.$$

Then,

$$P(\mathcal{E}^c) \leq \sum_{j=1}^{k} \frac{\|v\|}{\|u_j\|}. \tag{4}$$

Furthermore, for any unit vector $s \in \mathbb{S}^{d-1}$ and for any $p \geq 1$, we have

$$\mathbb{E}[|\langle X, s \rangle|^p | \mathcal{E}^c] \leq k 2^p \Gamma(1 + p/2), \tag{5}$$

where Γ is a gamma function.

Lemma A.2 *Let $X \sim \mathcal{N}(0, I_d)$. For any set of fixed vectors $u_1, ..., u_k \in \mathbb{R}^d$, and fixed constants $\alpha_1, ..., \alpha_k > 0$, define*

$$\mathcal{E} := \{|\langle X, u_j \rangle| \geq \alpha_j, \ \forall j = 1, ..., k\}.$$

Then,

$$P(\mathcal{E}^c) \leq \sum_{j=1}^{k} \frac{\alpha_j}{\|u_j\|}. \tag{6}$$

Furthermore, for any unit vector $s \in \mathbb{S}^{d-1}$ and for $p \geq 1$, we have

$$\mathbb{E}[|\langle X, s \rangle|^p | \mathcal{E}^c] \leq k 2^p \Gamma((1 + p)/2)/\sqrt{\pi}. \tag{7}$$

Proofs of these lemmas can be found in Appendix C. As a consequence of Lemma A.1, A.2, we can show the Lemma 4.2.

Lemma 4.2 *Let $X \sim \mathcal{N}(0, I_d)$. Suppose any fixed vector $v \in \mathbb{R}^d$, a set of vectors $u_1, ..., u_k \in \mathbb{R}^d$ such that $\|u_j\| \geq \|v\|$ for all j, and constants $\alpha_1, ..., \alpha_k > 0$. Then consider two events*

$$\mathcal{E} := \{|\langle X, u_j \rangle| \geq |\langle X, v \rangle|, \ \forall j = 1, ..., k\},$$
$$\mathcal{E}' := \{|\langle X, u_j \rangle| \geq \alpha_j, \ \forall j = 1, ..., k\}.$$

Then for any fixed unit vector $s \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[|\langle X, s \rangle|^2 | \mathcal{E}^c], \ \mathbb{E}[|\langle X, s \rangle|^2 | \mathcal{E}'^c] \leq C \log k, \tag{2}$$

for some universal constant $C > 0$.

Proof. We show for Lemma A.1 first. By Holder's inequality,

$$\mathbb{E}[|\langle X, s\rangle|^2|\mathcal{E}^c] \le \mathbb{E}[|\langle X, s\rangle|^{2p}|\mathcal{E}^c]^{1/p}\mathbb{E}[1|\mathcal{E}^c]^{1/q},$$

for any $p, q \ge 1$ such that $1/p + 1/q = 1$. We can take p as arbitrary as we want, say $p = \log k$, in order to get rid of k factor in equation (5). Then,

$$\begin{aligned}\mathbb{E}[|\langle X, s\rangle|^2|\mathcal{E}^c] &\le \mathbb{E}[|\langle X, s\rangle|^{2p}|\mathcal{E}^c]^{1/p}\mathbb{E}[1|\mathcal{E}^c]^{1/q} \le k^{1/p}(4^p\Gamma(1+p))^{1/p} \\ &\le 4e(\Gamma(1+p)^{1/2p})^2 \le C\log k,\end{aligned}$$

for some universal constant $C > 0$. We used the fact that $\Gamma(1+p) \le (p+1)^p$. The proof of Lemma A.2 can be written similarly. □

Remark 6 *These lemmas are modified from Balakrishnan et al. (2017); Yi et al. (2016) to involve multiple components and higher order moments. They are also used in proofs of finite-sample EM, to find sub-exponential norm Vershynin (2010) of random variables conditioned on specific events. Note that boundedness of any p^{th} moment by Gamma function implies sub-Gaussianity. We conjecture that k factor in (5) and (7) might be sub-optimal, and it will improve the SNR condition by $O(\log k)$ if resolved.*

A.2 Bounding B

Since $\|B\| = \sup_{s\in\mathbb{S}^{d-1}} \mathbb{E}_D[w_1\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]$, for any fixed unit vector s, we bound

$$\begin{aligned}B_s &:= |\mathbb{E}_D[w_1\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]| \\ &= |\mathbb{E}_D[w_1\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)] - \mathbb{E}_D[w_1^*\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]| \\ &= |\mathbb{E}_D[\Delta_w\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]| \\ &\le \pi_1^* \underbrace{|\mathbb{E}_{\mathcal{D}_1}[\Delta_w\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]|}_{B_1} + \sum_{j\ne 1}\pi_j^* \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]|}_{B_j}.\end{aligned}$$

We will then bound B_1 and B_j separately, as B_1 is the error term from its own component and B_j is the error from other components.

Term in B_j can be decoupled as

$$\begin{aligned}B_j &= |\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle] + \mathbb{E}_{\mathcal{D}_j}[\Delta w\langle X, s\rangle e]| \\ &\le \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle]|}_{b_1} + \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta w\langle X, s\rangle e]|}_{b_2}.\end{aligned}$$

Then for each $j = 1, ..., k$, we give a bound for B_j. We divide the cases between $\max_j \|\Delta_j\| > 1$ and $\max_j \|\Delta_j\| \le 1$. The proof for $\|\Delta_j\| \le 1$ will be given in Appendix D. We use D_m to denote $\max_j \|\Delta_j\|$ to simplify the notations. We also define $\rho_{jl} := \pi_l^*/\pi_j^*$ for $j \ne l$.

Case I. $\max_j \|\Delta_j\| > 1$**:**

$j \ne 1$: To bound first term, define four events as follows:

$$\begin{aligned}\mathcal{E}_1 &= \{|\langle X, \beta_j^* - \beta_1^*\rangle| \ge 4\sqrt{2}\tau_j\} \\ \mathcal{E}_2 &= \{4(|\langle X, \Delta_j\rangle| \vee |\langle X, \Delta_1\rangle|) \le |\langle X, \beta_j^* - \beta_1^*\rangle|\} \\ \mathcal{E}_3 &= \{|e| \le \tau_j\} \\ \mathcal{E} &= \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3.\end{aligned}$$

When all four events happen at the same time, it is a good sample: weights given to this sample is almost 0, as it comes from component j. For other events, we bound the probability of each event with respect to Δ_j and τ_j. We decide threshold parameter τ_j at the end of the stage.

$$\begin{aligned}b_1 \le\ & |\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_1^c\cap\mathcal{E}_2}]| \\ &+ |\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_2^c}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_3^c}]|.\end{aligned}$$

1. Event $\mathcal{E}$: Observe the value of the weight w_1. First note that

$$(\langle X, \beta_j^* - \beta_j\rangle + e)^2 \le 2|\langle X, \Delta_j\rangle|^2 + 2e^2 \le |\langle X, \beta_j^* - \beta_1^*\rangle|^2/8 + 2e^2$$
$$(\langle X, \beta_j^* - \beta_1\rangle + e)^2 \ge |\langle X, \beta_j^* - \beta_1^*\rangle - \langle X, \Delta_1\rangle|^2/2 - e^2 \ge (9/32)|\langle X, \beta_j^* - \beta_1^*\rangle|^2 - e^2.$$

   Then,

$$\begin{aligned} w_1 &\le \frac{\pi_1 \exp(-(Y - \langle X, \beta_1\rangle)^2/2)}{\pi_1 \exp(-(Y - \langle X, \beta_1\rangle)^2/2) + \pi_j \exp(-(Y - \langle X, \beta_j\rangle)^2/2)} \\ &= \frac{\pi_1 \exp(-(\langle X, \beta_j^* - \beta_1\rangle + e)^2/2)}{\pi_1 \exp(-(\langle X, \beta_j^* - \beta_1\rangle + e)^2/2) + \pi_j \exp(-(\langle X, \beta_j^* - \beta_j\rangle + e)^2/2)} \\ &\le (\pi_1/\pi_j) \exp\left(((\langle X, \beta_j^* - \beta_j\rangle + e)^2 - (\langle X, \beta_j^* - \beta_1\rangle + e)^2)/2\right) \\ &\le (\pi_1/\pi_j) \exp\left((-5|\langle X, \beta_j^* - \beta_1^*\rangle|^2/32 + 3e^2)/2\right) \\ &\le (\pi_1/\pi_j) \exp(-\tau_j^2). \end{aligned} \tag{8}$$

   Similarly, we get

$$\begin{aligned} w_1^* &\le (\pi_1^*/\pi_j^*) \exp\left((e^2 - (\langle X, \beta_j^* - \beta_1^*\rangle + e)^2)/2\right) \\ &\le (\pi_1^*/\pi_j^*) \exp\left(\left(e^2 - (|\langle X, \beta_j^* - \beta_1^*\rangle| - |e|)^2\right)/2\right) \\ &\le (\pi_1^*/\pi_j^*) \exp\left((\tau_j^2 - 16\tau_j^2)/2\right) \\ &\le (\pi_1^*/\pi_j^*) \exp\left(-\tau_j^2\right). \end{aligned}$$

   Note that due to our initialization condition for π_j for all j, $\rho_{j1} = \pi_1^*/\pi_j^* \le 3\pi_1/\pi_j$. Thus, $|\Delta_w| \le 3\rho_{j1} \exp\left(-\tau_j^2\right)$. From this inequality, we can get

$$\begin{aligned} |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}}]| &\le 3\rho_{j1} \exp\left(-\tau_j^2\right) \mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle|] \\ &\le 3\rho_{j1} \exp\left(-\tau_j^2\right) R_{j1}^*, \end{aligned}$$

   where the last inequality comes from Cauchy-Schwartz inequality.

2. Event $\mathcal{E}_1^c \cap \mathcal{E}_2$: In this case, from Lemma A.2,

$$P(\mathcal{E}_1^c \cap \mathcal{E}_2) \le P(\mathcal{E}_1^c) \le \frac{4\sqrt{2}\tau_j}{\|\beta_j^* - \beta_1^*\|}.$$

   Then, we proceed as

$$\begin{aligned} |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_1^c \cap \mathcal{E}_2}]| &\le 4\sqrt{2}\tau_j \mathbb{E}_{\mathcal{D}_j}[|\Delta_w \langle X, s\rangle \mathbb{1}_{\mathcal{E}_1^c \cap \mathcal{E}_2}|] \\ &\le 4\sqrt{2}\tau_j \mathbb{E}_{\mathcal{D}_j}[|\Delta_w \langle X, s\rangle \mathbb{1}_{\mathcal{E}_1^c}|] \\ &\le 4\sqrt{2}\tau_j \sqrt{\mathbb{E}[\Delta_w^2 | \mathcal{E}_1^c]} \sqrt{\mathbb{E}[\langle X, s\rangle^2 | \mathcal{E}_1^c]} P(\mathcal{E}_1^c) \\ &\le 4\sqrt{2}\tau_j P(\mathcal{E}_1^c) \le \frac{32\tau_j^2}{R_{j1}^*}. \end{aligned}$$

3. Event $\mathcal{E}_2^c$: Bound it as follows:

$$|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_2^c}]| \le \sqrt{\mathbb{E}[\Delta_w^2 \langle X, s\rangle^2 | \mathcal{E}_2^c]} \sqrt{\mathbb{E}[\langle X, \beta_j^* - \beta_1^*\rangle^2 | \mathcal{E}_2^c]} P(\mathcal{E}_2^c).$$

   Under this event, we note that

$$\langle X, \beta_j^* - \beta_1^*\rangle \le 4(|\langle X, \Delta_j\rangle| \vee |\langle X, \Delta_1\rangle|) \le 4(|\langle X, \Delta_j\rangle| + |\langle X, \Delta_1\rangle|).$$

$$\begin{aligned}\mathbb{E}[\langle X, \beta_j^* - \beta_1\rangle^2 | \mathcal{E}_2^c] &\le \mathbb{E}[32|\langle X, \Delta_j\rangle|^2 + 32|\langle X, \Delta_1\rangle|^2 | \mathcal{E}_2^c] \\ &\le 32(\mathbb{E}[|\langle X, \Delta_j\rangle|^2 | \mathcal{E}_2^c] + \mathbb{E}[|\langle X, \Delta_1\rangle|^2 | \mathcal{E}_2^c]) \\ &\le 512 D_m^2,\end{aligned}$$

where we used Lemma A.1 for bounding $\mathbb{E}[\langle X, \Delta_j\rangle^2 | \mathcal{E}_2^c]$.

Now plugging this into the above,

$$\begin{aligned}&\sqrt{\mathbb{E}[\langle X, s\rangle^2 | \mathcal{E}_2^c]}\sqrt{\mathbb{E}[\langle X, \beta_j^* - \beta_1^*\rangle^2 | \mathcal{E}_2^c]} P(\mathcal{E}_2^c) \\ &\qquad \le 64 D_m P(\mathcal{E}_2^c) \le 512 D_m \frac{D_m}{R_{j1}^*}.\end{aligned}$$

4. Event $\mathcal{E}_3^c$: Similarly,

$$\begin{aligned}|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_3^c}]| &\le \sqrt{\mathbb{E}[\Delta_w^2 \langle X, s\rangle^2 | \mathcal{E}_3^c]}\sqrt{\mathbb{E}[\langle X, \beta_j^* - \beta_1^*\rangle^2 | \mathcal{E}_3^c]} P(\mathcal{E}_3^c) \\ &\le \|\beta_j^* - \beta_1^*\| P(\mathcal{E}_3^c) \\ &\le 2R_{j1}^* \exp(-\tau_j^2/2).\end{aligned}$$

We used independence of e and X. Combining all,

$$b_1 \le O(\exp(-\tau_j^2/2)(1 \vee \rho_{j1}) R_{j1}^* + \tau_j^2/R_{j1}^* + D_m/R_{j1}^*) D_m. \tag{9}$$

Now we turn our attention to b_2. Recall $b_2 = |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e]|$. For this setup,

$$\begin{aligned}b_2 \le\ & |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1^c}]| \\ &+ |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_{2^c}}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_3^c}]|.\end{aligned}$$

Under good event $\mathcal{E}$, as previously we have $|\Delta_w| \le 3\rho_{j1}\exp(-\tau_j^2)$, thus

$$|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}}]| \le 3\rho_{j1}\exp(-\tau_j^2) \mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle e|] \le 3\rho_{j1}\exp(-\tau_j^2).$$

Similarly, we go through on the bad events. First,

$$|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1^c}]| \le \sqrt{\mathbb{E}_{\mathcal{D}_j}[\langle X, s\rangle^2 | \mathcal{E}_1^c]}\sqrt{\mathbb{E}_{\mathcal{D}_j}[e^2 | \mathcal{E}_1^c]} P(\mathcal{E}_1^c) \le c_1 \tau_j / R_{j1}^*,$$

where we used Lemma A.2 for bounding $\mathbb{E}_{\mathcal{D}_j}[\langle X, s\rangle^2 | \mathcal{E}_1^c]$.

Second,

$$|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_2^c}]| \le \sqrt{\mathbb{E}_{\mathcal{D}_j}[\langle X, s\rangle^2 | \mathcal{E}_2^c]}\sqrt{\mathbb{E}_{\mathcal{D}_j}[e^2 | \mathcal{E}_2^c]} P(\mathcal{E}_2^c) \le c_2 D_m / R_{j1}^*.$$

where we used Lemma A.1 for bounding $\mathbb{E}_{\mathcal{D}_j}[\langle X, s\rangle^2 | \mathcal{E}_2^c]$.

Finally,

$$|\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_3^c}]| \le \sqrt{\mathbb{E}_{\mathcal{D}_j}[\langle X, s\rangle^2 e^2]}\sqrt{P(\mathcal{E}_3^c)} \le c_3 \exp(-\tau_j^2/4).$$

Combining three items, we have

$$b_2 \le O((1 \vee \rho_{j1})\exp(-\tau_j^2/4) + \tau_j/R_{j1}^* + D_m/R_{j1}^*). \tag{10}$$

Now we set

$$\tau_j = c_\tau \sqrt{\log(R_{j1}^* k/(1 \wedge \rho_{j1}))},\ R_{j1}^* > c_r k \rho_{j1}^{-1} \log(R_{j1}^*).$$

With given good initialization $D_m/R_{j1}^* \le c_D \rho_{j1}/k$, we get $b_1 < c_b D_m \rho_{j1}/k$ and $b_2 \le c_{b'} D_m \rho_{j1}/k$ since $D_m \ge 1$. Combining (9) and (10), we get $B_j \le c_B D_m \rho_{j1}/k$ for some small universal constant $c_B < 1/4$ with large enough c_τ, c_r and small enough c_D.

$j = 1$: We only need to consider bounding $b_2 = |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e]|$. We define some events similarly, but each involves multiple factors in this case.

$$\mathcal{E}_1 = \{|\langle X, \beta_1^* - \beta_j\rangle| \geq 4\tau, \ \forall j \neq 1\}$$
$$\mathcal{E}_2 = \{4|\langle X, \Delta_1\rangle| \leq |\langle X, \beta_1^* - \beta_j\rangle|, \ \forall j \neq 1\}$$
$$\mathcal{E}_3 = \{|e| \leq \tau\},$$
$$\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3.$$

Then follow the same path as in cases $j \neq 1$,

$$\begin{aligned} b_2 \leq & |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1^c}]| \\ & + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_2 c}]| + |\mathbb{E}_{\mathcal{D}_j}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_3^c}]|. \end{aligned}$$

Then, on event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, for all $j \neq 1$, we have

$$w_j \leq (\pi_1/\pi_j) \exp\left((-(\langle X, \beta_1^* - \beta_j\rangle + e)^2 + (\langle X, \Delta_1\rangle + e)^2))/2\right) \leq 3\rho_{j1} \exp(-3\tau^2/2),$$

as before. Thus, $w_1 \geq 1 - 3k\rho_\pi \exp(-3\tau^2/2)$. Similarly, $w_1^* \geq 1 - 3k\rho_\pi \exp(-3\tau^2/2)$. Thus, Δ_w can be at most $k3\rho_\pi \exp(-3\tau^2/2)$. Then,

$$|\mathbb{E}_{\mathcal{D}_1}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}}]| \leq 3k\rho_\pi \exp(-3\tau^2/2)\mathbb{E}_{\mathcal{D}_1}[|\langle X, s\rangle e|] \leq 3k\rho_\pi \exp(-3\tau^2/2).$$

We can go over other events similarly.

$$|\mathbb{E}_{\mathcal{D}_1}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1^c}]| \leq \sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X, s\rangle^2 | \mathcal{E}_1^c]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[e^2 | \mathcal{E}_1^c]} P(\mathcal{E}_1^c) \leq c_1 \sqrt{\log k} \frac{k\tau}{R_{min}}.$$

$$|\mathbb{E}_{\mathcal{D}_1}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_2^c}]| \leq \sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X, s\rangle^2 | \mathcal{E}_2^c]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[e^2 | \mathcal{E}_2^c]} P(\mathcal{E}_2^c) \leq c_2 \sqrt{\log k} \frac{kD_m}{R_{min}}.$$

$$|\mathbb{E}_{\mathcal{D}_1}[\Delta_w \langle X, s\rangle e \mathbb{1}_{\mathcal{E}_3^c}]| \leq \sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X, s\rangle^2 e^2]}\sqrt{P(\mathcal{E}_3^c)} \leq c_3 \exp(-\tau^2/4).$$

For first two inequalites, we used Lemma A.1 and A.2. They all gives a bound for b_2 as,

$$b_2 \leq O(k\rho_\pi \exp(-\tau^2/4) + (k\sqrt{\log k})\tau/R_{min} + (k\sqrt{\log k})D_m/R_{min}). \tag{11}$$

Now we set $\tau = \Theta(\sqrt{\log(k\rho_\pi)})$, $R_{min} = \Omega(k \log(k\rho_\pi))$ and $D_m = O(R_{min}/(k\sqrt{\log k}))$, and we get $b_2 \leq c_B$ and $B_1 = b_2 \leq c_B D_m$.

Combining (9), (10), and (11), we get the first part of Lemma 4.1. We conclude

$$B = \pi_1^* B_1 + \sum_j \pi_j^* B_j \leq \pi_1^* c_B D_m + \sum_{j\neq 1} \pi_j^* c_B D_m \rho_{j1}/k = \pi_1^* \left(c_B + \sum_{j \neq 1} c_B/k\right) D_m \leq 2\pi_1^* c_B D_m,$$

where we used $\pi_j^* \rho_{j1} = \pi_1^*$. Thus $B \leq c_B' \pi_1^* D_m$ for some universal constant $c_B' \in (0, 1/4)$ with properly set constants in the proof.

Update for mixing weights. In this case $D_m \geq 1$, we will not focusing on improvement over the quality of π_j. Instead, we will only show that π_j stays in a neighborhood of the true parameter, *i.e.*, $|\pi_j - \pi_j^*| \leq \pi_j^*/2$. It can be actually very easily shown with reusing the results we derived for β. Observe that

$$\pi_1^+ - \pi_1^* = \mathbb{E}_{\mathcal{D}}[w_1 - w_1^*] = \mathbb{E}_{\mathcal{D}}[\Delta_w]. \tag{12}$$

Now we can proceed as before:

$$\mathbb{E}_{\mathcal{D}}[\Delta_w] = \pi_1^* \mathbb{E}_{\mathcal{D}_1}[\Delta_w] + \sum_{j \neq 1} \pi_j^* \mathbb{E}_{\mathcal{D}_j}[\Delta_w] \leq \pi_1^* \underbrace{|\mathbb{E}_{\mathcal{D}_1}[\Delta_w]|}_{P_1} + \sum_{j\neq 1} \pi_j^* \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_w]|}_{P_j}.$$

Moving along the same trajectory as in (10) for $j \neq 1$ case,

$$P_j \leq O\left((1+\pi_1^*/\pi_j^*)\exp(-\tau_j) + \tau_j/R_{j1}^* + D_m/R_{j1}^*\right).$$

With properly setting parameters similarly as in β case, we get $P_j \leq \rho_{j1}/4k$. For $j=1$ case, in fact, we can reuse the result for (11) as it is. To see this, for instance,

$$|\mathbb{E}_{\mathcal{D}_1}[\Delta_w \mathbb{1}_{\mathcal{E}_1^c}]| \leq \mathbb{E}_{\mathcal{D}_1}[|\mathbb{1}_{\mathcal{E}_1^c}|] = P(\mathcal{E}_1^c) \leq c_1 k\tau/R_{min}.$$

We can do for all cases similarly to get

$$P_1 \leq O\left(k\rho_\pi \exp(-\tau^2) + k\tau/R_{min} + kD_m/R_{min}\right).$$

By setting the parameters similarly as before, *i.e.*, $\tau = \Theta(\sqrt{\log(k\rho_\pi)})$, $R_{min} = \tilde{\Omega}(k)$, $D_m = O(R_{min}/k)$, we can get $P_1 \leq 1/4$ with properly set constants. Therefore, $|\pi_1^+ - \pi_1^*| \leq \pi_1^*/2$ as desired.

A.3 Bounding A

We will prove the following lemma in order to give a lower bound for minimum sinular value of A.

Lemma A.3 *There exists universal constants $c_1 \in (0, 1/2)$ and $c_2, c_3 > 0$, such that:*

$$\lambda_{min}(\mathbb{E}_{\mathcal{D}_1}[w_1 XX^T]) \geq 1 - \left(c_1 + c_2(k\log k)D_m/R_{min} + c_3(k\log^{3/2}(k\rho_\pi))/R_{min}\right).$$

We start it with a following observation.

$$\mathbb{E}_{\mathcal{D}}[w_1 XX^\top] \succeq \pi_1 \mathbb{E}_{\mathcal{D}_1}[w_1 XX^\top].$$

Thus, we will only focus on giving a constant lower bound for $\mathbb{E}_{\mathcal{D}_1}[w_1 XX^\top]$. We define good events as

$$\begin{aligned}
\mathcal{E}_1 &= \{|e| \leq \tau\} \\
\mathcal{E}_2 &= \{|\langle X, \beta_j - \beta_1^*\rangle| \geq 4|\langle X, \Delta_1\rangle|, \forall j \neq 1\} \\
\mathcal{E}_3 &= \{|\langle X, \beta_j - \beta_1^*\rangle| \geq 4\tau, \forall j \neq 1\}.
\end{aligned}$$

We will set $\tau = c_\tau\sqrt{\log(k\rho_\pi)}$ with some large constant $c_\tau > 0$ in this case. Let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$.

Using

$$\mathbb{E}_{\mathcal{D}_1}[w_1 XX^\top] = E_{\mathcal{D}_1}[XX^\top] - \underbrace{E_{\mathcal{D}_1}[(1-w_1)XX^\top \mathbb{1}_{\mathcal{E}}]}_{(i)} - \underbrace{E_{\mathcal{D}_1}[(1-w_1)XX^\top \mathbb{1}_{\mathcal{E}^c}]}_{(ii)},$$

we will give an upper bound to last two terms.

Under $\mathcal{E}$, it can be similarly shown as before that $(1-w_1) \leq 3k\rho_\pi \exp(-\tau^2)$. Thus, (i) is easily bounded:

$$\mathbb{E}_{\mathcal{D}_1}[(1-w_1)XX^\top \mathbb{1}_{\mathcal{E}}] \preceq 3k\rho_\pi \exp(-\tau^2)\mathbb{E}_{\mathcal{D}_1}[XX^\top \mathbb{1}_{\mathcal{E}}] \preceq 3k\rho_\pi \exp(-\tau^2)I.$$

We should split the cases for (ii). Observe that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_1}[(1-w_1)XX^\top \mathbb{1}_{\mathcal{E}^c}] &\preceq \mathbb{E}_{\mathcal{D}_1}[XX^\top \mathbb{1}_{\mathcal{E}^c}] \\
&\preceq \mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_1^c]P(\mathcal{E}_1^c) + \mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_2^c]P(\mathcal{E}_2^c) + \mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_3^c]P(\mathcal{E}_3^c)
\end{aligned}$$

We bound each one by one. First,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_1^c]P(\mathcal{E}_1^c) &= \mathbb{E}_{\mathcal{D}_1}[XX^\top||e| \geq \tau]P(e \geq \tau) \\
&= \mathbb{E}_{\mathcal{D}_1}[XX^\top]P(e \geq \tau) \preceq \exp(-\tau^2/2)I.
\end{aligned}$$

where in the first inequality we used independence of e and X.

For the second term,

$$\mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_2^c] \preceq c_1(\log k)I,$$

from Corollaray 4.2. Meanwhile, we have $P(\mathcal{E}_2^c) \le k\frac{4\|\Delta_1\|}{R_{min}}$. Thus,

$$\mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_2^c]P(\mathcal{E}_2^c) \preceq c_2(k\log k)\frac{D_m}{R_{min}}I.$$

Finally, we bound the operator norm for

$$\mathbb{E}_{\mathcal{D}_1}[XX^\top|\mathcal{E}_3^c] = \mathbb{E}_{\mathcal{D}_1}[XX^\top|\exists j \neq 1, \langle X, \beta_j - \beta_1^*\rangle \le 4\tau] \preceq c_3(\log k)I,$$

from Corollary 4.2. On one hand, $P(\mathcal{E}_3^c) \le k\frac{4\tau}{R_{min}}$. Now combining three pieces, we have

$$\|(ii)\|_{op} \le \exp(-\tau^2/2) + c_4(k\log k)\frac{D_m}{R_{min}} + c_5(k\log k)\frac{\tau}{R_{min}}.$$

Return to bounding $\mathbb{E}_{\mathcal{D}_1}[w_1XX^\top] = I - (i) - (ii)$, we have

$$\mathbb{E}_{\mathcal{D}_1}[w_1XX^\top] \succeq 1 - O\left(k\rho_\pi\exp(-\tau^2/2) + (k\log k)\frac{D_m}{R_{min}} + (k\log k)\frac{\tau}{R_{min}}\right).$$

Giving appropriate $\tau = c_\tau\sqrt{\log k\rho_\pi}$, $D_m/R_{min} \le 1/\tilde{O}(k)$, $R_{min} = \tilde{\Omega}(k)$, we have $\|\mathbb{E}_{\mathcal{D}_1}[w_1XX^\top]\|_{op} \ge 1/2$. Thus, $\|A^{-1}\|_{op} \le 2/\pi_1^*$.

Appendix B Proofs for Finite-Sample EM

B.1 Proofs for concentration of B

To couple it with population EM, we rearrange and write as

$$\beta_1^+ - \beta_1^* = \underbrace{\left(\frac{1}{n}\sum_i w_{1,i}X_iX_i^\top\right)}_{A_n}^{-1}\Bigg(\underbrace{\left(\frac{1}{n}\sum_i w_{1,i}X_i(y_i - \langle X_i, \beta_1^*\rangle) - \mathbb{E}_{\mathcal{D}}[w_1X(Y - \langle X, \beta_1^*\rangle)]\right)}_{e_B}$$
$$+ \underbrace{(\mathbb{E}_{\mathcal{D}}[w_1X(Y - \langle X, \beta_1^*\rangle)] - \mathbb{E}_{\mathcal{D}}[w_1^*X(Y - \langle X, \beta_1^*\rangle)])}_{B}\Bigg).$$

We will consider the following events for concentration result

$$\begin{aligned}
&\mathcal{E}_j = \{\text{sample comes from } j^{th} \text{ linear model}\}\\
&\mathcal{E}_{j,1} = \{|e| \le \tau_j\},\\
&\mathcal{E}_{j,2} = \{4(|\langle X, \Delta_1\rangle| \vee |\langle X, \Delta_j\rangle|) \le |\langle X, \beta_j^* - \beta_1^*\rangle|\},\\
&\mathcal{E}_{j,3} = \{|\langle X, \beta_j^* - \beta_1^*\rangle| \ge 4\sqrt{2}\tau_j\}\\
&\mathcal{E}_{j,good} = \mathcal{E}_{j,1} \cap \mathcal{E}_{j,2} \cap \mathcal{E}_{j,3}
\end{aligned}$$

then decompose each sample using the indicator functions of these events.

$$\begin{aligned}
w_{1,i}X_i(y_i - \langle X, \beta_1^*\rangle) = \Bigg(&\sum_{j\neq 1}^{k} w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}} + w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c}\\
&+ w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}^c} + w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3}^c}\\
&+ w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}} + w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c} + w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap(\mathcal{E}_{j,2}\cup\mathcal{E}_{j,3})^c}\Bigg)\\
&+ w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_1}.
\end{aligned}$$

We will bound the deviation under each event separately. Before getting into detailed analysis, we remind some basics on sub-exponential random variables.

From standard tail bound for sub-exponential random variable W with sub-exponential norm K, we have Vershynin (2010)

$$P\left(\left|\frac{1}{n}\sum_i W_i - \mathbb{E}[W]\right| \geq t\right) \leq 2\exp\left(-Cn\min(t/K,(t/K)^2)\right).$$

If W is a random vector in $\mathbb{R}^d$ with all elements being sub-exponential with same norm K, then

$$\begin{aligned} P\left(\left\|\frac{1}{n}\sum_i W_i - \mathbb{E}[W]\right\| \geq t\right) &\leq \sum_{j=1}^{d} 2P\left(\left|\frac{1}{n}\sum_i (W_i)_j - \mathbb{E}[(W)_j]\right| \geq t/\sqrt{d}\right) \\ &\leq 2d\exp\left(-Cn\min\left(\frac{t}{K\sqrt{d}},\left(\frac{t}{K\sqrt{d}}\right)^2\right)\right) \\ &= \exp\left(-Cn\min\left(\frac{t}{K\sqrt{d}},\left(\frac{t}{K\sqrt{d}}\right)^2\right) + C'\log d\right). \end{aligned} \tag{13}$$

Therefore, in order to achieve δ probability error bound, we should have

$$t = O\left(K\sqrt{d}\left(\frac{\log(d/\delta)}{n} \vee \sqrt{\frac{\log(d/\delta)}{n}}\right)\right). \tag{14}$$

Now we get into concentration of random variables multiplied with indicator functions. For each decomposed random variable, we will find the bound for deviations of empirical mean from true mean that holds with probability at least $1-\delta/k^2T$.

1. $w_{i,1}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}}$: We first check if the target random variable is sub-exponential random vector. For any fixed direction $s \in \mathbb{S}^{d-1}$, we will show $W_i = w_{1,i}\langle X_i, s\rangle\langle X_i, \beta_j^* - \beta_1^*\rangle | \mathcal{E}_j \cap \mathcal{E}_{j,good}$ is sub-exponential by bounding sub-exponential norm.

$$\begin{aligned} \|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p | \mathcal{E}_j \cap \mathcal{E}_{j,good}]^{1/p} \\ &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p | \mathcal{E}_{j,good}]^{1/p}. \end{aligned}$$

   Now recall (8) that how we bounded w_1. Under good event, we know that w_1 is less than 1 or $3\rho_{j1}\exp\left((-5/32\langle X, \beta_j^* - \beta_1^*\rangle^2 + 3e^2)/2\right) \leq 3\rho_{j1}\exp(-\tau_j^2)$. Thus,

$$\begin{aligned} \|W\|_{\psi_1} &= 3\rho_{j1}\exp(-\tau_j^2)\sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p | \mathcal{E}_{j,good}]^{1/p} \\ &\leq \frac{3}{P(\mathcal{E}_{j,good}|\mathcal{E}_j)}\rho_{j1}\exp(-\tau_j^2)\sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}]^{1/2p}\mathbb{E}_{\mathcal{D}_j}[|\langle X, \beta_j^* - \beta_1^*\rangle|^{2p}]^{1/2p} \\ &\leq C\rho_{j1}\exp(-\tau_j^2)R_{j1}^*, \end{aligned}$$

   with sufficiently large constant $C > 0$. We used the fact that $P(\mathcal{E}_{j,good}|\mathcal{E}_j) \geq 1/2$ given good enough initialization and SNR, and l_p-norm of Gaussian is bounded by $O(\sqrt{p})$.

   Now we got a sub-exponential norm of W, we are almost ready to apply our Proposition 5.3. In order to invoke Proposition 5.3, we need to choose proper n_e. First let us bound the probability of large deviation of Bernoulli random variables $Z_i = \mathbb{1}_{(X_i,y_i)\in\mathcal{E}_j\cap\mathcal{E}_{j,good}}$. Note that Bernstein's inequality for Bernoulli random variable is

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]| \geq t\right) \leq \exp\left(-\frac{nt^2}{2t + 2p/3}\right), \tag{15}$$

Observe that $p := P(\mathcal{E}_j \cap \mathcal{E}_{j,good}) \le P(\mathcal{E}_j) = \pi_j^*$. We can choose n_e by checking if the following holds:

$$P(\sum_i Z_i \ge n_e + 1) \le P\left(\frac{1}{n}\sum_i Z_i - p \ge t\right) \le \delta/(k^2 T).$$

By solving the equation (15) $= \delta/(k^2T)$, we get

$$t = O\left(\frac{\log(k^2T/\delta)}{n} + \sqrt{\frac{p\log(k^2T/\delta)}{n}}\right).$$

Therefore, right choice of $n_e = np + O\left(\log(k^2T/\delta) \vee \sqrt{np\log(k^2T/\delta)}\right)$.

We can also use Bernstein's inequality to get

$$P\left(\|\mathbb{E}[W]\| |\frac{1}{n}\sum_{i=1}^n Z_i - E[Z]| \ge t_2\right) \le \exp\left(-\frac{nt_2^2}{(2t_2+3p)\|W\|_{\psi_1}^2}\right), \tag{16}$$

where we used basic fact that $\|E[W]\| \le \|W\|_{\psi_1}$ from Vershynin (2010). For t_2, we set

$$\begin{aligned} t_2 &= \|W\|_{\psi_1} O\left(\frac{\log(k^2T/\delta)}{n} \vee \sqrt{\frac{p}{n}\log(k^2T/\delta)}\right) \\ &= \|W\|_{\psi_1} O\left(\sqrt{p \vee \frac{1}{n}}\sqrt{\frac{\log^2(k^2T/\delta)}{n}}\right). \end{aligned} \tag{17}$$

Then recall (13), we have

$$\begin{aligned} P\left(\|\frac{1}{n_e}\sum_{i=1}^{n_e} W_i - E[W]\| \ge \frac{n}{n_e}t_1\right) &\le \exp\left(-Cn_e \min\left(\frac{n^2t_1^2}{n_e^2 d\|W\|_{\psi_1}^2}, \frac{nt_1}{n_e\sqrt{d}\|W\|_{\psi_1}^2}\right) + C'\log d\right) \\ &= \exp\left(-C \min\left(\frac{n^2t_1^2}{n_e d\|W\|_{\psi_1}^2}, \frac{nt_1}{\sqrt{d}\|W\|_{\psi_1}^2}\right) + C'\log d\right), \end{aligned} \tag{18}$$

Therefore, proper scale of t_1 is

$$\begin{aligned} t_1 &= O\left(\|W\|_{\psi_1}\sqrt{\frac{n_e}{n}}\sqrt{\frac{d}{n}\log(dk^2T/\delta)} \vee \|W\|_{\psi_1}\frac{\sqrt{d}\log(dk^2T/\delta)}{n}\right) \\ &\le \|W\|_{\psi_1} O\left(\sqrt{p \vee \frac{1}{n}}\sqrt{\frac{d}{n}\log^2(dk^2T/\delta)}\right). \end{aligned} \tag{19}$$

Since $n = \tilde{\Omega}(1/\pi_{min})$ as we will use, with probability at least $1 - 3\delta/(k^2T)$,

$$\begin{aligned} &\|\frac{1}{n}\sum_i w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}} - E[w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}}]\| \\ &\qquad = O\left(\rho_{j1}R_{j1}^*\exp(-\tau_j^2)\sqrt{\pi_j^*}\sqrt{\frac{d}{n}\log^2(dk^2T/\delta)}\right) \end{aligned}$$

2. $w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c}$: It corresponds to the case where the noise power is larger than τ_j. The probability of this event is $p := P(\mathcal{E}_j \cap \mathcal{E}_{j,1}^c) \le 2\pi_j^*\exp(-\tau_j^2/2)$. In this case, $W_i = w_{1,i}\langle X_i, s\rangle\langle X_i, \beta_j^* - \beta_1^*\rangle | \mathcal{E}_j \cap \mathcal{E}_{j,1}^c$ is

bounded as

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_j \cap \mathcal{E}_{j,1}^c]^{1/p} \\
&= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_{j,1}^c]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_{j,1}^c]^{1/p} \\
&= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p]^{1/p} \\
&\leq CR_{j1}^*,
\end{aligned}$$

for some constant C, where the last equality comes from the fact that X and e are independent. While we want to reuse (17) and (19) to decide deviation of means under this event, we also need to cancel out the norm of $W = O(R_{j1}^*)$. We consider two cases: $1/n < p^{1/c}$ and $1/n > p^{1/c}$ for some number $c \geq 2$.

If $1/n < p^{1/c}$, then $\sqrt{p \vee 1/n} \leq p^{1/2c} = 2\exp(-\tau_j^2/(4c))$. We can just plug in this bound into (17) and (19) to get the deviation

$$\begin{aligned}
&\|\frac{1}{n}\sum_i w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c} - E[w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c}]\| \\
&\qquad = O\left(R_{j1}^*\exp(-\tau_j^2/(4c))\sqrt{\frac{d}{n}\log^2(dk^2T/\delta)}\right),
\end{aligned}$$

with probability at least $1 - \delta/(k^2T)$.

On the other side, if $1/n > p^{1/c}$, then we will set $n_e = 0$, *i.e.*, no sample fell into this event. This is true with probability $1 - np = 1 - 1/n^{c-1}$. The statistical error is thus just

$$\mathbb{E}[W]p = O(R_{j1}^*\exp(-\tau_j^2/2)) \leq O(R_{j1}^*\exp(-\tau_j^2/4)/n).$$

By setting $c = 4$ and $n \geq (k^2T/\delta)^{1/3}$, this will hold with probability at least $1 - \delta/(k^2T)$.

3. $w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}^c}$: Under this event, we first note that $|w_{1,i}\langle X_i, s\rangle\ \langle X_i, \beta_j^* - \beta_1^*\rangle| \leq 4|\langle X_i, s\rangle|(|\langle X_i, \Delta_j\rangle| + |\langle X_i, \Delta_j\rangle|)$. In turn,

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_j \cap \mathcal{E}_{j,1} \cap \mathcal{E}_{j,2}^c]^{1/p} \\
&= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_{j,1} \cap \mathcal{E}_{j,2}^c]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_{j,2}^c]^{1/p} \\
&\leq 4\sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}\left[|\langle X, s\rangle(\langle X, \Delta_1\rangle + \langle X, \Delta_j\rangle)|^p|\ \mathcal{E}_{j,2}^c\right]^{1/p} \\
&\underset{(i)}{\leq} 4\sup_{p\geq 1} p^{-1}\left(\sqrt{\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}|\mathcal{E}_{j,2}^c]}\sqrt{|\langle X, \Delta_j\rangle|^{2p}|\mathcal{E}_{j,2}^c]}\right)^{1/p} \\
&\quad + 4\sup_{p\geq 1} p^{-1}\left(\sqrt{\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}|\mathcal{E}_{j,2}^c]}\sqrt{|\langle X, \Delta_1\rangle|^{2p}|\mathcal{E}_{j,2}^c]}\right)^{1/p} \\
&\underset{(ii)}{\leq} c_2 D_m,
\end{aligned}$$

where (i) we used Minkowski's inequality and Cauchy-Schwartz inequality, then (ii) we invoked Lemma A.1. Recall that $D_m = \max_j \|\Delta_j\|$. Thus $W = w_1 X\langle X, \beta_j^* - \beta_1^*\rangle|(\mathcal{E}_j \cap \mathcal{E}_{j,2}^c)$ is sub-exponential with parameter at most $c_2 D_m$. We can also check that $p := P(\mathcal{E}_j \cap \mathcal{E}_{j,2}^c) \leq O(\pi_j^* D_m/R_{j1}^*)$.

We choose $n_e = np + O(\log(k^2T/\delta) \vee \sqrt{np\log(k^2T/\delta)})$ as before. Using (17) and (19),

$$t_1 = O\left(D_m\sqrt{p \vee \frac{\log(dk^2T/\delta)}{n}}\sqrt{\frac{d}{n}\log(dk^2T/\delta)}\right),$$

$$t_2 = O\left(D_m\sqrt{\frac{p\log(k^2T/\delta)}{n}} \vee D_m\frac{\log(k^2T/\delta)}{n}\right). \tag{20}$$

We can see that $n = \Omega(dk\log(dk^2T/\delta)/\pi_1^*)$ suffices to ensure $t_1, t_2 < O(D_m\pi_1^*/k)$ since $p \leq O(\pi_j^*/(k\rho_\pi)) = O(\pi_1^*/k)$ by the initialization condition. Overall, $t_1 + t_2$ is bounded by

$$O\left(D_m\sqrt{\frac{\pi_j^* D_m}{R_{j1}^*} \vee \frac{\log(dk^2T/\delta)}{n}}\sqrt{\frac{d}{n}\log(dk^2T/\delta)}\right).$$

4. $w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3}^c}$: In this case, we define W as

$$W_i = w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle\mathbb{1}_{\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}}|(\mathcal{E}_j\cap\mathcal{E}_{j,3}^c).$$

In other words, we are leaving some events in the indicator. Then, we can restart from bounding the sub-exponential norm of W.

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p\mathbb{1}_{\mathcal{E}_{j,1}\cap\mathcal{E}_{j,2}}|\cap\mathcal{E}_{j,3}^c]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{X\sim\mathcal{N}(0,I)}[\left(\mathbb{E}_{Y\sim\mathcal{N}(\langle X,\beta_j^*\rangle,1)}[w_1]\right)|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p\mathbb{1}_{\mathcal{E}_{j,2}}|\mathcal{E}_{j,3}^c]^{1/p}.
\end{aligned}$$

Then, use the following bound for expectation of w_1 when $\mathcal{E}_{j,2}$ is true,

$$\begin{aligned}
\mathbb{E}_{Y\sim\mathcal{N}(\langle X,\beta_j^*\rangle,1)}[w_1] &\leq \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[\min\left(3\rho_{j1}\exp\left((-5/32\langle X, \beta_j^* - \beta_1^*\rangle^2 + 3e^2)/2\right), 1\right)\right] \\
&\leq \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[3\rho_{j1}\exp\left(-5/32\langle X, \beta_j^* - \beta_1^*\rangle^2 + 3e^2\right)\mathbb{1}_{3e^2\leq|\langle X,\beta_j^*-\beta_1^*\rangle|^2/32|}\right] \\
&+ \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[\mathbb{1}_{3e^2\geq\langle X,\beta_j^*-\beta_1^*\rangle^2/32}\right] \\
&\leq \mathbb{E}_{e\sim\mathcal{N}(0,1)}\left[3\rho_{j1}\exp\left(-\langle X, \beta_j^* - \beta_1^*\rangle^2/16\right)\right] + P(3e^2 \geq |\langle X, \beta_j^* - \beta_1^*\rangle|^2/32) \\
&\leq 5(1\vee\rho_{j1})\exp(-\langle X, \beta_j^* - \beta_1^*\rangle^2/192).
\end{aligned} \tag{21}$$

Then we compute the upper bound for $\exp(-\langle X, \beta_j^* - \beta_1^*\rangle^2/192)|\langle X, \beta_j^* - \beta_1^*\rangle|^p$. Letting $|\langle X, \beta_j^* - \beta_1^*\rangle| = a$, and find a maximum for $-a^2/192 + p\log a$ by finding a zero point in its derivative. We get a maximum at $a = \sqrt{96p}$ with value $(96p)^{p/2}\exp(-p/2)$. Now plug this upper bound to continue bounding norm of W.

$$\begin{aligned}
\|W\|_{\psi_1} &\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[\left(\mathbb{E}_{Y\sim\mathcal{N}(\langle X,\beta_j^*\rangle,1)}[w_1]\right)\mathbb{1}_{\mathcal{E}_{j,2}}|\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p|\mathcal{E}_{j,1}^c\right]^{1/p} \\
&\leq 5(1\vee\rho_{j1})\sup_{p\geq 1} p^{-1}(96p)^{1/2}\exp(-1/2)\mathbb{E}_{X\sim\mathcal{N}(0,I)}\left[|\langle X, s\rangle|^p|\mathcal{E}_{j,1}^c\right]^{1/p} \\
&\leq C(1\vee\rho_{j1}),
\end{aligned}$$

with sufficiently large constant $C > 0$ and Lemma A.2 for the final inequality.

The probability of this event $p := P(\mathcal{E}_j\cap\mathcal{E}_{j,2}^c) \leq 4\sqrt{2}\pi_j^*\tau_j/R_{j1}^*$. Again we use Proposition 5.3 to get a deviation of this random variable. We can set t_1 and t_2 as

$$t_1 = O\left((1\vee\rho_{j1})\sqrt{p \vee \frac{\log(dk^2T/\delta)}{n}}\sqrt{\frac{d\log(dk^2T/\delta)}{n}}\right),$$

$$t_2 = O\left((1\vee\rho_{j1})\sqrt{\frac{p\log(k^2T/\delta)}{n}} \vee (1\vee\rho_{j1})\frac{\log(k^2T/\delta)}{n}\right).$$

With probability at least $1 - \delta/k^2T$, we conclude that the deviation of sum under this event is

$$O\left((1 \vee \rho_{j1})\left(\sqrt{\frac{\log(dk^2T/\delta)}{n}} \vee \sqrt{\frac{\pi_j^* \tau_j}{R_{j1}^*}}\right)\sqrt{\frac{d}{n}\log(dk^2T/\delta)}\right).$$

Now we will bound terms for $w_{1,i}X_ie_i$, it is almost exact repetition of previous procedures when it comes from $j^{th} \neq 1$ component.

1. $w_{i,1}X_ie_i\mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,good}}$, $j \neq 1$: One can show that following the exact same procedure for the first case we handled for $w_{i,1}X_i\langle X, \beta_j^* - \beta_1^*\rangle$. In this case, $W_i = w_{i,1}X_ie_i|\mathcal{E}_j \cap \mathcal{E}_{j,good}$, we can get $\|W\|_{\psi_1} \leq C\rho_{j1}\exp(-\tau_j^2)$. To see this,

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle e|^p|\mathcal{E}_j \cap \mathcal{E}_{j,good}]^{1/p} \\
&= \sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle|^p\mathcal{E}_{j,good}]^{1/p} \\
&\leq 3/P(\mathcal{E}_{j,good}|\mathcal{E}_j)\rho_{j1}\exp(-\tau_j^2)\sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle e|^p]^{1/p} \\
&\leq C\rho_{j1}\exp(-\tau_j^2).
\end{aligned}$$

Following the same trick we used with Proposition 5.3, (see (17) and (19)), we get

$$\|\frac{1}{n}\sum_i w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,good}} - \mathbb{E}[w_1Xe\mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,good}}]\| = O\left(\rho_{j1}\exp(-\tau_j^2)\sqrt{\pi_j^*}\sqrt{\frac{d}{n}\log(dk^2T/\delta)}\right).$$

2. $w_{i,1}X_ie_i\mathbb{1}_{\mathcal{E}_j \cap \mathcal{E}_{j,1}^c}$, $j \neq 1$: The challenge here is how to bound $\mathbb{E}_{e \sim \mathcal{N}(0,1)}[|e|^p||e| \geq \tau_j]$. We will use the standard lower bound for Gaussian tail bound:

$$P(e \geq \tau_j) \geq \frac{\tau_j}{\tau_j^2 + 1}\frac{1}{\sqrt{2\pi}}\exp(-\tau_j^2/2) \geq \frac{\exp(-\tau_j^2/2)}{3\tau_j}.$$

Now for the sub-exponential norm of $W = w_1Xe|\mathcal{E}_j \cap \mathcal{E}_{j,1}^c$ is given as

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle e|^p|\mathcal{E}_{j,1}^c]^{1/p} \\
&\leq \sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^p]^{1/p}\mathbb{E}_{\mathcal{D}_j}[|e|^p|\mathcal{E}_{j,1}^c]^{1/p} \\
&= \sup_{p \geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^p]^{1/p}\left(\mathbb{E}_{\mathcal{D}_j}[|e|^p\mathbb{1}_{\mathcal{E}_{j,1}^c}]/P(\mathcal{E}_{j,1}^c)\right)^{1/p}.
\end{aligned}$$

where in the inequality we used the independence of X and e. $\mathbb{E}_{e \sim \mathcal{N}(0,1)}[|e|^p\mathbb{1}_{e \geq \tau_j}]$ can be upper-bounded as follows:

$$\begin{aligned}
\mathbb{E}_{e \sim \mathcal{N}(0,1)}[|e|^p\mathbb{1}_{e \geq \tau_j}] &= \mathbb{E}_{e \sim \mathcal{N}(0,1)}[|e|^p\mathbb{1}_{2\tau_j \geq |e| \geq \tau_j}] + \mathbb{E}_{e \sim \mathcal{N}(0,1)}[|e|^p\mathbb{1}_{|e| \geq 2\tau_j}] \\
&\leq (2\tau_j)^pP(|e| \geq \tau_j) + \sqrt{\mathbb{E}[|e|^{2p}]}\sqrt{P(|e| \geq 2\tau_j)}.
\end{aligned}$$

For the comparison of $\sqrt{P(|e| \geq 2\tau_j)}$ and $P(|e| \geq \tau_j)$, the standard lower and upper bounds for Gaussian tail are useful. That is,

$$\begin{aligned}
P(e \geq \tau_j) &\geq \frac{x}{x^2 + 1}\frac{1}{\sqrt{2\pi}}\exp(-\tau_j^2/2), \\
P(e \geq 2\tau_j) &\leq \exp(-2\tau_j^2).
\end{aligned}$$

Thus,

$$\sqrt{P(|e| \geq 2\tau_j)}/P(|e| \geq \tau_j) \leq 8\tau_j\exp(-\tau_j^2/2).$$

Now we can plug those values we found to proceed

$$\begin{aligned}
\|W\|_{\psi_1} &\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^p]^{1/p}\left((2\tau_j)^p + \sqrt{\mathbb{E}[|e|^{2p}]}8\tau_j\exp(-\tau_j^2/2)\right)^{1/p} \\
&\leq c_0 \sup_{p\geq 1} p^{-1/2}\left(2\tau_j + \sqrt{\mathbb{E}[|e|^{2p}]}^{1/p}(8\tau_j\exp(-\tau_j^2/2))^{1/p}\right) \\
&\leq C\tau_j,
\end{aligned}$$

for some universal constant c_0, C. Now we have the sub-exponential norm of W, we can follow the procedure for $w_{1,i}X_i\langle X_i, \beta_j^* - \beta_1^*\rangle \mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}^c}$. Similarly to previously guaranteed in Remark 7, the deviation will be given as

$$O\left(\tau_j\exp(-\tau_j^2/(4c))\sqrt{\frac{d}{n}\log^2(dk^2T/\delta)}\right).$$

Again, we may set $c = 4$ and $n > (k^2T/\delta)^{1/3}$ to get $\delta/(k^2T)$ probability bound.

3. $w_{i,1}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,1}\cap(\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3})^c}$, $j \neq 1$: For this case, we set $W_i = w_{i,1}X_ie_i\mathbb{1}_{\mathcal{E}_{j,1}}|\mathcal{E}_j\cap(\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3})^c$ and find that

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle e|^p\mathbb{1}_{\mathcal{E}_{j,1}}|\mathcal{E}_j\cap(\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3})^c]^{1/p} \\
&= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|w_1\langle X, s\rangle e|^p\mathcal{E}_{j,1}|(\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3})^c]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle e|^p|\mathcal{E}_{j,2}^c\cup\mathcal{E}_{j,3}^c]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\left(\sqrt{\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}|\mathcal{E}_{j,2}^c\cup\mathcal{E}_{j,3}^c]}\sqrt{\mathbb{E}[|e|^{2p}|\mathcal{E}_{j,2}^c\cup\mathcal{E}_{j,3}^c]}\right)^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\left(\sqrt{\mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}|\mathcal{E}_{j,2}^c] + \mathbb{E}_{\mathcal{D}_j}[|\langle X, s\rangle|^{2p}|\mathcal{E}_{j,3}^c]}\sqrt{\mathbb{E}[|e|^{2p}]}\right)^{1/p} \\
&\leq C,
\end{aligned}$$

for some constant $C > 0$. The probability is bounded as $P(\mathcal{E}_j\cap(\mathcal{E}_{j,2}\cap\mathcal{E}_{j,3})^c) \leq \pi_j^* O\left(D_m/R_{j1}^* + \tau_j/R_{j1}^*\right)$, so we can bound the deviation in this case as

$$\|\frac{1}{n}\sum_i w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,2}^c} - E[w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,2}^c}]\| = O\left(\sqrt{\frac{\pi_j^*}{k\rho_\pi}\vee\frac{\log(dk^2T/\delta)}{n}}\sqrt{\frac{d\log(dk^2T/\delta)}{n}}\right).$$

given our initialization and SNR condition.

4. $w_{i,1}X_ie_i\mathbb{1}_{\mathcal{E}_1}$ $(j = 1)$: Finally, it is the easiest case since

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}}[|w_1\langle X, s\rangle e|^p|\mathcal{E}_1]^{1/p} \\
&= \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_1}[|w_1\langle X, s\rangle e|^p]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\mathbb{E}_{\mathcal{D}_1}[|\langle X, s\rangle e|^p]^{1/p} \\
&\leq \sup_{p\geq 1} p^{-1}\left(\sqrt{\mathbb{E}_{\mathcal{D}_1}[|\langle X, s\rangle|^{2p}]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[|e|^{2p}]}\right)^{1/p} \\
&\leq c_3,
\end{aligned}$$

for some constant c_3. We can apply the same trick and get

$$\|\frac{1}{n}\sum_i w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_1} - E[w_{1,i}X_ie_i\mathbb{1}_{\mathcal{E}_1}]\| = O\left(\sqrt{\pi_1^*\vee\frac{\log(dk^2T/\delta)}{n}}\sqrt{\frac{d\log(dk^2T/\delta)}{n}}\right).$$

Now we collect every scale of deviations from each item, and conclude that with probability $1 - \delta/kT$ (by taking union bound over $O(k)$ items), we have

$$\begin{aligned} e_B &= \frac{1}{n}\sum_i w_{1,i} X_i (Y_i - \langle X, \beta_1^* \rangle) - \mathbb{E}_{\mathcal{D}}[w_1 X (Y - \langle X, \beta_1^* \rangle)] \\ &\leq \sqrt{\frac{d}{n}\log^2(dk^2T/\delta)} \Bigg(\sum_{j\neq 1}^{k} \sqrt{\pi_j^*}\rho_{j1} R_{j1}^* \exp(-\tau_j^2) + R_{j1}^* \exp(-\tau_j^2/16) + D_m \sqrt{\frac{\pi_j^* D_m}{R_{j1}^*} \vee \frac{1}{n}} \\ &+ (1 \vee \rho_{j1}) \sqrt{\frac{1}{n} \vee \frac{\pi_j^* \tau_j}{R_{j1}^*}} + \sqrt{\pi_j^*}\rho_{j1}\exp(-\tau_j^2) + \tau_j \exp(-\tau_j^2/16) + \sqrt{\frac{\pi_j^*}{k\rho_\pi} \vee \frac{1}{n}} \Bigg) \\ &+ \sqrt{\frac{d\log(dk^2T/\delta)}{n}} \sqrt{\pi_1^*}. \end{aligned} \tag{22}$$

As we set $\tau_j = c_\tau \sqrt{\log(k\rho_\pi R_{j1}^*)}$, SNR and initialization condition

$$\begin{aligned} R_{j1}^* &= \Omega\left(k\rho_\pi \ \log(\rho_\pi k R_{j1}^*)\right) = \tilde{\Omega}(k), \\ D_m/R_{j1}^* &\leq O(1/(k\rho_\pi)), \end{aligned}$$

and sample complexity

$$n \gg \left(k/\pi_{min}^* (d/\epsilon^2) \log^2(dk^2T/\delta)\right) \vee (k^2T/\delta)^{1/3}, \tag{23}$$

every term inside the summation in (22) is now less than $O(\sqrt{\pi_1^*/k})$ or $O(D_m\sqrt{\pi_1^*/k})$. Thus,

$$e_B \leq O\left(\epsilon\sqrt{\frac{\pi_{min}}{k}} \left(k(1+D_m)\sqrt{\pi_1^*/k}\right) + \epsilon \pi_1^* \right).$$

We can conclude that $e_B \leq \pi_1^* D_m \epsilon + \pi_1^* \epsilon$ with probability at least $1 - \delta/kT$ (changing δ to $c\delta$ with some constant c).

Remark 7 *The high probability result is usually given as $1 - \exp(-cn)$, but it is also enough to show that it holds with probability $1 - 1/n^c$ for some constant $c > 0$. The choice of 3 is rather arbitrary and we could have picked any other larger constant with slight constant penalty in SNR requirement.*

B.2 Concentration of A

As we only are interested in lower bound of the minimum eigenvalue, we only need to consider the concentration of $w_{1,i} X_i X_i^\top \mathbb{1}_{\mathcal{E}_j}$ since $\frac{1}{n}\sum_i w_{1,i} X_i X_i^\top \succeq \frac{1}{n}\sum_i w_{1,i} X_i X_i^\top \mathbb{1}_{\mathcal{E}_1}$. The concentration comes from standard concentration argument for random matrix with sub-exponential norm Vershynin (2010). For any fixed $s \in \mathbb{S}^{d-1}$, we have

$$\|w_1 \langle X, s \rangle^2\|_{\psi_1} \leq 2\|\langle X, s\rangle\|_{\psi_2}^2 \leq K,$$

with some universal constant K, since w_1 is bounded in [0,1]. Using this and (1/2) covering-net argument over unit sphere, and the same argument we used with Proposition 5.3, we get

$$\left\|\frac{1}{n}\sum_i w_{1,i} X_i X_i^\top \mathbb{1}_{\mathcal{E}_1} - \mathbb{E}_{\mathcal{D}}[w_1 X X^\top \mathbb{1}_{\mathcal{E}_1}]\right\|_{op} \leq O\left(\sqrt{\pi_1^*}\sqrt{\frac{d\log(k^2T/\delta)}{n}}\right),$$

with high probability. As we see in the proof of Appendix A.3, $\mathbb{E}_{\mathcal{D}}[w_1 X X^\top \mathbb{1}_{\mathcal{E}_1}] \succeq (\pi_1^*/2) I$ with good initialization and SNR. Thus,

$$\frac{1}{n}\sum_i w_{1,i} X_i X_i^\top \succeq \frac{\pi_1^*}{2} I - \sqrt{\frac{\pi_1^* d\log(k^2T/\delta)}{n}} I \succeq \left(\frac{\pi_1^*}{2} - \epsilon\pi_1^*\right) I,$$

given $n = \Omega(d\log(k^2T/\delta)/\pi_{min})$. Thus, we can get $\|A_n^{-1}\|_{op} \leq 2/\pi_1^*$ with high probability.

B.3 Concentration of Mixing Weights

We can again use the per-sample decomposition strategy. The target we will bound is the error $\left|\frac{1}{n}\sum_i w_{1,i} - \mathbb{E}_{\mathcal{D}}[w_1]\right|$. As before, decompose $w_{1,i}$ as

$$w_{1,i} = \left(\sum_{j>1}^{k} w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}} + w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}^c}\right) + w_{1,i}\mathbb{1}_{\mathcal{E}_1}.$$

It is the repetition of proofs for other quantities we have considered so far.

1. $w_{i,1}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}}$, $j \neq 1$: The difference is, now in all cases W is a sub-Gaussian random variable. Note that w_1 is always less than 1.

$$\begin{aligned}
\|W\|_{\psi_1} &= \sup_{p\geq 1} p^{-1/2}\mathbb{E}_{\mathcal{D}}[|w_1|^p|\mathcal{E}_j\cap\mathcal{E}_{j,good}]^{1/p} \\
&= \sup_{p\geq 1} p^{-1/2}\mathbb{E}_{\mathcal{D}_j}[|w_1|^p|\mathcal{E}_{j,good}]^{1/p} \\
&\leq C\rho_{j1}\exp(-\tau_j^2),
\end{aligned}$$

   Following the same trick we used with Proposition 5.3,with probability at least $1-\delta/(k^2T)$, we get

$$|\frac{1}{n}\sum_i w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}} - E[w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}}]| = O\left(\rho_{j1}\exp(-\tau_j^2)\sqrt{\pi_j^*}\sqrt{\frac{1}{n}\log(k^2T/\delta)}\right).$$

2. $w_{i,1}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,good}^c}$, $j \neq 1$: For this case, we set $W = w_{i,1}|\mathcal{E}_j\cap\mathcal{E}_{j,good}^c$ and find that

$$\|W\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\mathbb{E}_{\mathcal{D}}[|w_1|^p\mathcal{E}_j\cap\mathcal{E}_{j,good}^c]^{1/p} \leq 1.$$

   The probability is bounded as $P(\mathcal{E}_j\cap\mathcal{E}_{j,good}^c) \leq \pi_j^* O\left(\exp(-\tau_j^2/2) + D_m/R_{j1}^* + \tau_j/R_{j1}^*\right) \leq O(\pi_j^*/(k\rho_\pi))$, so we can bound the deviation in this case as

$$|\frac{1}{n}\sum_i w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,2}^c} - E[w_{1,i}\mathbb{1}_{\mathcal{E}_j\cap\mathcal{E}_{j,2}^c}]| = O\left(\sqrt{\frac{\pi_j^*}{k\rho_\pi}\vee\frac{\log(k^2T/\delta)}{n}}\sqrt{\frac{\log(k^2T/\delta)}{n}}\right).$$

   given our initialization and SNR condition.

3. $w_{i,1}\mathbb{1}_{\mathcal{E}_1}$: Finally, it is the easiest case since

$$\|W\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\mathbb{E}_{\mathcal{D}}[|w_1|^p|\mathcal{E}_1]^{1/p} \leq 1,$$

   We can apply the same trick and get

$$|\frac{1}{n}\sum_i w_{1,i}\mathbb{1}_{\mathcal{E}_1} - E[w_{1,i}\mathbb{1}_{\mathcal{E}_1}]| = O\left(\sqrt{\pi_1^*\vee\frac{\log(k^2T/\delta)}{n}}\sqrt{\frac{\log(k^2T/\delta)}{n}}\right).$$

Now combining this all, given $n = \Omega(k\epsilon^{-2}/\pi_{min})$ we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_i w_{1,i} - \mathbb{E}_{\mathcal{D}}[w_1]\right| &\leq \sqrt{\frac{1}{n}\log(k^2T/\delta)}\left(\sum_{j>1}^{k}\rho_{j1}\exp(-\tau_j^2)\sqrt{\pi_j^*} + \sqrt{\frac{\pi_1^*}{k}}\right) + \sqrt{\frac{\pi_1^*\log(k^2T/\delta)}{n}} \\
&\leq \epsilon\sqrt{\frac{\pi_1^*}{k}}\left(\sum_{j>1}^{k}\frac{\sqrt{\rho_{j1}}\sqrt{\pi_1^*}}{k\rho_\pi} + \sqrt{\frac{\pi_1^*}{k}}\right) + \epsilon\pi_1^* \leq O(\pi_1^*\epsilon).
\end{aligned}$$

This implies the concentration of mixing weights in relative scale.

Appendix C Proof of Auxiliary Lemmas

Lemma A.1 *Let $X \sim \mathcal{N}(0, I_d)$. For any fixed vector $v \in \mathbb{R}^d$, and a set of vectors $u_1, ..., u_k \in \mathbb{R}^d$ such that $\|u_j\| \geq \|v\|$ for all j, we define*

$$\mathcal{E} := \{|\langle X, u_j \rangle| \geq |\langle X, v \rangle|, \ \forall j = 1, ..., k\}.$$

Then,

$$P(\mathcal{E}^c) \leq \sum_{j=1}^{k} \frac{\|v\|}{\|u_j\|}. \tag{4}$$

Furthermore, for any unit vector $s \in \mathbb{S}^{d-1}$ and for any $p \geq 1$, we have

$$\mathbb{E}[|\langle X, s \rangle|^p | \mathcal{E}^c] \leq k 2^p \Gamma(1 + p/2), \tag{5}$$

where Γ is a gamma function.

Proof. Equation (4) is a consequence of Lemma 6 in Yi et al. (2016) and elementary rule of union bounds.

For (5), we first look at p^{th} moment conditioned on only one event. Recall that in Yi et al. (2016), only the case for $p = 2$ is proven. Without loss of generality, due to the rotational invariance of standard Gaussian distribution, we can assume $\text{span}(u, v_1) = \text{span}(\boldsymbol{e}_1, \boldsymbol{e}_2)$.

Change first two coordiantes of X, x_1, x_2 to combination of r Rayleigh distribution and θ uniformly distributed over $[0, 2\pi)$. Then define $Y = \langle s_{3:d}, X_{3:d} \rangle$ where $s_{3:d}, X_{3:d}$ be partial vectors of s and X from third coordinate. Then $Y \sim \mathcal{N}(0, \|s_{3:d}\|^2)$, and r, θ, Y are all independent.

Now note that the event $\mathcal{E}_1 = |\langle X, u_1 \rangle| \geq |\langle X, v \rangle|$ only depends on θ. Then,

$$\begin{aligned}
\mathbb{E}[\langle X, s \rangle^p | \mathcal{E}_1^c] &= \mathbb{E}[|s_1 r \cos\theta + s_2 r \sin\theta + Y|^p | \mathcal{E}_1^c] \\
&= \frac{\mathbb{E}[|s_1 r \cos\theta + s_2 r \sin\theta + Y|^p \mathbb{1}_{\mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&= \frac{\mathbb{E}_\theta[\mathbb{E}_{r,Y}[|r s_1 \cos\theta + r s_2 \sin\theta + Y|^p | \theta] \ \mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&= \frac{\mathbb{E}_\theta[(\mathbb{E}_{r,Y}[|r s_1 \cos\theta + r s_2 \sin\theta + Y|^p | \theta]^{1/p})^p \ \mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&\underset{(i)}{\leq} \frac{\mathbb{E}_\theta[(\mathbb{E}_r[|r s_1 \cos\theta + r s_2 \sin\theta|^p | \theta]^{1/p} + \mathbb{E}_Y[|Y|^p | \theta]^{1/p})^p \ \mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&\underset{(ii)}{\leq} \frac{\mathbb{E}_\theta[(\mathbb{E}_r[r^p | s_1 \cos\theta + s_2 \sin\theta|^p | \theta]^{1/p} + \mathbb{E}_Y[|Y|^p]^{1/p})^p \ \mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&\leq \frac{\mathbb{E}_\theta[\mathbb{E}_r[r^p]^{1/p} \|s_{1:2}\| + \mathbb{E}_Y[|Y|^p]^{1/p})^p \ \mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&\underset{(iii)}{=} \frac{(\mathbb{E}_r[r^p]^{1/p} \|s_{1:2}\| + \mathbb{E}_Y[|Y|^p]^{1/p})^p \ \mathbb{E}_\theta[\mathbb{1}_{\theta \in \mathcal{E}_1^c}]}{P(\mathcal{E}_1^c)} \\
&= (\mathbb{E}[r^p]^{1/p} \|s_{1:2}\| + \mathbb{E}[|Y|^p]^{1/p})^p,
\end{aligned}$$

where (i) we used Minkowski's inequality, (ii) we used independence of θ and Y, and (iii) used independence of all terms from θ.

Then, since $r \sim \text{Rayleigh}(1)$ and $Y \sim \mathcal{N}(0, \|s_{3:d}\|^2)$, we have an exact value for each p^{th} moments from well-known distribution properties. That is,

$$\mathbb{E}[\langle X, s \rangle^p | \mathcal{E}^c] \leq \left(\|s_{1:2}\| \sqrt{2} \Gamma(1 + p/2)^{1/p} + \|s_{3:d}\| \sqrt{2} \left(\Gamma((p+1)/2) / \sqrt{\pi} \right)^{1/p} \right)^p.$$

Now since $\Gamma(1 + p/2) \geq 2\Gamma((p+1)/2)/\sqrt{\pi}$ for $p \geq 1$, and

$$\|s_{1:2}\| + \|s_{3:d}\| \leq \sqrt{\|s_{1:2}\|^2 + \|s_{3:d}\|^2} \sqrt{2} = \sqrt{2}$$

since s is an unit vector, we conclude that

$$\mathbb{E}[\langle X, s\rangle^p | \mathcal{E}_1^c] \leq 2^p \Gamma(1 + p/2).$$

Now we move on to conditioning on $\mathcal{E}^c$. It comes from elementary property of union of the events,

$$\begin{aligned}
\mathbb{E}[|\langle X, s\rangle|^p | \mathcal{E}^c] &= \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}^c}]}{P(\mathcal{E}^c)} \leq \frac{\mathbb{E}[|\langle X, s\rangle|^p \sum_i \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}^c)} \\
&= \sum_i \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}^c)} \leq \sum_i \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}_i^c)} \\
&\leq k 2^p \Gamma(1 + p/2),
\end{aligned}$$

where we used $P(\mathcal{E}^c) \geq P(\mathcal{E}_i^c)$, and $\mathbb{1}_{\mathcal{E}^c} \leq \sum_i \mathbb{1}_{\mathcal{E}_i^c}$ since $\mathcal{E}^c = \cup_i \mathcal{E}_i^c$. The claim follows. □

Lemma A.2 *Let $X \sim \mathcal{N}(0, I_d)$. For any set of fixed vectors $u_1, ..., u_k \in \mathbb{R}^d$, and fixed constants $\alpha_1, ..., \alpha_k > 0$, define*

$$\mathcal{E} := \{|\langle X, u_j\rangle| \geq \alpha_j, \ \forall j = 1, ..., k\}.$$

Then,

$$P(\mathcal{E}^c) \leq \sum_{j=1}^k \frac{\alpha_j}{\|u_j\|}. \tag{6}$$

Furthermore, for any unit vector $s \in \mathbb{S}^{d-1}$ and for $p \geq 1$, we have

$$\mathbb{E}[|\langle X, s\rangle|^p | \mathcal{E}^c] \leq k 2^p \Gamma((1 + p)/2)/\sqrt{\pi}. \tag{7}$$

Proof. Equation (6) is a direct consequence of lemma 9(v) in Balakrishnan et al. (2017) and union bound.

We start of (7) with the same strategy in proof of A.1. Let us consider only one comparison first. Let $\mathcal{E}_1 = \{|\langle X, u_1\rangle \geq \alpha_1\}$. Without loss of generality (by rotational invariance of standard Gaussian), let $u_1 = \boldsymbol{e}_1$ and $Y = \langle x_{2:d}, s_{2:d}\rangle$.

$$\begin{aligned}
\mathbb{E}[|\langle X, s\rangle|^p | \mathcal{E}_1^c] &= \mathbb{E}[|s_1 x_1 + Y|^p | (|x_1| \leq \alpha_1)] \\
&= \frac{\mathbb{E}[|s_1 x_1 + Y|^p \mathbb{1}_{x_1 \leq \alpha_1}]}{P(|x_1| \leq \alpha_1)} \\
&\leq \frac{\mathbb{E}[\mathbb{E}[|s_1 x_1 + Y|^p | x_1] \mathbb{1}_{x_1 \leq \alpha_1}]}{P(x_1 \leq \alpha_1)} \\
&\leq \frac{\mathbb{E}[(\mathbb{E}[|s_1 x_1|^p | x_1]^{1/p} + \mathbb{E}[|Y|^p | x_1]^{1/p})^p | x_1] \mathbb{1}_{x_1 \leq \alpha_1}]}{P(x_1 \leq \alpha_1)} \\
&\leq \frac{\mathbb{E}[(|s_1 x_1| + \mathbb{E}[|Y|^p]^{1/p})^p \mathbb{1}_{x_1 \leq \alpha_1}]}{P(x_1 \leq \alpha_1)} \\
&\leq \frac{\mathbb{E}[(|s_1 \alpha_1| + \sqrt{2}\|s_{2:d}\|(\Gamma((1 + p)/2)/\sqrt{\pi})^{1/p})^p \mathbb{1}_{x_1 \leq \alpha_1}]}{P(x_1 \leq \alpha_1)} \\
&= \left(|s_1 \alpha_1| + \sqrt{2}\|s_{2:d}\|(\Gamma((1 + p)/2)/\sqrt{\pi})^{1/p}\right)^p \\
&\leq 2^p \Gamma((1 + p)/2)/\sqrt{\pi}.
\end{aligned}$$

The rest of the proof follows by decomposing union events into separate events as before.

$$\begin{aligned}
\mathbb{E}[|\langle X, s\rangle|^p | \mathcal{E}^c] &= \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}^c}]}{P(\mathcal{E}^c)} \leq \frac{\mathbb{E}[|\langle X, s\rangle|^p \sum_i \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}^c)} \\
&= \sum_i \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}^c)} \leq \sum_i \frac{\mathbb{E}[|\langle X, s\rangle|^p \mathbb{1}_{\mathcal{E}_i^c}]}{P(\mathcal{E}_i^c)} \\
&\leq k 2^p \Gamma((1 + p)/2)/\sqrt{\pi}.
\end{aligned}$$

□

Proposition C.1 *Let X be a random d-dimensional vector, and A be an event defined in the same probability space with $p = P(X \in A) > 0$. Let random variable $Y = X|A$,* i.e., *X conditioned on event A, and $Z = \mathbb{1}_{X \in A}$. Let X_i, Y_i, Z_i be the i.i.d. samples from corresponding distributions. Then, equation (3) holds for any $0 \le n_e \le n$ and $t_1 + t_2 = t$.*

Proof. We are interested in bounding the following probability

$$P\left(\| \sum_i \left(X_i \mathbb{1}_A - \mathbb{E}[X_i \mathbb{1}_A] \right) \|_2 \ge nt \right).$$

We will upper bound this probability by spliting it with conditioning on every possible set of Bernoulli variables Z_i, then arrange them.

$$P\left(\left\| \sum_i \left(X_i \mathbb{1}_A - \mathbb{E}[X_i \mathbb{1}_A] \right) \right\| \ge nt \right) = \sum_{\{Z_i\}_1^n \in \{0,1\}^n} P\left(\| \sum_i (X_i Z_i - \mathbb{E}[X_i \mathbb{1}_A]) \| \ge nt \middle| \{Z_i\}_1^n \right) P(\{Z_i\}_1^n).$$

Note that $X_i Z_i = 0$ when $Z_i = 0$, and $X_i Z_i = X_i|A = Y_i$ when $Z_i = 1$. Now we divide the cases into when $\sum_i Z_i \le n_e$ and $\sum_i Z_i > n_e$.

$$\begin{aligned}
&\sum_{\{Z_i\}_1^n \in \{0,1\}^n} P\left(\|(\sum_{i:Z_i=1} X_i) - n\mathbb{E}[X_i|A]P(A)\| \ge nt \middle| \{Z_i\}_1^n \right) P(\{Z_i\}_1^n) \\
&\le \sum_{\{Z_i\}_1^n \in \{0,1\}^n, \sum_i Z_i \le n_e} P\left(\|(\sum_{i:Z_i=1} X_i) - n\mathbb{E}[X|A]P(A)\| \ge nt \middle| \{Z_i\}_1^n \right) P(\{Z_i\}_1^n) + P\left(\sum_i Z_i \ge n_e + 1 \right).
\end{aligned}$$

We can decouple the first term above into two terms as the following:

$$\begin{aligned}
&P\left(\|(\sum_{i:Z_i=1} X_i) - n\mathbb{E}[X|A]P(A)\| \ge nt \middle| \{Z_i\}_1^n \right) \\
&= P\left(\| \sum_{i:Z_i=1} (X_i - \mathbb{E}[X|A]) + \mathbb{E}[X|A](\sum_i Z_i - nP(A)) \| \ge nt \middle| \{Z_i\}_1^n \right) \\
&\le P\left(\| \sum_{i:Z_i=1} (X_i - \mathbb{E}[X|A]) \| \ge nt_1 \middle| \{Z_i\}_1^n \right) + P\left(\| \mathbb{E}[X|A](\sum_i Z_i - nP(A)) \| \ge nt_2 \middle| \{Z_i\}_1^n \right).
\end{aligned}$$

where $t_1 + t_2 = t$. Then we observe that conditioned on $Z_i = 1$, X_i is actually Y_i, and we can discard all X_i for i such that $Z_i = 0$. Thus, the first expression is simplified to

$$P\left(\| \sum_{i:Z_i=1} (X_i - \mathbb{E}[X|A]) \| \ge nt_1 \middle| \{Z_i\}_1^n, \sum_i Z_i = m \right) = P\left(\left\| \sum_{j=1}^m (Y_j - \mathbb{E}[Y]) \right\| \ge nt_1 \right), \tag{24}$$

Here, j is a new index variable, and now the condition is only on the sum of Z_i, which is m. Now we are ready to wrap up the result:

$$
\begin{aligned}
&P\left(\left\|\sum_i \left(X_i \mathbb{1}_A - \mathbb{E}[X_i \mathbb{1}_A]\right)\right\| \geq nt\right) \\
&\leq \sum_{\{Z_i\}_1^n \in \{0,1\}^n, \sum_i Z_i \leq n_e} P\left(\left\|\sum_{j=1}^m (Y_j - \mathbb{E}[Y])\right\| \geq nt_1\right) P(\{Z_i\}_1^n, \sum_i Z_i = m) \\
&+ \sum_{\{Z_i\}_1^n \in \{0,1\}^n, \sum_i Z_i \leq n_e} P\left(\|\mathbb{E}[Y]\| \left|\sum_i Z_i - nP(A)\right| \geq nt_2 \middle| \{Z_i\}_1^n\right) P(\{Z_i\}_1^n) \\
&+ P\left(\sum_i Z_i \geq n_e + 1\right) \\
&\leq \max_{m \leq n_e} P\left(\left\|\sum_{j=1}^m (Y_j - \mathbb{E}[Y])\right\| \geq nt_1\right) \\
&+ P\left(\|\mathbb{E}[Y]\| \left|\sum_i Z_i - nP(A)\right| \geq nt_2\right) + P\left(\sum_i Z_i \geq n_e + 1\right),
\end{aligned}
$$

where the last inequality we used the fact $\sum_{\{Z_i\}_1^n \in \{0,1\}^n} P(\{Z_i\}_1^n) = 1$, and (24) is only conditioned on the sum of Z_i being less than n_e. We divide by n in conditions inside the first two probabilities, and we get the theorem. □

Appendix D Defered Proof: Bounding B for population EM when $D_m \leq 1$

Case II. $\max_j \|\Delta_j\| \leq 1$**:** We use mean-value theorem to reformulate Δ_w. We additionally define a symbol $\delta_j := \pi_j - \pi_j^*$. Denote $\beta_j^u = \beta_j^* + u\Delta_j$ and $\pi_j^u = \pi_j^* + u\delta_j$ for $u \in [0,1]$, and let w_1^u be the weight in E-step constructed with β_j^u and π_j^u. Then, by mean-value theorem, for some $u \in [0,1]$, $B = \|\mathbb{E}_{\mathcal{D}}[\Delta_w^u \langle X, s\rangle (Y - \langle X, \beta_1^*\rangle)]\|$ where

$$
\begin{aligned}
\Delta_w^u = &\underbrace{-w_1^u(1-w_1^u)(\langle X, \beta_1^u\rangle - Y)\langle X, \Delta_1\rangle + \sum_{l\neq 1} w_1^u w_l^u (\langle X, \beta_l^u\rangle - Y)\langle X, \Delta_l\rangle}_{\Delta_{w,1}} \\
&\underbrace{-w_1^u(1-w_1^u)\delta_1/\pi_1^u + \sum_{l\neq 1} w_1^u w_l^u \delta_l/\pi_l^u}_{\Delta_{w,2}},
\end{aligned}
$$

for some $u \in [0,1]$. Note that $\delta_j/\pi_j^u \leq 2\delta_j/\pi_j^* \leq 1$ guaranteed by initialization condition and the result for $D_m \geq 1$. Let us now redefine $D_m = \max(\max_j \|\Delta_j\|, \max_j \delta_j/\pi_j^*)$. Then for each j, we can decompose the target term as

$$
|\mathbb{E}_{\mathcal{D}_j}[\Delta_w^u \langle X, s\rangle (Y - \langle X, \beta_1^*\rangle)]| \leq \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,1}\langle X, s\rangle (Y - \langle X, \beta_1^*\rangle)]|}_{E_1} + \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,2}\langle X, s\rangle (Y - \langle X, \beta_1^*\rangle)]|}_{E_2}.
$$

We will bound E_1 and E_2 separately as we proceed.

$j \neq 1$**:**

Bounding E1. We first consider bounding the first term.

$$
\begin{aligned}
E_1 &= |\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,1}\langle X, s\rangle (Y - \langle X, \beta_1^*\rangle)]| \\
&\leq \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,1}\langle X, s\rangle \langle X, \beta_j^* - \beta_1^*\rangle]|}_{b_1} + \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,1}\langle X, s\rangle e]|}_{b_2},
\end{aligned}
$$

As before, we will first bound b_1. It is a bit complicated as it involves many algebraic terms, but the idea is the same.

$$b_1 \le \underbrace{\left|\mathbb{E}_{\mathcal{D}_j}[w_1^u(\langle X, \beta_j^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle]\right|}_{d_1} + \underbrace{\left|\mathbb{E}_{\mathcal{D}_j}\left[\sum_{l=1}^k w_1^u w_l^u(\langle X, \beta_j^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right]\right|}_{d_2}.$$

We bound d_2 first. Consider the following good events:

$$\mathcal{E}_1 = \{|\langle X, \Delta_l\rangle| \le D_m\tau_j, \forall l\} \cap \{|e| \le \tau_j\},\ \mathcal{E}_2 = \{|\langle X, \beta_j^* - \beta_1^u\rangle| \ge 4\tau_j\}.$$

We will set $\tau_j = c_\tau\left(\sqrt{\log(R_{j1}^* k\rho_\pi)}\right)$ for some large constant $c_\tau > 0$.

Under event $\mathcal{E}_1$, when $l \ne j$, we claim $|w_l^u(\langle X, \beta_j^* - \beta_l^u\rangle + e)| \le \rho_{jl}|\exp(-6\tau_j^2)4\tau_j| + w_l^u 4\tau_j$. Let us denote $r := (\langle X, \beta_j^* - \beta_l^u\rangle + e)$. Then

$$\begin{aligned} w_l^u &\le \rho_{jl}\exp\left(\frac{-(\langle X, \beta_j^* - \beta_l^u\rangle + e)^2 + (\langle X, \beta_j^* - \beta_j^u\rangle + e)^2}{2}\right) \\ &= \rho_{jl}\exp\left((\langle X, \beta_j^* - \beta_j^u\rangle + e)^2/2\right)\exp\left(-(\langle X, \beta_j^* - \beta_l^u\rangle + e)^2/2\right) \\ &= \rho_{jl}\exp\left(2\tau_j^2\right)\exp\left(-r^2/2\right). \end{aligned}$$

Thus $|w_l^u r| \le \rho_{jl}\exp(2\tau_j^2) r\exp(-r^2/2)$. The function $f(r) = r\exp(-r^2/2)$ is maximized when $r = 1$, and decreasing afterward. Therefore, we can conclude that whenever $r > 4\tau_j$,

$$|w_l^u r| \le \rho_{jl}\exp\left(2\tau_j^2\right)\sup_{r\ge 4\tau_j} r\exp\left(-r^2/2\right) \le \rho_{jl}\exp(2\tau_j^2)4\tau_j\exp(-8\tau_j^2) \le \rho_{jl}4\tau_j\exp(-6\tau_j^2).$$

When $4\tau_j > r$, we have $|w_l^u r| \le w_l^u 4\tau_j$. Thus, we have $|w_l^u r| \le 4\rho_{jl}\tau_j\exp(-6\tau_j^2) + |w_l^u 4\tau_j|$.

For $l = j$, under event $\mathcal{E}_1$, we know $|\langle X, \Delta_j\rangle + e| \le 2\tau_j$. Thus, it is also true for $j = l$ that $|w_l^u r| \le (4\tau_j\exp(-6\tau_j^2) \vee |w_l^u|4\tau_j)$.

Now we plugging these relations into d_2, we get

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_j}\left[\sum_l w_1^u w_l^u(\langle X, \beta_j^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right] \\ &\le \rho_\pi\mathbb{E}_{\mathcal{D}_j}\left[\sum_l \left|w_1^u\exp(-6\tau_j^2)4\tau_j\langle X, \Delta_l\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right] \\ &+ \mathbb{E}_{\mathcal{D}_j}\left[\sum_l \left|w_1^u w_l^u 4\tau_j\langle X, \Delta_l\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right] \\ &+ \mathbb{E}_{\mathcal{D}_j}\left[\left|\sum_l w_1^u w_l^u(\langle X, \beta_j^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1^c}\right] \\ &\le 4\rho_\pi D_m\tau_j^2\exp(-6\tau_j^2)\underbrace{\mathbb{E}_{\mathcal{D}_j}\left[\sum_l \left|w_1^u\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right]}_{(i)} \\ &+ 4D_m\tau_j^2\underbrace{\mathbb{E}_{\mathcal{D}_j}\left[\left|(\sum_l w_l^u)w_1^u\langle X, s\rangle\langle X, \beta_j^* - \beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right]}_{(ii)} \end{aligned}$$

$$+\underbrace{\mathbb{E}_{\mathcal{D}_j}\left[\left|\sum_l w_1^u w_l^u(\langle X,\beta_j^*-\beta_l^u\rangle+e)\langle X,\Delta_l\rangle\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\mathbb{1}_{\mathcal{E}_1^c}\right|\right]}_{(iii)}.$$

For (i),

$$\mathbb{E}_{\mathcal{D}_j}\left[\sum_l \left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right] \leq \sum_l \sqrt{\mathbb{E}_{\mathcal{D}_j}[\langle X,\beta_j^*-\beta_1^*\rangle^2]}\sqrt{\mathbb{E}_{\mathcal{D}_j}[\langle X,s\rangle^2]}=\sum_l R_{j1}^*=kR_{j1}^*.$$

For (ii),

$$\mathbb{E}_{\mathcal{D}_j}\left[\left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1}\right]=\mathbb{E}_{\mathcal{D}_j}\left[\left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] + \mathbb{E}_{\mathcal{D}_j}\left[\left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right].$$

Under event $\mathcal{E}_1\cap\mathcal{E}_2$, it is easy to see that

$$|\langle X,\beta_j^*-\beta_j^u\rangle+e|\leq 2\tau_j,\ |\langle X,\beta_j^*-\beta_1^u\rangle+e|\geq 3\tau_j,\ w_1^u\leq\rho_{j1}\exp(-2\tau_j^2),$$

thus

$$\mathbb{E}_{\mathcal{D}_j}\left[\left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right]\leq\rho_{j1}\exp(-2\tau_j^2)R_{j1}^*.$$

For the second term:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_j}\left[\left|w_1^u\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\right|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right] &\leq \mathbb{E}[|\langle X,s\rangle||\langle X,\beta_j^*-\beta_1^u\rangle+u\langle X,\Delta_1\rangle|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}]\\ &\leq \mathbb{E}[|\langle X,s\rangle|(5\tau_j)\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}]\\ &\leq 5\tau_j\mathbb{E}[|\langle X,s\rangle||\mathcal{E}_2^c]P(\mathcal{E}_2^c)\\ &\leq c_1\tau_j^2/R_{j1}^*.\end{aligned}$$

For (iii), note that $P(\mathcal{E}_1^c)\leq 2k\exp(-\tau_j^2/2)$. Then,

$$\begin{aligned}\text{(iii)} &\leq \mathbb{E}_{\mathcal{D}_j}\left[\left|\sum_l w_1^u w_l^u(\langle X,\beta_j^*-\beta_l^u\rangle+e)\langle X,\Delta_l\rangle\langle X,s\rangle\langle X,\beta_j^*-\beta_1^*\rangle\mathbb{1}_{\mathcal{E}_1^c}\right|\right]\\ &\leq \sum_l\sqrt{\mathbb{E}_{\mathcal{D}_j}[{w_l^u}^2\langle X,\beta_j^*-\beta_l^u\rangle+e)^2\langle X,\Delta_l\rangle^2}\sqrt{\mathbb{E}_{\mathcal{D}_j}\left[\langle X,s\rangle^2\langle X,\beta_j^*-\beta_1^*\rangle^2\mathbb{1}_{\mathcal{E}_1^c}\right]}\\ &\leq \sum_l\sqrt{\mathbb{E}_{\mathcal{D}_j}[(w_l^u)^2(\langle X,\beta_j^*-\beta_l^u\rangle+e)^2\langle X,\Delta_l\rangle^2}\sqrt[8]{\mathbb{E}_{\mathcal{D}_j}[\langle X,s\rangle^8]}\sqrt[8]{\mathbb{E}_{\mathcal{D}_j}\left[\langle X,\beta_j^*-\beta_1^*\rangle^8\right]}\sqrt[4]{P(\mathcal{E}_1^c)}\\ &\leq cR_{j1}^*\sqrt[4]{k}\exp(-\tau_j^2/8)\left(\sum_l\sqrt{\mathbb{E}_{\mathcal{D}_j}[(w_l^u)^2(\langle X,\beta_j^*-\beta_l^u\rangle+e)^2\langle X,\Delta_l\rangle^2}\right).\end{aligned}\tag{25}$$

In order to bound (25), we need the following equation which we defer to prove in D:

Lemma D.1 *If $D_m\leq 1$, for $j\neq l$,*

$$\mathbb{E}_{\mathcal{D}_j}\left[(w_l^u)^2\langle X,(\beta_j^*-\beta_l^u+e)\rangle^2\langle X,\Delta_l\rangle^2\right]\leq O\left((\rho_{jl}R_{jl}^*)^2\exp(-\tau_l^2/2)+\tau_l^3/R_{jl}^*\right)\|\Delta_l\|^2,\tag{26}$$

which is less than $O(\|\Delta_l\|^2)$ with $\tau_l=O(\sqrt{\log(R_{jl}^*\rho_\pi)})$.

If $j=l$, we have

$$\mathbb{E}_{\mathcal{D}_j}\left[(w_j^u)^2\langle X,(\beta_j^*-\beta_j^u+e)\rangle^2\langle X,\Delta_j\rangle^2\right]\leq O\left(\tau_j^2+(\|\Delta_j\|^2+1)\sqrt{k}\exp(-\tau_j^2/4)\right)\|\Delta_j\|^2,\tag{27}$$

which is less than $O(\|\Delta_j\|^2\log k)$ with $\tau_j=O(\sqrt{\log k})$.

Then, we can bound (25) by

$$
\begin{aligned}
(25) &\leq O\left(R_{j1}^* \sqrt[4]{k} \exp(-\tau_j^2/8) (\sum_{l \neq j} D_m + \sqrt{\log k} D_m) \right) \\
&\leq O\left(R_{j1}^* k^{5/4} \exp(-\tau_j^2/8) D_m \right).
\end{aligned}
$$

Combining all results, we have

$$
d_2 \leq O\left((\rho_\pi k \tau_j^2 + k^{5/4}) \exp(-\tau_j^2/8) R_{j1}^* + \tau_j^4 / R_{j1}^* \right) D_m.
$$

Then, we set $\tau_j = \Theta\left(\sqrt{\log(R_{j1}^* k \rho_\pi)}\right)$ to get $d_2 \leq c_d D_m/(k\rho_\pi)$ along with $R_{j1}^* \geq R_{min} \geq \tilde{\Omega}(k\rho_\pi)$.

Now for d_1, we follow the exactly same path, while the only difference is that it does not involve summation over all components.

$$
\begin{aligned}
d_1 &= \mathbb{E}_{\mathcal{D}_j} \left[w_1^u (\langle X, \beta_j^* - \beta_1^u \rangle + e) \langle X, \Delta_1 \rangle \langle X, s \rangle \langle X, \beta_j^* - \beta_1^* \rangle \right] \\
&\leq \rho_\pi \exp(-6\tau_j^2) 4\tau_j \left[|\langle X, \Delta_1 \rangle \langle X, s \rangle \langle X, \beta_j^* - \beta_1^* \rangle| \mathbb{1}_{\mathcal{E}_1} \right] \\
&+ 4\tau_j \mathbb{E}_{\mathcal{D}_j} \left[|w_1^u \langle X, \Delta_1 \rangle \langle X, s \rangle \langle X, \beta_j^* - \beta_1^* \rangle| \mathbb{1}_{\mathcal{E}_1} \right] \\
&+ \mathbb{E} \left[|w_1^u (\langle X, \beta_j^* - \beta_1^u \rangle + e) \langle X, \Delta_1 \rangle \langle X, s \rangle \langle X, \beta_j^* - \beta_1^* \rangle| \mathbb{1}_{\mathcal{E}_1^c} \right] \\
&\leq O\left((\rho_\pi \tau_j^2 + \sqrt[4]{k}) \exp(-\tau_j^2/8) R_{j1}^* + \tau_j^4 / R_{j1}^* \right) D_m,
\end{aligned}
$$

where we can set τ_j the same, and we get $d_1 \leq c_{d'} D_m/(k\rho_\pi)$. Therefore we complete the proof for $b_1 \leq c_b D_m/(k\rho_\pi)$.

The bound for b_2 is replicate of the proof for b_1 except that, at the end of inequality we get $\sqrt{\mathbb{E}[\langle X, s \rangle^2 e^2]}$ instead of $\sqrt{\mathbb{E}[\langle X, s \rangle^2 \langle X, \beta_j^* - \beta_1^* \rangle^2]}$. Specifically, we start from

$$
\begin{aligned}
b_2 &\leq \underbrace{|\mathbb{E}_{\mathcal{D}_j}[w_1^u (\langle X, \beta_j^* - \beta_1^u \rangle + e) \langle X, \Delta_1 \rangle \langle X, s \rangle e]|}_{d_1} \\
&+ \underbrace{|\mathbb{E}_{\mathcal{D}_j}[\sum_{l=1}^k w_1^u w_l^u (\langle X, \beta_j^* - \beta_l^u \rangle + e) \langle X, \Delta_l \rangle \langle X, s \rangle e]|}_{d_2}.
\end{aligned}
$$

For d_2, applying the same argument, we get

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_j} \left[\sum_l w_1^u w_l^u (\langle X, \beta_j^* - \beta_l^u \rangle + e) \langle X, \Delta_l \rangle \langle X, s \rangle e \right] \\
&\leq 4\rho_\pi D_m \tau_j^2 \exp(-6\tau_j^2) \underbrace{\mathbb{E}_{\mathcal{D}_j} \left[\sum_l |w_1^u \langle X, s \rangle e| \mathbb{1}_{\mathcal{E}_1} \right]}_{(i)} + 4 D_m \tau_j^2 \underbrace{\mathbb{E}_{\mathcal{D}_j} \left[\left| (\sum_l w_l^u) w_1^u \langle X, s \rangle e \right| \mathbb{1}_{\mathcal{E}_1} \right]}_{(ii)} \\
&+ \underbrace{\mathbb{E}_{\mathcal{D}_j} \left[\left| \sum_l w_1^u w_l^u (\langle X, \beta_j^* - \beta_l^u \rangle + e) \langle X, \Delta_l \rangle \langle X, s \rangle e \mathbb{1}_{\mathcal{E}_1^c} \right| \right]}_{(iii)}.
\end{aligned}
$$

Then, we can go through exactly same path to bound each (i), (ii), (iii). Finally, set $\tau_j = \Theta\left(\sqrt{\log(R_{j1}^* k \rho_\pi)}\right)$ as before and we get the bound $E_1 \leq c_b D_m/(k\rho_\pi)$ for $j \neq 1$.

Bounding E_2, the term from mismatch in mixing weights. Recall that

$$
\begin{aligned}
E_2 &= |\mathbb{E}_{\mathcal{D}_j}[\Delta_{w,2}\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)]|, \\
\Delta_{w,2} &= -w_1^u(1-w_1^u)\delta_1/\pi_1^u + \sum_{l\neq 1} w_1^u w_l^u \delta_l/\pi_l^u \\
&\le \left| w_1^u(1-w_1^u) + \sum_{l\neq 1} w_1^u w_l^u \right| 2D_m = 2w_1^u D_m.
\end{aligned}
$$

Hence, $E_2 \le 2D_m \mathbb{E}_{\mathcal{D}_j}[w_1^u|\langle X, s\rangle(Y - \langle X, \beta_1^*\rangle)|]$. We have already seen similar equation when we handle $D_m \ge 1$. Only difference is that Δ_w is now changed to w_1^u, but we can observe that we can reuse the exactly same procedure. (Remember the only property we needed for Δ_w was that it to be less than $\exp(\cdot)$ under good events, which is also true for w_1^u). Following the procedure to derive equation (9) and (10), E_2 can be bounded by

$$
O\left(\exp(-\tau_j^2/4)(1 \vee \rho_{j1})R_{j1}^* + \tau_j^2/R_{j1}^* + D_m/R_{j1}^*\right) D_m,
$$

which the same choice of parameters $\tau_j = \Theta(\sqrt{\log(R_{j1}^* k\rho_\pi)}$ gives $E_2 \le c_b D_m/(k\rho_\pi)$ with the same SNR condition.

$j = 1$**:**

Bounding E1. We define events

$$
\begin{aligned}
\mathcal{E}_1 &= \{|\langle X, \Delta_j\rangle| \le D_m\tau, \forall j\} \cap \{|e| \le \tau\} \\
\mathcal{E}_2 &= \{|\langle X, \beta_1^* - \beta_j^u\rangle| \ge 4\tau, \forall j \neq 1\}.
\end{aligned}
$$

For bounding E_1, same as when $D_m \ge 1$, $b_1 = 0$. Thus, we consider b_2 only, which is

$$
\begin{aligned}
b_2 &= |\mathbb{E}_{\mathcal{D}_1}[\Delta_w\langle X, s\rangle e]| \\
&\le \Bigg| \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[w_1^u(1-w_1^u)(\langle X, \beta_1^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle e\right]}_{d_1} \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[\sum_{l\neq 1} w_1^u w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e\right]}_{d_2} \Bigg|.
\end{aligned}
$$

First part of the proof follows the path for $j \neq 1$.

$$
\begin{aligned}
d_2 &\le \mathbb{E}_{\mathcal{D}_1}\left[\left|\sum_{l\neq 1} w_1^u w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1}\right|\right] \\
&\le \mathbb{E}_{\mathcal{D}_1}\left[\sum_{l\neq 1} \left|w_1^u \rho_\pi \exp(-6\tau^2) 4\tau\langle X, \Delta_l\rangle\langle X, s\rangle e\right| \mathbb{1}_{\mathcal{E}_1}\right] + \mathbb{E}_{\mathcal{D}_1}\left[\sum_{l\neq 1} \left|w_1^u w_l^u 4\tau\langle X, \Delta_l\rangle\langle X, s\rangle e\right| \mathbb{1}_{\mathcal{E}_1}\right] \\
&\quad + \mathbb{E}_{\mathcal{D}_1}\left[\left|\sum_{l\neq 1} w_1^u w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e\right| \mathbb{1}_{\mathcal{E}_1^c}\right] \\
&\le 4\rho_\pi D_m\tau^2\exp(-6\tau^2) \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[\sum_{l} |w_1^u\langle X, s\rangle e| \mathbb{1}_{\mathcal{E}_1}\right]}_{(i)} + 4D_m\tau^2 \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[\left|(\sum_{l\neq 1} w_l^u) w_1^u\langle X, s\rangle e\right| \mathbb{1}_{\mathcal{E}_1}\right]}_{(ii)} \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[\left|\sum_{l\neq 1} w_1^u w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e \mathbb{1}_{\mathcal{E}_1^c}\right|\right]}_{(iii)}.
\end{aligned}
$$

For (i),

$$\mathbb{E}_{\mathcal{D}_1}\left[\sum_l |w_1^u \langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1}\right] \le \sum_k \mathbb{E}_{\mathcal{D}_1}[|e\langle X, s\rangle|] \le k.$$

(ii), we use event $\mathcal{E}_2$ as before,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_1}\left[|(1-w_1^u)w_1^u\langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1}\right] &= \mathbb{E}_{\mathcal{D}_1}\left[|(1-w_1^u)w_1^u\langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\ &+ \mathbb{E}_{\mathcal{D}_1}\left[|(1-w_1^u)w_1^u\langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right].\end{aligned}$$

Under event $\mathcal{E}_1\cap\mathcal{E}_2$, it is now easy to show that $w_l^u \le 3\rho_\pi \exp(-2\tau^2)$ for all $l \ne 1$. Thus, $1-w_1^u \le 3k\rho_\pi \exp(-2\tau^2)$, and

$$\mathbb{E}_{\mathcal{D}_1}\left[|(1-w_1^u)w_1^u\langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \le 3k\rho_\pi \exp(-2\tau^2)\mathbb{E}_{\mathcal{D}_1}[|\langle X, s\rangle e|] \le 3k\rho_\pi \exp(-2\tau^2).$$

For $\mathcal{E}_1\cap\mathcal{E}_2^c$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_1}\left[|(1-w_1^u)w_1^u\langle X, s\rangle e|\, \mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right] &\le \mathbb{E}_{\mathcal{D}_1}[|\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_2^c}] \\ &\le \sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X, s\rangle^2|\mathcal{E}_2^c]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[e^2|\mathcal{E}_2^c]}P(\mathcal{E}_2^c) \\ &\le c_1\sqrt{\log k}\frac{k\tau}{R_{min}}.\end{aligned}$$

For (iii),

$$\begin{aligned}(iii) &= \mathbb{E}_{\mathcal{D}_1}\left[\left|\sum_{l\ne1} w_1^u w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e\mathbb{1}_{\mathcal{E}_1^c}\right|\right] \\ &\le \sum_{l\ne1}\mathbb{E}_{\mathcal{D}_1}\left[\left|w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e\mathbb{1}_{\mathcal{E}_1^c}\right|\right],\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_1}\left[\left|w_l^u(\langle X, \beta_1^* - \beta_l^u\rangle + e)\langle X, \Delta_l\rangle\langle X, s\rangle e\mathbb{1}_{\mathcal{E}_1^c}\right|\right] \le& \sqrt{\mathbb{E}_{\mathcal{D}_1}[(w_l^u)^2(\langle X, \beta_1^* - \beta_l^u\rangle + e)^2\langle X, \Delta_l\rangle^2]} \\ & \sqrt[4]{\mathbb{E}_{\mathcal{D}_1}[\langle X, s\rangle^4 e^4]}\sqrt[4]{P(\mathcal{E}_1^c)},\end{aligned}$$

For bounding $\sqrt{\mathbb{E}_{\mathcal{D}_1}[(w_l^u)^2(\langle X, \beta_1^* - \beta_l^u\rangle + e)^2\langle X, \Delta_l\rangle^2]}$ for $l \ne 1$, we can again use Lemma D.1. We also have that $P(\mathcal{E}_1^c) \le k\exp(-\tau^2/2)$. Then,

$$(iii) \le c_2 k\sqrt[4]{k}\exp(-\tau^2/8)D_m.$$

Combining all,

$$d_2 \le O\left((k^{5/4} + k\tau^2)\exp(-\tau^2/8) + k\sqrt{\log k}\tau^3/R_{min}\right)D_m. \tag{28}$$

Along with our choice $\tau = \Theta(\sqrt{\log(k\rho_\pi)})$ and $R_{min} = \tilde{\Omega}(k)$, we get $d_2 \le c_d D_m$.

For bounding d_1, (all constants $c_1, c_2, ...$ are renewed)

$$\begin{aligned}d_1 &= \mathbb{E}_{\mathcal{D}_1}\left[w_1^u(1-w_1^u)(\langle X, \beta_1^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle e\right] \\ &\le \mathbb{E}_{\mathcal{D}_1}\left[|w_1^u(1-w_1^u)(\langle X, \beta_1^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\ &+ \mathbb{E}_{\mathcal{D}_1}\left[|w_1^u(1-w_1^u)(\langle X, \beta_1^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_1^c}\right] \\ &+ \mathbb{E}_{\mathcal{D}_1}\left[|w_1^u(1-w_1^u)(\langle X, \beta_1^* - \beta_1^u\rangle + e)\langle X, \Delta_1\rangle\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right] \\ &\le k\rho_\pi\exp(-\tau^2/2)\underbrace{\mathbb{E}_{\mathcal{D}_1}\left[|(|u\langle X, \Delta_1\rangle| + |e|)\langle X, \Delta_1\rangle\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right]}_{(i)} \\ &+ \underbrace{\mathbb{E}_{\mathcal{D}_1}\left[(|u\langle X, \Delta_1\rangle| + |e|)\langle X, \Delta_1\rangle\langle X, s\rangle e|\mathbb{1}_{\mathcal{E}_1^c}\right]}_{(ii)}\end{aligned}$$

$$+\underbrace{\mathbb{E}_{\mathcal{D}_1}\left[(|u\langle X,\Delta_1\rangle|+|e|)\langle X,\Delta_1\rangle\langle X,s\rangle e|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2^c}\right]}_{(iii)}.$$

$$\begin{aligned}(i) &= \mathbb{E}_{\mathcal{D}_1}\left[|(|u\langle X,\Delta_1\rangle|+|e|)\langle X,\Delta_1\rangle\langle X,s\rangle e|\mathbb{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\ &\le \mathbb{E}_{\mathcal{D}_1}\left[\langle X,\Delta_1\rangle^2|\langle X,s\rangle e|\right]+\mathbb{E}_{\mathcal{D}_1}|\left[\langle X,\Delta_1\rangle\langle X,s\rangle e^2|\right] \\ &\le c_1\|\Delta_1\|(1+\|\Delta_1\|)\le 2c_1D_m.\end{aligned}$$

$$\begin{aligned}(ii) &\le \mathbb{E}_{\mathcal{D}_1}\left[\langle X,\Delta_1\rangle^2|\langle X,s\rangle e|\mathbb{1}_{\mathcal{E}_1^c}\right]+\mathbb{E}_{\mathcal{D}_1}\left[|\langle X,\Delta_1\rangle\langle X,s\rangle e^2|\mathbb{1}_{\mathcal{E}_1^c}\right] \\ &= \sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X,\Delta_1\rangle^4\langle X,s\rangle^2e^2]}\sqrt{P(\mathcal{E}_1^c)}+\sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X,\Delta_1\rangle^2\langle X,s\rangle^2e^4]}\sqrt{P(\mathcal{E}_1^c)} \\ &\le c_2\sqrt{k}\|\Delta_1\|\exp(-\tau^2/4).\end{aligned}$$

$$\begin{aligned}(iii) &\le D_m^2\tau^2\sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X,s\rangle^2|\mathcal{E}_2^c]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[e^2|\mathcal{E}_2^c]}P(\mathcal{E}_2^c) \\ &\quad + D_m\tau\sqrt{\mathbb{E}_{\mathcal{D}_1}[\langle X,s\rangle^2|\mathcal{E}_2^c]}\sqrt{\mathbb{E}_{\mathcal{D}_1}[e^4|\mathcal{E}_2^c]}P(\mathcal{E}_2^c) \\ &\le c_3\sqrt{\log k}D_m\frac{k\tau^3}{R_{min}},\end{aligned}$$

where we applied Corollary 4.2. (i), (ii), (iii) gives a bound for d_1 as

$$d_1\le O\left(k\rho_\pi\exp(-\tau^2/4)+k\sqrt{\log k}\tau^3/R_{min}\right)D_m. \tag{29}$$

Now combining (28) and (29) we get the bound for $E_1\le c_eD_m$, with the choice of $\tau=\Theta(\sqrt{\log(k\rho_\pi)})$ and high SNR $\Omega\tilde{(}k)$.

Bounding E_2, the term from mismatch in mixing weights. When $j=1$,

$$\Delta_{w,2}=-w_1^u(1-w_1^u)\delta_1/\pi_1^u+\sum_{l\neq 1}w_1^uw_l^u\delta_l/\pi_l^u\le\left|w_1^u(1-w_1^u)+\sum_{l\neq 1}w_1^uw_l^u\right|D_m=2w_1^u(1-w_1^u)D_m.$$

Hence, $E_2\le D_m\mathbb{E}_{\mathcal{D}_j}[(1-w_1^u)|\langle X,s\rangle(Y-\langle X,\beta_1^*\rangle)|]$. Again, we have already seen similar equation when we handle $D_m\ge 1$. Following the procedure to derive equation (11), E_2 can be bounded by

$$O\left(k\rho_\pi\exp(-\tau^2/4)+(k\sqrt{\log k})\tau/R_{min}+(k\sqrt{\log k})D_m/R_{min}\right)D_m,$$

which the same choice of parameters $\tau=\Theta(\sqrt{\log(k\rho_\pi)})$ gives $E_2\le c_bD_m$ with the SNR condition $\tilde{\Omega}(k)$.

Summing up everything, for $j\neq 1$ we have $B_j\le O(D_m/(k\rho_\pi))$, and for $j=1$ we have $B_1\le O(D_m)$. Thus, $B\le\pi_1^*B_1+\sum_{j\neq 1}\pi_j^*B_j\le c_BD_m\pi_1^*$ for some constant $c_B\in(0,1/8)$ by properly setting constants in the proof. That is $\|\beta_1^+-\beta_1^*\|\le c_BD_m\pi_1^*$.

Update for mixing weights. The procedure is exactly same for proving the bound for $\|\beta_1^+-\beta_1^*\|$. It is actually easier since it does not involve additional terms $\langle X,s\rangle$ and $Y-\langle X,\beta_1^*\rangle$ as can be seen in (12). Thus we can follow the exact same procedure, getting $|\pi_1^+-\pi_1^*|/\pi_1^*\le c_BD_m$.

Proof of Lemma D.1

Proof. If $j \neq l$, we define a new event with new parameter τ_l,

$$\mathcal{E}_{1,l} = \{|\langle X, \Delta_l \rangle| \le D_m \tau_l\} \cap \{|e| \le \tau_l\}$$
$$\mathcal{E}_{2,l} = \{|\langle X, \beta_j^* - \beta_l^u \rangle| \ge 4\tau_l\}.$$

Under event $\mathcal{E}_{1,l}$, we can show that

$$|w_l^u|^2 \langle X, (\beta_j^* - \beta_l^u + e)\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}} \le (\rho_{jl} \exp(-6\tau_l^2) 4\tau_l)^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}} + (w_l^u 4\tau_l)^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}^c}.$$

Now we can bound (26) as,

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_j}\left[(w_l^u)^2 \langle X, (\beta_j^* - \beta_l^u + e)\rangle^2 \langle X, \Delta_l\rangle^2\right] \\ &\quad\le \mathbb{E}_{\mathcal{D}_j}\left[16\rho_{jl} \exp(-12\tau_l^2)\tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}}\right] \\ &\quad+ \mathbb{E}_{\mathcal{D}_j}\left[16 (w_l^u)^2 \tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}^c}\right] \\ &\quad+ \mathbb{E}_{\mathcal{D}_j}\left[(w_l^u)^2 \langle X, (\beta_j^* - \beta_l^u + e)\rangle^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right]. \end{aligned}$$

We do similarly bound each term:

$$\mathbb{E}_{\mathcal{D}_j}\left[16 \exp(-12\tau_l^2)\tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}}\right] \le c_1 \rho_{jl} \exp(-12\tau_l^2)\tau_l^2 \|\Delta_l\|^2,$$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_j}\left[16 (w_l^u)^2 \tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l} \cap \mathcal{E}_{2,l}^c}\right] &\le 16\tau_l^2 \mathbb{E}_{\mathcal{D}_j}\left[(w_l^u)^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{2,l}^c}\right] \\ &\le 16\tau_l^2 \mathbb{E}_{\mathcal{D}_j}\left[\langle X, \Delta_l\rangle^2 | \mathcal{E}_{2,l}^c\right] P(\mathcal{E}_{2,l}^c) \\ &\le c_2 \tau_l^2 \|\Delta_l\|^2 \tau_l / R_{jl}^*, \end{aligned}$$

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_j}\left[[(w_l^u)^2 \langle X, (\beta_j^* - \beta_l^u + e)\rangle^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right] \\ &\quad\le \mathbb{E}_{\mathcal{D}_j}\left[2\langle X, \beta_j^* - \beta_l^u\rangle^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right] + \mathbb{E}_{\mathcal{D}_j}\left[2e^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right] \\ &\quad\le 2\sqrt{\mathbb{E}_{\mathcal{D}_j}\left[\langle X, \beta_j^* - \beta_l^u\rangle^4 \langle X, \Delta_l\rangle^4\right]} \sqrt{P(\mathcal{E}_{1,l}^c)} + 2\sqrt{\mathbb{E}_{\mathcal{D}_j}\left[e^4 \langle X, \Delta_l\rangle^4\right]} \sqrt{P(\mathcal{E}_{1,l}^c)} \\ &\quad\le c_3 (R_{jl}^*)^2 \|\Delta_l\|^2 \exp(-\tau_l^2/2) + c_4 \|\Delta_l\|^2 \exp(-\tau_l^2/2). \end{aligned}$$

Set $\tau_l = \Theta(\sqrt{\log(R_{jl}^* \rho_\pi)})$. Then every terms will be canceled out and we get

$$(26) \le O(\|\Delta_l\|^2).$$

If $l = j$, then

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_j}\left[(w_l^u)^2 \langle X, (\beta_j^* - \beta_l^u + e)\rangle^2 \langle X, \Delta_l\rangle^2\right] \\ &\quad\le \mathbb{E}_{\mathcal{D}_j}\left[4\tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}}\right] \\ &\quad+ \mathbb{E}_{\mathcal{D}_j}\left[(\langle X, \Delta_l\rangle + e)^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right]. \end{aligned}$$

Each term is easy to bound.

$$\mathbb{E}_{\mathcal{D}_j}\left[4\tau_l^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}}\right] \le O(\tau_l^2 D_m^2).$$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_j}\left[(\langle X, \Delta_l\rangle + e)^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right] &\le \mathbb{E}_{\mathcal{D}_j}\left[2\langle X, \Delta_l\rangle^4 + 2e^2 \langle X, \Delta_l\rangle^2 \mathbb{1}_{\mathcal{E}_{1,l}^c}\right] \\ &\le 2\sqrt{\mathbb{E}_{\mathcal{D}_j}\left[\langle X, \Delta_l\rangle^8\right]} \sqrt{P(\mathcal{E}_{1,l}^c)} + 2\sqrt{\mathbb{E}_{\mathcal{D}_j}\left[e^4 \langle X, \Delta_l\rangle^4\right]} \sqrt{P(\mathcal{E}_{1,l}^c)} \\ &\le O((\|\Delta_l\|^4 + \|\Delta_l\|^2)\sqrt{k} \exp(-\tau_l^2/4)). \end{aligned}$$

We set $\tau_l = O(\sqrt{\log k})$ and get

$$(26) \le O(\|\Delta_l\|^2 \log k).$$

$\square$