Supplement to "Structure-Preserving Embedding of Multi-layer Networks"

Anonymous Author(s)
Affiliation
Address
`email`

This supplement file contains the appendixes of the paper "Structure-Preserving Embedding of Multi-layer Networks". In Appendix A, we summarize the projected gradient descent algorithm developed in Section 2 of the paper. Appendix B contains the detailed cross-validation procedure in selecting the tuning parameter λ_n. Additional simulation studies are provided in Appendix C. In Appendix D, we provide an eigenvalue plot of the WAT dataset discussed in Section 4.2 of the paper. All technical proofs and necessary lemmas are included in Appendix E.

A Summary of the projected gradient descent algorithm

For easy of presentation, we denote the projection result from Step 1 to Step 3 discussed in Section 2.3 as $P_{\Omega_{\boldsymbol{\alpha}} \times \Omega_{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$. The developed projected gradient descent algorithm can be summarized in Algorithm 1.

Algorithm 1: Projected gradient descent (PGD)

Input : Adjacency tensor $\boldsymbol{\mathcal{A}}$, sparsity factor s_n, number of communities K, embedding dimension R, constraint parameter ξ, tuning parameter λ_n, learning rate η, number of iterations T.

Output : Estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, estimated vertex community memberships and community centers.

Initialize $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ and hence obtain $\boldsymbol{Z}^{(0)}$, $\boldsymbol{C}^{(0)}$ by $(1+\delta)$-approximation K-means algorithm. Set t=0.

while $t < T$ **do**

  $\tilde{\boldsymbol{\alpha}}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \eta \nabla_{\boldsymbol{\alpha}} \mathcal{L}_\lambda(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{\mathcal{A}})$, $\tilde{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \nabla_{\boldsymbol{\beta}} \mathcal{L}_\lambda(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{\mathcal{A}})$;

  $\left(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}\right) = P_{\Omega_{\boldsymbol{\alpha}} \times \Omega_{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\alpha}}^{(t+1)}, \tilde{\boldsymbol{\beta}}^{(t+1)})$;

  Apply $(1+\delta)$-approximation K-means algorithm to $\boldsymbol{\alpha}^{(t+1)}$ to obtain $\boldsymbol{Z}^{(t+1)}$ and $\boldsymbol{C}^{(t+1)}$.

  if $\frac{|\mathcal{L}_\lambda(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}; \boldsymbol{\mathcal{A}}) - \mathcal{L}_\lambda(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{\mathcal{A}})|}{\mathcal{L}_\lambda(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}; \boldsymbol{\mathcal{A}})} < 10^{-6}$ **then**

    break.

  end

  t = t+1.

end

B Selecting λ_n

In this appendix, we provide the detailed tuning procedure for selecting λ_n. Specifically, let $\Lambda = \{\lambda_{n1}, ..., \lambda_{nQ}\}$ be the set of Q candidates of λ_n, p_0 be the fraction of training data, and κ be the number of repetitions. For each repetition κ_0, we first sample the training data from the adjacency tensor $\boldsymbol{\mathcal{A}}$ such that $a_{i,j,m}$ will be sampled independently with probability p_0, for any $i \leq j$. Denote

Δ be the index set of the training data. For each candidate $\lambda_{nq} \in \Lambda$, we apply Algorithm 1 to solve for $(\boldsymbol{\alpha}^{\kappa_0,q}, \boldsymbol{\beta}^{\kappa_0,q}) \in \Omega_{\boldsymbol{\alpha}} \times \Omega_{\boldsymbol{\beta}}$ that minimizes

$$\frac{1}{|\Delta|} \sum_{(i,j,m)\in\Delta} L(\theta_{i,j,m}, a_{i,j,m}) + \lambda_{nq} J(\boldsymbol{\alpha}), \tag{1}$$

where $|\Delta|$ is the cardinality of Δ. We then evaluate the negative log-likelihood over the held-out set

$$l^{\kappa_0,q} = \frac{1}{|\Delta^c|} \sum_{(i,j,m)\in\Delta^c} L(\theta^{\kappa_0,q}_{i,j,m}, a_{i,j,m}),$$

where Δ^c is the complement of Δ and $\theta^{\kappa_0,q}_{i,j,m} = \mathcal{I} \times_1 (\boldsymbol{\alpha}^{\kappa_0,q}_{i,\cdot})^T \times_2 (\boldsymbol{\alpha}^{\kappa_0,q}_{j,\cdot})^T \times_3 (\boldsymbol{\beta}^{\kappa_0,q}_{m,\cdot})^T$. Finally, we select λ_n from Λ such that it minimizes the averaged held-out loss over κ repetitions; that is, $\lambda_n = \lambda_{nq^*}$ with

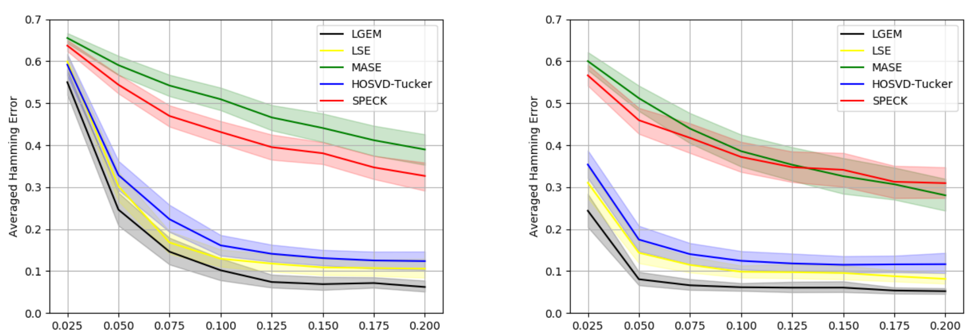
$$q^* = \arg\min_{q\in[Q]} \frac{1}{\kappa} \sum_{\kappa_0=1}^{\kappa} l^{\kappa_0,q}.$$

We remark that when solving (1), one needs to replace $\boldsymbol{\mathcal{T}}$ by $\boldsymbol{\mathcal{T}} * \boldsymbol{\mathcal{B}}$, to obtain the corresponding gradients associated with the training data in the PGD algorithm. Herein, $\boldsymbol{\mathcal{B}} \in \{0,1\}^{n\times n\times M}$ is the binary indicator tensor associated with Δ such that $\boldsymbol{\mathcal{B}}_{i,j,m} = 1$ if and only if $(i,j,m) \in \Delta$. Similarly, when estimating s_n inside the cross-validation process by equation (7) labeled in the paper, one need to replace the coefficient $\frac{1}{nM}$ by $\frac{1}{nMp_0}$ and $\boldsymbol{\mathcal{A}}$ by $\boldsymbol{\mathcal{A}} * \widetilde{\boldsymbol{\mathcal{B}}}$, where $\widetilde{\boldsymbol{\mathcal{B}}}$ is a symmetrization version of $\boldsymbol{\mathcal{B}}$ such that $\widetilde{\boldsymbol{\mathcal{B}}}_{i,j,m} = \widetilde{\boldsymbol{\mathcal{B}}}_{j,i,m} = \boldsymbol{\mathcal{B}}_{i,j,m}$, for $i \leq j$, $m \in [M]$.

C Additional simulation studies

As network gets sparser or community sizes gets more unbalanced, it becomes more difficult to differentiate vertices community memberships based on the observed multi-layer network. In this Appendix, we provide additional simulation studies of two scenarios. In Scenario I, we study the performances of TLSM and its competitors on networks with various sparsity, while in Scenario II, we study the performances of TLSM and its competitors on networks with various levels of unbalanced structures.

Scenario I: The multi-layer network generating process is the same as that descried in the paper, except that we vary $(n, s_n) \in \{200, 400\} \times \{0.025i : i \in [8]\}$ and fix $(M, K) = (5, 4)$. The averaged Hamming errors with 95% confidence intervals over 50 replications of all methods are plotted in Figure 1.



,

Figure 1: The averaged Hamming errors with 95% confidence intervals over 50 replications against various values of s_n in Scenario I with $n = 200$ (Left) and 400 (Right).

Scenario II: The multi-layer network generating process is the same as that descried in the paper, except that we generate $\psi \sim \text{Multi}(1, \boldsymbol{\pi})$ and vary $n \in \{200, 400\}$ while fixing $(M, K) = (5, 4)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4) = (0.25+p, 0.25+p, 0.25-p, 0.25-p)$ with $p \in \{1/24, 1/12, 1/8, 1/6\}$. The averaged Hamming errors with 95% confidence intervals over 50 replications of all methods are plotted in Figure 2.

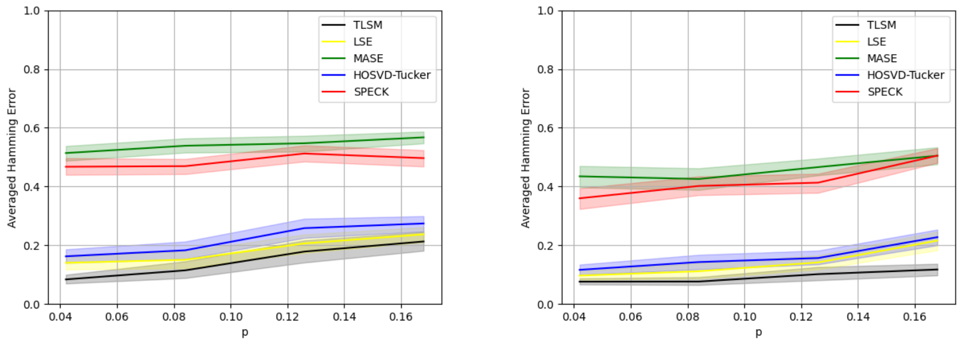


Figure 2: The averaged Hamming error with 95% confidence interval over 50 replications against various values of p in Scenario II with $n = 200$ (Left) and 400(Right).

It is evident that TLSM consistently outperforms the other competitors in both scenarios. In Scenario I, as s_n becomes larger, the averaged hamming errors of all methods decrease as expected, and TLSM and LSE perform the best even for relatively small s_n. In Scenario II, the averaged hamming errors of all methods increase gradually when the networks get more and more unbalanced, whereas TLSM appears to be more robust against the unbalancedness.

D Eigenvalue plot of the WAT dataset

In this appendix, we provide a leading singular value plot of the mode-1 matricization of the WAT dataset as in Figure 3. Note that the mode-1 matricization of a tensor is to unfold it into a matrix by stacking its mode-1 fibers as the columns of its matricization. It is clear from Figure 3 that the 7th leading singular value of the mode-1 matricization of the WAT dataset is an elbow point, which suggests there are 6 potential communities among the vertices. We hence set $K = 6$ in our analysis at Section 4.2. Such an eigen-gap investigation approach has been popularly employed to determine the number of communities for a network data in literature [1, 4] when it is unknown.

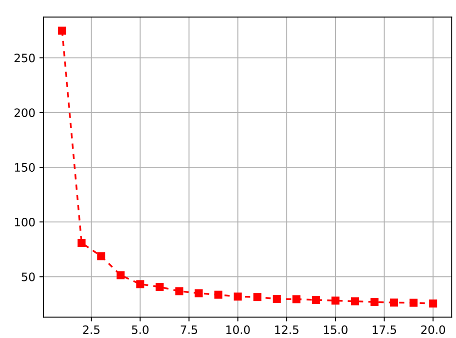


Figure 3: The first 20 leading singular values of the mode-1 matricization of the WAT dataset.

E Technical proofs

All Technical proofs and necessary lemmas are included in this appendix.

To begin with, we define the followings. For a Bernoulli random variable Y with expectation $p = s_n(1+\exp(-\theta))^{-1}$ and probability mass function $p(y;\theta)$, the discrete Hellinger distance between $p(y;\theta)$ and $p(y;\theta^*)$ is defined as

$$d(\theta,\theta^*) = \Big[\big(p^{1/2}-(p^*)^{1/2}\big)^2 + \big((1-p)^{1/2}-(1-p^*)^{1/2}\big)^2\Big]^{1/2},$$

and the deviation of $\mathbf{\Theta}$ from $\mathbf{\Theta}^*$ can be assessed by the averaged squared Hellinger distance,

$$D^2(\mathbf{\Theta},\mathbf{\Theta}^*) = \frac{1}{\varphi(n,M)}\sum_{m=1}^{M}\sum_{i\le j} d^2(\theta_{i,j,m},\theta^*_{i,j,m}).$$

In the proof of the main result, we will use the follow inequality several times.

Lemma 1. *Let $\boldsymbol{\mathcal{I}}$ be the order three R-dimensional identity matrix. For any matrix $\boldsymbol{A}\in\mathbb{R}^{n\times R}$, $\boldsymbol{B}\in\mathbb{R}^{n\times R}$ and $\boldsymbol{C}\in\mathbb{R}^{M\times R}$, we have*

$$||\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}||_F \le \min\{\sqrt{M}||\boldsymbol{C}||_{vec(\infty)}, ||\boldsymbol{C}||_F\}||\boldsymbol{A}||_F||\boldsymbol{B}||_F,$$

where $||\boldsymbol{C}||_{vec(\infty)}$ is the l_∞-norm of the vectorization of $\boldsymbol{C}$.

Proof of Lemma 1. The general Hölder inequality yields that the absolute value of the (i_1,i_2,i_3)-th entry of $\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}$ is upper bounded as

$$\big|(\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C})_{i_1,i_2,i_3}\big| = \big|\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}^T_{i_1,.}\times_2\boldsymbol{B}^T_{i_2,.}\times_3\boldsymbol{C}^T_{i_3,.}\big| \le ||\boldsymbol{A}_{i_1,.}||||\boldsymbol{B}_{i_2,.}||||\boldsymbol{C}_{i_3,.}||_\infty.$$

Consequently,

$$\begin{aligned}||\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}||_F^2 &= \sum_{i_1,i_2,i_3}\big|(\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C})_{i_1,i_2,i_3}\big|^2 \qquad (2)\\ &\le M||\boldsymbol{C}||^2_{\text{vec}(\infty)}\sum_{i_1,i_2}||\boldsymbol{A}_{i_1,.}||^2||\boldsymbol{B}_{i_2,.}||^2 = M||\boldsymbol{C}||^2_{\text{vec}(\infty)}||\boldsymbol{A}||_F^2||\boldsymbol{B}||_F^2.\end{aligned}$$

Besides, the Cauchy-Schwarz inequality implies that the absolute value of the (i_1,i_2,i_3)-th entry of $\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}$ is upper bounded as

$$\big|(\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C})_{i_1,i_2,i_3}\big| = \big|\sum_j \boldsymbol{A}_{i_1,j}\boldsymbol{B}_{i_2,j}\boldsymbol{C}_{i_3,j}\big| \le ||\boldsymbol{A}_{i_1,.}*\boldsymbol{B}_{i_2,.}||||\boldsymbol{C}_{i_3,.}||,$$

where $\boldsymbol{A}_{i_1,.}*\boldsymbol{B}_{i_2,.}$ is the Hadamard product between $\boldsymbol{A}_{i_1,.}$ and $\boldsymbol{B}_{i_2,.}$. Note that

$$||\boldsymbol{A}_{i_1,.}*\boldsymbol{B}_{i_2,.}|| = \sqrt{\sum_j \boldsymbol{A}^2_{i_1,j}\boldsymbol{B}^2_{i_2,j}} \le \sqrt{||\boldsymbol{A}_{i_1,.}||^2||\boldsymbol{B}_{i_2,.}||^2} = ||\boldsymbol{A}_{i_1,.}||||\boldsymbol{B}_{i_2,.}||,$$

which leads to

$$\big|(\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C})_{i_1,i_2,i_3}\big| \le ||\boldsymbol{A}_{i_1,.}||||\boldsymbol{B}_{i_2,.}||||\boldsymbol{C}_{i_3,.}||.$$

It then follows that

$$||\boldsymbol{\mathcal{I}}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}||_F^2 \le \sum_{i_1,i_2,i_3}||\boldsymbol{A}_{i_1,.}||^2||\boldsymbol{B}_{i_2,.}||^2||\boldsymbol{C}_{i_3,.}||^2 = ||\boldsymbol{A}||_F^2||\boldsymbol{B}||_F^2||\boldsymbol{C}||_F^2. \qquad (3)$$

Finally, the desired result immediately follows form (2) and (3). □

Proof of Proposition 1. Denote $S = \{\mathbf{\Theta}\in\Omega | KL(\mathbf{\Theta}^*||\mathbf{\Theta}) \ge 4\epsilon_n\}$. Let

$$\begin{aligned}I := {}& P\Big(\sup_S\big(\mathcal{L}_\lambda(\mathbf{\Theta}^*;\boldsymbol{\mathcal{A}}) - \mathcal{L}_\lambda(\mathbf{\Theta};\boldsymbol{\mathcal{A}})\big) \ge -\epsilon_n\Big)\\ ={}& P\Big(\sup_S\frac{1}{\varphi(n,M)}\sum_{m=1}^{M}\sum_{i\le j}\big(L(\theta^*_{i,j,m};a_{i,j,m}) - L(\theta_{i,j,m};a_{i,j,m})\big) + \lambda_n\big(J(\mathbf{\Theta}^*) - J(\mathbf{\Theta})\big) \ge -\epsilon_n\Big).\end{aligned}$$

We now decompose S as follows. Let $S_u = \{\mathbf{\Theta}\in\Omega | 2^{u+1}\epsilon_n \le KL(\mathbf{\Theta}^*||\mathbf{\Theta}) < 2^{u+2}\epsilon_n\}$, for $u = 1,2,\ldots$. It immediately follows that $S = \bigcup_{u=1}^{+\infty} S_u$, and then

$$\begin{aligned}I \le{}& \sum_{u=1}^{+\infty} P\Big(\sup_{S_u}\frac{1}{\varphi(n,M)}\sum_{m=1}^{M}\sum_{i\le j}\big(L(\theta^*_{i,j,m};a_{i,j,m}) - L(\theta_{i,j,m};a_{i,j,m})\big) + \lambda_n\big(J(\mathbf{\Theta}^*) - J(\mathbf{\Theta})\big) \ge -\epsilon_n\Big)\\ :={}& \sum_{u=1}^{+\infty} I_u.\end{aligned}$$

Define an empirical process $\nu_{n,M}(\Theta, \mathcal{A}) = \frac{1}{\varphi(n,M)} \sum_{m=1}^{M} \sum_{i \le j} \Big(L(\theta^*_{i,j,m}; a_{i,j,m}) - L(\theta_{i,j,m}; a_{i,j,m}) - \mathbb{E}\big(L(\theta^*_{i,j,m}; a_{i,j,m}) - L(\theta_{i,j,m}; a_{i,j,m})\big)\Big)$, for some independent but not identical data. It then follows that

$$I_u \le P\Big(\sup_{S_u} \nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) \ge \inf_{S_u} \Big(KL(\boldsymbol{\Theta}^* || \boldsymbol{\Theta}) + \lambda_n \big(J(\boldsymbol{\Theta}) - J(\boldsymbol{\Theta}^*) \big) \Big) - \epsilon_n \Big).$$

Since $\inf_{S_u} \Big(KL(\boldsymbol{\Theta}^* || \boldsymbol{\Theta}) + \lambda_n \big(J(\boldsymbol{\Theta}) - J(\boldsymbol{\Theta}^*) \big) \Big) - \epsilon_n \ge 2^{u+1}\epsilon_n - \epsilon_n - \epsilon_n \ge 2^u \epsilon_n$ and Lemma 2 shows that $\mathbb{E} \sup_{S_u} \nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) \le 2^{u-1}\epsilon_n$ when n is large enough, we have

$$I_u \le P\big(\sup_{S_u} \nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) \ge 2^u \epsilon_n \big) \le P\Big(\sup_{S_u} \nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) \ge \mathbb{E} \sup_{S_u} \nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) + 2^{u-1}\epsilon_n \Big).$$

Let Y be a Bernoulli random variable with expectation $p = s_n \big(1 + \exp(-\theta)\big)^{-1}$, we have

$$\begin{aligned} \mathbb{E}\big(L(\theta; Y) - L(\theta^*; Y)\big) &= -2p^* \log\big((\frac{p}{p^*})^{1/2} \big) - 2(1-p^*) \log\big((\frac{1-p}{1-p^*})^{1/2} \big) \\ &\ge -2p^* \big(\frac{p^{1/2}}{(p^*)^{1/2}} - 1 \big) - 2(1-p^*) \big(\frac{(1-p)^{1/2}}{(1-p^*)^{1/2}} - 1 \big) \\ &= \big(p^{1/2} - (p^*)^{1/2} \big)^2 + \big((1-p)^{1/2} - (1-p^*)^{1/2} \big)^2, \end{aligned}$$

where $p^* = s_n \big(1 + \exp(-\theta^*)\big)^{-1}$. It immediately follows that $D^2(\boldsymbol{\Theta}, \boldsymbol{\Theta}^*) \le KL(\boldsymbol{\Theta}^* || \boldsymbol{\Theta})$. Moreover, by Lagrange's mean value theorem, we further have

$$\begin{aligned} &\mathbb{E}\big(L(\theta; Y) - L(\theta^*; Y)\big)^2 \\ =& 4p^* \Big(\log(p^{1/2}) - \log\big((p^*)^{1/2}\big) \Big)^2 + 4(1-p^*) \Big(\log\big((1-p)^{1/2}\big) - \log\big((1-p^*)^{1/2}\big) \Big)^2 \\ =& 4p^* \eta_1^{-1} \big((p^*)^{1/2} - p^{1/2} \big)^2 + 4(1-p^*)(1-\eta_2)^{-1} \big((1-p^*)^{1/2} - (1-p)^{1/2} \big)^2, \end{aligned}$$

where η_1 and η_2 are some real numbers between p and p^*. Since $(1-\xi)s_n \le p, p^* \le \xi s_n$, we have $p^* \eta_1^{-1} \le \frac{\xi}{1-\xi}$ and $(1-p^*)(1-\eta_2)^{-1} \le \frac{1-(1-\xi)s_n}{1-\xi s_n} \le \frac{\xi}{1-\xi}$, which leads to $\mathbb{E}\big(L(\theta; Y) - L(\theta^*; Y)\big)^2 \le \frac{4\xi}{1-\xi} d^2(\theta, \theta^*)$. On the set S_u, we have $KL(\boldsymbol{\Theta}^* || \boldsymbol{\Theta}) < 2^{u+2}\epsilon_n$. Therefore, the variance of $\nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}})$ can be bounded as

$$\begin{aligned} Var\big(\nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}})\big) &\le \frac{4}{\varphi^2(n,M)} \sum_{m=1}^{M} \sum_{i \le j} \mathbb{E}\big(L(\theta_{i,j,m}; a_{i,j,m}) - L(\theta^*_{i,j,m}; a_{i,j,m}) \big)^2 \\ &\le \frac{4\xi}{(1-\xi)\varphi(n,M)} D^2(\boldsymbol{\Theta}, \boldsymbol{\Theta}^*) \le \frac{4\xi}{(1-\xi)\varphi(n,M)} KL(\boldsymbol{\Theta}^* || \boldsymbol{\Theta}) < \frac{\xi 2^{u+4}\epsilon_n}{(1-\xi)\varphi(n,M)}. \end{aligned}$$

Also note that $|L(\theta; Y) - L(\theta^*; Y)|$ can be upper bounded as

$$\big| L(\theta; Y) - L(\theta^*; Y) \big| \le \max\{ \big| \log \frac{1 + \exp(-\theta)}{1 + \exp(-\theta^*)} \big|, \big| \log \frac{1 - s_n \big(1 + \exp(-\theta)\big)^{-1}}{1 - s_n \big(1 + \exp(-\theta^*)\big)^{-1}} \big| \} \le \log 2 + \frac{\xi}{1-\xi},$$

where the last inequality comes from the fact that $|\theta| \le \frac{\xi}{1-\xi}$. It follows that

$$\frac{1}{2(\log 2 + \frac{\xi}{1-\xi})} \Big(L(\theta^*_{i,j,m}; a_{i,j,m}) - L(\theta_{i,j,m}; a_{i,j,m}) - \mathbb{E}\big(L(\theta^*_{i,j,m}; a_{i,j,m}) - L(\theta_{i,j,m}; a_{i,j,m}) \big) \Big) \in [-1, 1].$$

Denote $\tilde{\nu}_{n,M}(\boldsymbol{\theta}; \boldsymbol{\mathcal{A}}) = \varphi(n,M)\nu_{n,M}(\boldsymbol{\Theta}; \boldsymbol{\mathcal{A}})/(2\log 2 + \frac{2\xi}{1-\xi})$. By the concentration inequality in Theorem 1.1 of [2], we have

$$\begin{aligned} I_u &\le \exp\Big(- \frac{\big(\varphi(n,M)2^{u-1}\epsilon_n/(2\log 2 + \frac{2\xi}{1-\xi})\big)^2}{2\Big(2\mathbb{E}\sup_{S_u}\tilde{\nu}_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) + \sup_{S_u} Var\big(\tilde{\nu}_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}})\big)\Big) + 3\frac{\varphi(n,M)}{2(\log 2 + R^{3/2}\xi^2)} * 2^{u-1}\epsilon_n}\Big) \\ &< \exp\Big(- \frac{\big(\varphi(n,M)2^{u-1}\epsilon_n/(2\log 2 + \frac{2\xi}{1-\xi})\big)^2}{2\Big(\frac{\varphi(n,M)}{(\log 2 + \frac{\xi}{1-\xi})} * 2^{u-1}\epsilon_n + \frac{\xi\varphi(n,M)}{4(1-\xi)(\log 2 + R^{3/2}\xi^2)^2} * 2^{u+4}\epsilon_n\Big) + 3\frac{\varphi(n,M)}{2(\log 2 + \frac{\xi}{1-\xi})} * 2^{u-1}\epsilon_n}\Big) \\ &= \exp\Big(- \frac{2^u\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi} + 28\log 2}\Big). \end{aligned}$$

Denote $\zeta = \exp\big(- \frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi} + 28\log 2}\big)$. We have

$$I \le \sum_{u=1}^{+\infty} \exp\Big(- \frac{2^u\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi} + 28\log 2}\Big) \le \sum_{u=1}^{+\infty}\zeta^u = \frac{\zeta}{1-\zeta}.$$

As a result, $I \le (1+I)\zeta \le 2\zeta$. □

Lemma 2. *Let the set S_u and the empirical process $\nu_{n,M}(\boldsymbol{\Theta}; \boldsymbol{\mathcal{A}})$ be defined in the proof of Proposition 1. If $\frac{(n+M)R}{\varphi(n,M)\epsilon_n}\log\sqrt{\frac{1}{\epsilon_n}} \le c_1$, for some constant c_1 that depends on ξ only, then for any $u = 1, 2, \ldots$, we have $E\big(\sup_{S_u}\nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}})\big) \le 2^{u-1}\epsilon_n$.*

Proof of Lemma 2. Denote $f(\theta_{i,j,m}; a_{i,j,m}) = L(\theta^*_{i,j,m}; a_{i,j,m}) - L(\theta_{i,j,m}; a_{i,j,m})$, and hence $\nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) = \varphi^{-1}(n,M)\sum_{m=1}^M\sum_{i\le j}\big(f(\theta_{i,j,m}; a_{i,j,m}) - \mathbb{E}f(\theta_{i,j,m}; a_{i,j,m})\big)$. Let $\boldsymbol{\mathcal{A}}' = (a'_{i,j,m})$ be an independent copy of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\tau} = (\tau_{i,j,m})$ be a collection of independent Rademacher random variables. By the standard symmetrization argument, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mathcal{A}}}\sup_{S_u}\nu_{n,M}(\boldsymbol{\Theta}, \boldsymbol{\mathcal{A}}) &\le \frac{1}{\varphi(n,M)}\mathbb{E}_{\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{A}}'}\sup_{S_u}\sum_{m=1}^M\sum_{i\le j}\big(f(\theta_{i,j,m}; a_{i,j,m}) - f(\theta_{i,j,m}; a'_{i,j,m})\big) \\ &\le \frac{2}{\varphi(n,M)}\mathbb{E}_{\boldsymbol{\mathcal{A}},\boldsymbol{\tau}}\sup_{S_u}|\sum_{m=1}^M\sum_{i\le j}\tau_{i,j,m}f(\theta_{i,j,m}; a_{i,j,m})|. \end{aligned}$$

Denote $X(\boldsymbol{\Theta}; \boldsymbol{\mathcal{A}}) = \varphi^{-1/2}(n,M)\sum_{m=1}^M\sum_{i\le j}\tau_{i,j,m}f(\theta_{i,j,m}; a_{i,j,m})$ as the conditional Rademacher process. For any $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)} \in S_u$, and $\omega \in \mathbb{R}$, we have $\mathbb{E}_{\boldsymbol{\tau}|\boldsymbol{\mathcal{A}}}\exp\Big(\omega\big(X(\boldsymbol{\Theta}^{(1)}; \boldsymbol{\mathcal{A}}) - X(\boldsymbol{\Theta}^{(2)}; \boldsymbol{\mathcal{A}})\big)\Big) \le \exp\big(\frac{1}{2}\omega^2\rho^2(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}; \boldsymbol{\mathcal{A}})\big)$, where

$$\rho^2(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}; \boldsymbol{\mathcal{A}}) = \frac{1}{\varphi(n,M)}\sum_{m=1}^M\sum_{i\le j}\big(f(\theta^{(1)}_{i,j,m}; a_{i,j,m}) - f(\theta^{(2)}_{i,j,m}; a_{i,j,m})\big)^2,$$

showing that $X(\boldsymbol{\Theta}; \boldsymbol{\mathcal{A}})$ is a sub-Gaussian process with respect to ρ when $\boldsymbol{\mathcal{A}}$ is given. Thus, by Theorem 3.11 of [3], there exists a positive constant c_4, such that

$$\varphi^{-1/2}(n,M)\mathbb{E}_{\boldsymbol{\mathcal{A}},\boldsymbol{\tau}}\sup_{S_u}|\sum_{m=1}^M\sum_{i\le j}\tau_{i,j,m}f(\theta_{i,j,m}; a_{i,j,m})| \le \frac{c_4}{2}\mathbb{E}_{\boldsymbol{\mathcal{A}}}\int_0^{\mathrm{diam}(S_u)} H^{1/2}(\varepsilon; S_u, \rho)d\varepsilon,$$

where $\mathrm{diam}(S_u)$ is the diameter of S_u and $H(\varepsilon; S_u, \rho)$ is the metric entropy. Note that $|\frac{dL(\theta_{i,j,m}; a_{i,j,m})}{d\theta_{i,j,m}}| = |\frac{\exp(-\theta_{i,j,m})}{1 - s_n + \exp(-\theta_{i,j,m})}(p_{i,j,m} - a_{i,j,m})| < 1$. Thus, both $L(\theta_{i,j,m}; a_{i,j,m})$ and $f(\theta_{i,j,m}; a_{i,j,m})$ are Lipschitz continuous with Lipschitz constant 1. Thus, for any $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)} \in S_u$, we have

$$\rho^2(\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}; \boldsymbol{\mathcal{A}}) \le \frac{1}{\varphi(n,M)}\sum_{m=1}^M\sum_{i\le j}|\theta^{(1)}_{i,j,m} - \theta^{(2)}_{i,j,m}|^2 \le \frac{1}{\varphi(n,M)}||\boldsymbol{\Theta}^{(1)} - \boldsymbol{\Theta}^{(2)}||_F^2.$$

By the triangle inequality and Lemma 1,

$$\begin{aligned}\rho(\mathbf{\Theta}^{(1)},\mathbf{\Theta}^{(2)};\boldsymbol{\mathcal{A}}) &\leq \frac{1}{\varphi^{1/2}(n,M)}\Big(||\boldsymbol{\mathcal{I}}\times_1(\boldsymbol{\alpha}^{(1)}-\boldsymbol{\alpha}^{(2)})\times_2\boldsymbol{\alpha}^{(1)}\times_3\boldsymbol{\beta}^{(1)}||_F \\ &\quad + ||\boldsymbol{\mathcal{I}}\times_1\boldsymbol{\alpha}^{(2)}\times_2(\boldsymbol{\alpha}^{(1)}-\boldsymbol{\alpha}^{(2)})\times_3\boldsymbol{\beta}^{(1)}||_F + ||\boldsymbol{\mathcal{I}}\times_1\boldsymbol{\alpha}^{(2)}\times_2\boldsymbol{\alpha}^{(2)}\times_3(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}||_F\Big) \\ &\leq \frac{\left(\sqrt{\min\{M,R\}}||\boldsymbol{\alpha}^{(1)}-\boldsymbol{\alpha}^{(2)}||_F(||\boldsymbol{\alpha}^{(1)}||_F+||\boldsymbol{\alpha}^{(2)}||_F)+\min\{2\sqrt{M},||\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}||_F\}||\boldsymbol{\alpha}^{(2)}||_F^2\right)}{\varphi^{1/2}(n,M)} \\ &\leq \frac{n\log\frac{\xi}{1-\xi}}{\varphi^{1/2}(n,M)}\Big(2\sqrt{\min\{M,R\}}||\frac{1}{\sqrt{\log\frac{\xi}{1-\xi}}}(\boldsymbol{\alpha}^{(1)}-\boldsymbol{\alpha}^{(2)})||_F+\sqrt{R}\min\{2\sqrt{\frac{M}{R}},\frac{1}{\sqrt{R}}||\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}||_F||\}\Big).\end{aligned}$$

This leads to

$$H(\varepsilon;S_u,\rho)\leq H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{4n\sqrt{\min\{M,R\}}\log\frac{\xi}{1-\xi}};B(nR),||\cdot||\Big)+H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{2n\sqrt{R}\log\frac{\xi}{1-\xi}};B^h(MR),h\Big),$$

where $B(nR)$ is the unit ball with respect to the l_2-norm in $\mathbb{R}^{nR}$, $B^h(MR)$ is the Euclidean ball in $\mathbb{R}^{MR}$ with radius $\min\{2\sqrt{\frac{M}{R}},1\}$, h is a truncated distance such that $h(\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}) = \min\{2\sqrt{\frac{M}{R}},\frac{1}{\sqrt{R}}||\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}||_F\}$, and $H(\cdot;\cdot,\cdot)$ is the metric entropy.

In the case that $2\sqrt{M}\geq\sqrt{R}$, we have

$$\begin{aligned}H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{2n\sqrt{R}\log\frac{\xi}{1-\xi}};B^h(MR),h\Big) &= H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{2n\sqrt{R}\log\frac{\xi}{1-\xi}};B(MR),||\cdot||\Big) \\ &\leq MR\log\frac{6n\sqrt{R}\log\frac{\xi}{1-\xi}}{\varphi^{1/2}(n,M)\epsilon}\leq MR\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon},\end{aligned}$$

where $B(MR)$ is the unit ball with respect to the l_2 norm in $\mathbb{R}^{MR}$. In the case that $2\sqrt{M}<\sqrt{R}$, we have

$$H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{2n\sqrt{R}\log\frac{\xi}{1-\xi}};B^h(MR),h\Big)=H\Big(\frac{\varphi^{1/2}(n,M)\varepsilon}{2n\sqrt{R}\log\frac{\xi}{1-\xi}}\cdot\frac{\sqrt{R}}{2\sqrt{M}};B(MR),||\cdot||\Big)\leq MR\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}.$$

Thus, $H(\varepsilon;S_u,\rho)$ can be bounded as

$$H(\varepsilon;S_u,\rho)\leq nR\log\frac{12n\sqrt{\min\{M,R\}}\log\frac{\xi}{1-\xi}}{\varphi^{1/2}(n,M)\epsilon}+MR\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}\leq(n+M)R\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}.$$

By concavity,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mathcal{A}}}\sup_{S_u}\nu_{n,M}(\mathbf{\Theta};\boldsymbol{\mathcal{A}}) &\leq \frac{c_4}{\varphi^{1/2}(n,M)}\mathbb{E}_{\boldsymbol{\mathcal{A}}}\int_0^{\mathrm{diam}(S_u)}\sqrt{(n+M)R\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}}d\varepsilon \\ &\leq c_4\sqrt{\frac{(n+M)R}{\varphi(n,M)}}\int_0^{\sqrt{\mathbb{E}_{\boldsymbol{\mathcal{A}}}\mathrm{diam}^2(S_u)}}\sqrt{\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}}d\varepsilon.\end{aligned}$$

Furthermore, according to the same argument of bounding the variance of $\nu_{n,M}(\mathbf{\Theta},\boldsymbol{\mathcal{A}})$, we have $\mathbb{E}_{\boldsymbol{\mathcal{A}}}\rho^2(\mathbf{\Theta}^{(1)},\mathbf{\Theta}^{(2)};\boldsymbol{\mathcal{A}})\leq 2\big(\mathbb{E}_{\boldsymbol{\mathcal{A}}}\rho^2(\mathbf{\Theta}^{(1)},\mathbf{\Theta}^{*};\boldsymbol{\mathcal{A}})+\mathbb{E}_{\boldsymbol{\mathcal{A}}}\rho^2(\mathbf{\Theta}^{(2)},\mathbf{\Theta}^{*};\boldsymbol{\mathcal{A}})\big)\leq 2(\frac{\xi}{1-\xi}2^{u+2}\epsilon_n +$

$\frac{\xi}{1-\xi}2^{u+2}\epsilon_n) = \frac{\xi}{1-\xi}2^{u+4}\epsilon_n$, implying that $\mathbb{E}_{\boldsymbol{\mathcal{A}}}\mathrm{diam}^2(S_u) \leq \frac{\xi}{1-\xi}2^{u+4}\epsilon_n$. Thus,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mathcal{A}}}\sup_{S_u}\nu_{n,M}(\boldsymbol{\Theta},\boldsymbol{\mathcal{A}}) \leq c_4\sqrt{\frac{(n+M)R}{\varphi(n,M)}}\int_0^{\sqrt{\frac{\xi}{1-\xi}2^{(u+4)}\epsilon_n}}\sqrt{\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\epsilon}}d\varepsilon \\ \leq \frac{12\sqrt{2}c_4\sqrt{(n+M)R}\log\frac{\xi}{1-\xi}}{\sqrt{\varphi(n,M)\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\sqrt{\frac{\xi}{1-\xi}2^{u+4}\epsilon_n}}}}\int_{12\sqrt{2}\log\frac{\xi}{1-\xi}/\sqrt{\frac{\xi}{1-\xi}2^{u+4}\epsilon_n}}^{+\infty}\frac{\log\varepsilon}{\varepsilon^2}d\varepsilon \\ = c_4\sqrt{\frac{2^{u+4}(n+M)R\frac{\xi}{1-\xi}\epsilon_n}{\varphi(n,M)\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\sqrt{\frac{\xi}{1-\xi}2^{u+4}\epsilon_n}}}}\Big(1+\log\frac{12\sqrt{2}\log\frac{\xi}{1-\xi}}{\sqrt{\frac{\xi}{1-\xi}2^{u+4}\epsilon_n}}\Big) \\ \leq c_5\sqrt{\frac{2^{u+4}(n+M)R\epsilon_n}{\varphi(n,M)}\log\sqrt{\frac{1}{\epsilon_n}}},\end{aligned}$$

for some positive constant c_5 that depends on ξ only. Finally,

$$\mathbb{E}_{\boldsymbol{\mathcal{A}}}\sup_{S_u}\nu_{n,M}(\boldsymbol{\Theta},\boldsymbol{\mathcal{A}}) \leq 4\sqrt{2}c_5\sqrt{\frac{(n+M)R}{\varphi(n,M)\epsilon_n}\log\sqrt{\frac{1}{\epsilon_n}}}\cdot 2^{u-1}\epsilon_n \leq 2^{u-1}\epsilon_n, \tag{4}$$

where the second inequality follows from the condition that $\frac{(n+M)R}{\varphi(n,M)\epsilon_n}\log\sqrt{\frac{1}{\epsilon_n}} \leq c_1$ with c_1 taking to be $\frac{1}{32c_5^2}$. □

Proof of Theorem 1. By definition of $\widehat{\boldsymbol{\Theta}}$, it follows from Proposition 1 that

$$\begin{aligned}P(D^2(\widehat{\boldsymbol{\Theta}},\boldsymbol{\Theta}^*) \geq 4\epsilon_n) &\leq P\big(KL(\boldsymbol{\Theta}^*||\widehat{\boldsymbol{\Theta}}) \geq 4\epsilon_n\big) \\ &\leq P\big(\sup_{\{\Theta\in\Omega|KL(\Theta^*||\Theta)\geq 4\epsilon_n\}}\mathcal{L}_\lambda(\Theta^*)-\mathcal{L}_\lambda(\Theta) \geq -\epsilon_n\big) \\ &\leq 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big).\end{aligned}$$

That is, with probability at least $1-2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big)$, $D^2(\widehat{\boldsymbol{\Theta}},\boldsymbol{\Theta}^*) \leq 4\epsilon_n$.

Next, we bound the F-norm of the different between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$. Let $g(x) = \log\frac{x^2}{s_n-x^2}$. By Lagrange's mean value theorem, for any $\boldsymbol{\Theta}\in\Omega$,

$$|\theta_{i,j,m}-\theta^*_{i,j,m}| = |g(p_{i,j,m}^{1/2})-g\big((p^*_{i,j,m})^{1/2}\big)| \leq \max\{\frac{2}{\sqrt{(1-\xi)s_n}\xi},\frac{2}{\sqrt{\xi s_n}(1-\xi)}\}|p_{ijm}^{1/2}-(p^*_{ijm})^{1/2}|.$$

Moreover, $\xi > 1/2$ implies that $\max\{\frac{2}{\sqrt{(1-\xi)s_n}\xi},\frac{2}{\sqrt{\xi s_n}(1-\xi)}\} = \frac{2}{\sqrt{\xi s_n}(1-\xi)}$. It then follows that $\frac{1}{\varphi(n,M)}\sum_{m=1}^M\sum_{i\leq j}(\theta_{ijm}-\theta^*_{ijm})^2 \leq \frac{4}{s_n\xi(1-\xi)^2}D^2(\boldsymbol{\Theta},\boldsymbol{\Theta}^*)$. Particularly, for the estimator $\widehat{\boldsymbol{\Theta}}$, we have

$$\frac{1}{n^2M}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_F^2 \leq \frac{8\varphi(n,M)}{n^2Ms_n\xi(1-\xi)^2}D^2(\widehat{\boldsymbol{\Theta}};\boldsymbol{\Theta}^*) \leq \frac{8}{s_n\xi(1-\xi)^2}D^2(\widehat{\boldsymbol{\Theta}};\boldsymbol{\Theta}^*) \leq \frac{32\epsilon_n}{s_n\xi(1-\xi)^2},$$

whit probability at least $1-2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big)$. □

Lemma 3. *Under the conditions of Theorem 1 and Assumption B, then there exists an absolute constant c_3 that depends on ξ only, such that*

$$\begin{aligned}&\frac{1}{n\sqrt{M}}||\boldsymbol{\mathcal{I}}\times_1\hat{\boldsymbol{Z}}\hat{\boldsymbol{C}}\times_2\hat{\boldsymbol{Z}}\hat{\boldsymbol{C}}\times_3\hat{\boldsymbol{\beta}}-\boldsymbol{\mathcal{I}}\times_1\boldsymbol{Z}^*\boldsymbol{C}^*\times_2\boldsymbol{Z}^*\boldsymbol{C}^*\times_3\boldsymbol{\beta}^*||_F \\ &\leq\Big(\frac{4\sqrt{2}}{(1-\xi)\sqrt{\xi}}+c_3\sqrt{\frac{(1+\delta)\min\{M,R\}}{M}}\Big)\sqrt{\epsilon_n s_n^{-1}},\end{aligned}$$

with probability at least $1-2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big)-n^{-2}$.

Proof of Lemma 3. We first provide a probabilistic upper bound for $J(\hat{\boldsymbol{\alpha}})$. Note that

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\mathcal{A}}) &= \frac{1}{\varphi(n,M)} \sum_{m=1}^{M} \sum_{i \le j} L(\theta^*_{i,j,m}; a_{i,j,m}) \\
&= \frac{1}{\varphi(n,M)} \sum_{m=1}^{M} \sum_{i \le j} \big(a_{i,j,m} \log \frac{1-p^*_{i,j,m}}{p_{i,j,m}} + \log \frac{1}{1-p^*_{i,j,m}} \big).
\end{aligned}
$$

Denote $X_{i,j,m} = a_{i,j,m} \log \frac{1-p^*_{i,j,m}}{p^*_{i,j,m}} + \log \frac{1}{1-p^*_{i,j,m}}$, for $i \le j$, $m \in [M]$. It follows that $\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\mathcal{A}})$ is the average of $\varphi(n,M)$ independent two-value random variables with $|X_{i,j,m}| \le c_6 \log \frac{1}{s_n}$, $\mathbb{E}X_{i,j,m} \le c_6 s_n \log \frac{1}{s_n}$ and $\mathbb{E}X^2_{i,j,m} \le c_6 s_n (\log \frac{1}{s_n})^2$, where c_6 is a constant that depends on ξ only. By Bernstein inequality, for any $t > 0$,

$$
P\big(\frac{1}{\varphi(n,M)} \sum_{m=1}^{M} \sum_{i \le j} (X_{i,j,m} - \mathbb{E}X_{i,j,m}) > t\big) \le \exp\Big\{ - \frac{\frac{1}{2}\varphi^2(n,M)t^2}{c_5 \varphi(n,M) s_n (\log \frac{1}{s_n})^2 + c_5 \varphi(n,M) t \log \frac{1}{s_n}/3} \Big\}.
$$

Taking $t = \sqrt{6c_6}\varphi^{-1/2}(n,M) s_n^{1/2} (\log \frac{1}{s_n})(\log n)^{1/2}$, with probability at least $1 - n^{-2}$, we have

$$
\lambda_n J(\hat{\boldsymbol{\alpha}}) < \mathcal{L}_\lambda(\widehat{\boldsymbol{\Theta}}; \boldsymbol{\mathcal{A}}) \le \mathcal{L}_\lambda(\boldsymbol{\Theta}^*; \boldsymbol{\mathcal{A}}) + \epsilon_n \le \frac{1}{\varphi(n,M)} \sum_{m=1}^{M} \sum_{i \le j} \mathbb{E}X_{i,j,m} + t + \epsilon_n \le c_6 s_n \log \frac{1}{s_n} + t + \epsilon_n.
$$

Clearly $t = o(s_n \log \frac{1}{s_n})$ and $\epsilon_n = o(s_n \log \frac{1}{s_n})$. Thus, the assumption $\lambda_n \epsilon_n s_n^{-2} (\log s_n^{-1})^{-1} \ge c_2$ immediately implies that $J(\hat{\boldsymbol{\alpha}}) \le \frac{(c_7-1)^2}{4} \epsilon_n s_n^{-1}$, for some constant $c_7 > 1$, with probability at least $1 - n^{-2}$.

We now turn to bound the difference between $\boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_2 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_3 \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\mathcal{I}} \times_1 \boldsymbol{Z}^*\boldsymbol{C}^* \times_2 \boldsymbol{Z}^*\boldsymbol{C}^* \times_3 \boldsymbol{\beta}^*$. Applying the triangle inequality and Lemma 1 yields that

$$
\begin{aligned}
& \frac{1}{n\sqrt{M}} \big\| \boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{\alpha}} \times_2 \hat{\boldsymbol{\alpha}} \times_3 \hat{\boldsymbol{\beta}} - \boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_2 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_3 \hat{\boldsymbol{\beta}} \big\|_F \\
\le & \frac{1}{n\sqrt{M}} ||\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}}||_F (||\hat{\boldsymbol{\alpha}}||_F + ||\hat{\boldsymbol{Z}}\hat{\boldsymbol{C}}||_F) \min\{\sqrt{M}, \sqrt{R}\} \\
\le & 2\sqrt{\frac{\log \frac{\xi}{1-\xi}}{M}} \sqrt{(1+\delta) J(\hat{\boldsymbol{\alpha}})} \min\{\sqrt{M}, \sqrt{R}\} \\
\le & (c_7 - 1) \sqrt{\frac{\min\{M,R\}}{M}} \sqrt{(1+\delta)\epsilon_n s_n^{-1} \log \frac{\xi}{1-\xi}},
\end{aligned}
\tag{5}
$$

with probability at least $1 - 2\exp\Big(- \frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi} + 28 \log 2} \Big) - n^{-2}$. Similarly,

$$
\begin{aligned}
& \frac{1}{n\sqrt{M}} \big\| \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{\alpha}^* \times_2 \boldsymbol{\alpha}^* \times_3 \boldsymbol{\beta}^* - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{Z}^*\boldsymbol{C}^* \times_2 \boldsymbol{Z}^*\boldsymbol{C}^* \times_3 \boldsymbol{\beta}^* \big\|_F \\
\le & 2\sqrt{\frac{\min\{M,R\}}{M}} \sqrt{J(\boldsymbol{\alpha}^*) \log \frac{\xi}{1-\xi}} = o\Big(\sqrt{\frac{\min\{M,R\}}{M}} \sqrt{\epsilon_n s_n^{-1} \log \frac{\xi}{1-\xi}} \Big),
\end{aligned}
\tag{6}
$$

where the equality follows from $\lambda_n J(\boldsymbol{\alpha}^*) \le \epsilon_n$ and Assumption B. Finally, by (5), (6) and Theorem 1, we have

$$
\begin{aligned}
& \frac{1}{n\sqrt{M}} ||\boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_2 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_3 \hat{\boldsymbol{\beta}} - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{Z}^*\boldsymbol{C}^* \times_2 \boldsymbol{Z}^*\boldsymbol{C}^* \times_3 \boldsymbol{\beta}^*||_F \\
\le & \frac{1}{n\sqrt{M}} ||\boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_2 \hat{\boldsymbol{Z}}\hat{\boldsymbol{C}} \times_3 \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}||_F + \frac{1}{n\sqrt{M}} ||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*||_F \\
& \qquad + \frac{1}{n\sqrt{M}} ||\boldsymbol{\Theta}^* - \boldsymbol{\mathcal{I}} \times_1 \boldsymbol{Z}^*\boldsymbol{C}^* \times_2 \boldsymbol{Z}^*\boldsymbol{C}^* \times_3 \boldsymbol{\beta}^*||_F \\
\le & \Big(\frac{4\sqrt{2}}{(1-\xi)\sqrt{\xi}} + c_7 \sqrt{\frac{\min\{M,R\}}{M}} \sqrt{(1+\delta) \log \frac{\xi}{1-\xi}} \Big) \sqrt{\epsilon_n s_n^{-1}}.
\end{aligned}
$$

The desired result follows by taking $c_3 = c_7\sqrt{\log \frac{\xi}{1-\xi}}$.

$\square$

Lemma 4. *Let $\widehat{\boldsymbol{\mathcal{B}}} = \boldsymbol{\mathcal{I}} \times_1 \hat{\boldsymbol{C}} \times_2 \hat{\boldsymbol{C}} \times_3 \hat{\boldsymbol{\beta}}$ be the estimation counterpart of $\boldsymbol{\mathcal{B}}^*$. Denote $\boldsymbol{\mathcal{M}}^* = \boldsymbol{\mathcal{B}}^* \times_2 \boldsymbol{Z}^*$ and $\widehat{\boldsymbol{\mathcal{M}}} = \widehat{\boldsymbol{\mathcal{B}}} \times_2 \hat{\boldsymbol{Z}}$. Under the conditions of Lemma 3 and Assumption A and C, then with probability at least $1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2}$, the following event F holds. F: for any $k \in [K]$, there exists an unique $k' \in [K]$, such that*

$$\frac{1}{\sqrt{nM}}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{k',\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F = o(\gamma_n\sqrt{\frac{n_{\min}K}{n}}).$$

Proof of Lemma 4. Denote F_0 be the event that there exists $k \in [K]$ such that $\frac{1}{\sqrt{nM}}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{k',\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F \geq \gamma_n\sqrt{\frac{n_{\min}K}{n}}$, for any $k' \in [K]$ and sufficiently large n and M. It follows that

$$\begin{aligned}\frac{1}{n\sqrt{M}}\big\|\widehat{\boldsymbol{\mathcal{B}}} \times_1 \hat{\boldsymbol{Z}} \times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^* \times_1 \boldsymbol{Z}^* \times_2 \boldsymbol{Z}^*\big\|_F &\geq \frac{1}{n\sqrt{M}}\Big(\sum_{i\in N_k^*}\big\|(\widehat{\boldsymbol{\mathcal{M}}} \times_1 \hat{\boldsymbol{Z}})_{i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F^2\Big)^{1/2}\\ &\geq \gamma_n\sqrt{\frac{n_{\min}K}{n}}\sqrt{\frac{n_k}{n}} \geq \frac{\gamma_n n_{\min}\sqrt{K}}{n}.\end{aligned}$$

However, it follows from Lemma 3 and the Assumption C that

$$\begin{aligned}P(F_0) &\leq P\Big(\frac{1}{n\sqrt{M}}\big\|\widehat{\boldsymbol{\mathcal{B}}} \times_1 \hat{\boldsymbol{Z}} \times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^* \times_1 \boldsymbol{Z}^* \times \boldsymbol{Z}^*\big\|_F \geq \frac{\gamma_n n_{\min}\sqrt{K}}{n}\Big)\\ &\leq 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) + n^{-2}.\end{aligned}$$

Therefore, with probability at least $1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2}$, F_0^c, the complement of F_0 holds; that is; the existence holds with high probability.

We now prove the uniqueness under F_0^c. Assume there exist $k_1 \neq k_2 \in [K]$ such that $\frac{1}{\sqrt{nM}}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{k_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F = o(\gamma_n\sqrt{\frac{n_{\min}K}{n}})$, for $i \in [2]$. By existence, there exists $a \in [K]$ and $b_1 \neq b_2 \in [K]$ such that $\frac{1}{\sqrt{nM}}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{a,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{b_j,\cdot,\cdot}\big\|_F = o(\gamma_n\sqrt{\frac{n_{\min}K}{n}})$, for $j \in [2]$. The triangle inequality implies that

$$\frac{1}{\sqrt{nM}}\|\boldsymbol{\mathcal{M}}^*_{b_1,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{b_2,\cdot,\cdot}\|_F \leq \frac{1}{\sqrt{nM}}\Big(\big\|\widehat{\boldsymbol{\mathcal{M}}}_{a,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{b_1,\cdot,\cdot}\big\|_F + \big\|\widehat{\boldsymbol{\mathcal{M}}}_{a,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{b_2,\cdot,\cdot}\big\|_F\Big) = o(\gamma_n\sqrt{\frac{n_{\min}K}{n}}).$$

On the other hand, Assumption A implies that

$$\frac{1}{\sqrt{nM}}\|\boldsymbol{\mathcal{M}}^*_{b_1,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{b_2,\cdot,\cdot}\|_F \geq \sqrt{\frac{n_{\min}}{nM}}\|\boldsymbol{\mathcal{B}}^*_{b_1,\cdot,\cdot} - \boldsymbol{\mathcal{B}}^*_{b_2,\cdot,\cdot}\|_F \geq \sqrt{\frac{n_{\min}K}{n}}\gamma_n,$$

which is a contradiction. Hence, F_0^c also implies uniqueness, showing that F holds with probability at least $1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2}$. $\square$

Proof of Theorem 2. Based on Lemma 4, with probability at least $1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2}$, there exists a permutation $\pi^* \in S_K$ such that for each $k \in [K]$, $\frac{1}{\sqrt{nM}}\|\widehat{\boldsymbol{\mathcal{M}}}_{\pi^*(k),\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\|_F = o(\gamma_n\sqrt{\frac{n_{\min}K}{n}})$. It then suffices to show that with probability at least $1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2}$, it holds true that $\min_{\pi\in S_K}\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\psi_i^* \neq \pi(\hat{\psi}_i)\} \leq \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}$. Let $\hat{N}_k = \{i : \hat{\psi}_i = k\}$, for $k \in [K]$. Note that

$$\min_{\pi\in S_K}\sum_{i=1}^n \mathbf{1}\{\psi_i^* \neq \pi(\hat{\psi}_i)\} = \min_{\pi\in S_K}\sum_{k=1}^K |N_k^* \setminus \hat{N}_{\pi^{-1}(k)}| = \min_{\pi\in S_K}\sum_{k=1}^K |N_k^* \setminus \hat{N}_{\pi(k)}|,$$

where the last equality follows from the fact that π^{-1} is also a permutation in S_K. It then suffices to show that with probability at least $1-2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big)-n^{-2}$, $\frac{1}{n}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}$ for the particular permutation π^*. Let F denote the same event in Lemma 4. In fact, by Lemma 3, we have

$$
\begin{aligned}
&P\big(\frac{1}{n}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}|F\big) = P\big(\frac{n_{\min}K\gamma_n^2}{n^2}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| = \frac{c_\xi^2 \epsilon_n}{s_n}|F\big)\\
&\ge P\big(\frac{n_{\min}K\gamma_n^2}{n^2}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le \frac{1}{n^2M}\big\|\widehat{\boldsymbol{\mathcal{B}}}\times_1 \hat{\boldsymbol{Z}}\times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^*\times_1 \boldsymbol{Z}^*\times_2 \boldsymbol{Z}^*\big\|_F^2|F\big)+\\
&\qquad P\big(\frac{1}{n^2M}\big\|\widehat{\boldsymbol{\mathcal{B}}}\times_1 \hat{\boldsymbol{Z}}\times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^*\times_1 \boldsymbol{Z}^*\times_2 \boldsymbol{Z}^*\big\|_F^2 \le \frac{c_\xi^2\epsilon_n}{s_n}|F\big)-1\\
&= P\big(\frac{n_{\min}K\gamma_n^2}{n^2}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le \frac{1}{n^2M}\big\|\widehat{\boldsymbol{\mathcal{B}}}\times_1 \hat{\boldsymbol{Z}}\times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^*\times_1 \boldsymbol{Z}^*\times_2 \boldsymbol{Z}^*\big\|_F^2|F\big),
\end{aligned}
$$

where the last equality comes from the fact that the event F is based on the resultant inequality of Lemma 3. Furthermore, note that

$$
\begin{aligned}
&\frac{1}{n^2M}\big\|\widehat{\boldsymbol{\mathcal{B}}}\times_1 \hat{\boldsymbol{Z}}\times_2 \hat{\boldsymbol{Z}} - \boldsymbol{\mathcal{B}}^*\times_1 \boldsymbol{Z}^*\times_2 \boldsymbol{Z}^*\big\|_F^2 \ge \frac{1}{n^2M}\sum_{k=1}^K \sum_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \big\|(\widehat{\boldsymbol{\mathcal{M}}}\times_1 \hat{\boldsymbol{Z}})_{i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F^2\\
&\ge \frac{1}{n^2M}\sum_{k=1}^K \sum_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \left(\frac{\big\|\boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F^2}{2} - \|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\|_F^2\right)\\
&\ge \sum_{k=1}^K \frac{|N_k^*\setminus \hat{N}_{\pi^*(k)}|}{n^2M} \min_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \left(\frac{1}{2}\big\|\boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F^2 - \big\|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\big\|_F^2\right)\\
&\ge \sum_{k=1}^K \frac{|N_k^*\setminus \hat{N}_{\pi^*(k)}|}{n^2M}\Big(\frac{1}{2}n_{\min}MK\gamma_n^2 - \max_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \big\|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\big\|_F^2\Big).
\end{aligned}
$$

Here we use the fact that $\min_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \frac{1}{nM}\big\|\boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{k,\cdot,\cdot}\big\|_F^2 \ge \frac{n_{\min}K}{n}\gamma_n^2$ according to Assumption A. Consequently,

$$
\begin{aligned}
&P\big(\frac{1}{n}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}|F\big)\\
&\ge P\Big(\frac{n_{\min}K\gamma_n^2}{n^2}\sum_{k=1}^K |N_k^*\setminus \hat{N}_{\pi^*(k)}| \le\\
&\qquad \sum_{k=1}^K \frac{|N_k^*\setminus \hat{N}_{\pi^*(k)}|}{n^2M}\Big(\frac{1}{2}n_{\min}MK\gamma_n^2 - \max_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \big\|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\big\|_F^2\Big)|F\Big)\\
&\ge P\Big(\bigcap_{k=1}^K (\{\frac{n_{\min}K\gamma_n^2}{n^2} \le \frac{n_{\min}K\gamma_n^2}{2n^2} - \max_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \frac{1}{n^2M}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\big\|_F^2\}\bigcap F)\Big)\\
&\ge P\Big(\bigcap_{k=1}^K (\{\max_{i\in N_k^*\setminus \hat{N}_{\pi^*(k)}} \frac{1}{nM}\big\|\widehat{\boldsymbol{\mathcal{M}}}_{\hat{\psi}_i,\cdot,\cdot} - \boldsymbol{\mathcal{M}}^*_{(\pi^*)^{-1}(\hat{\psi}_i),\cdot,\cdot}\big\|_F^2 = o(\frac{n_{\min}K\gamma_n^2}{n})\}\bigcap F)\Big) = 1,
\end{aligned}
$$

where the last equality is suggested by Lemma 4. Finally, by the definition of conditional probability,

$$
\begin{aligned}
P\Big(\frac{1}{n}\sum_{k=1}^{K}|N_k^*\setminus \hat{N}_{\pi^*(k)}| \leq \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}\Big) &= P\Big(\frac{1}{n}\sum_{k=1}^{K}|N_k^*\setminus \hat{N}_{\pi^*(k)}| \leq \frac{c_\xi^2 n\epsilon_n}{n_{\min}K\gamma_n^2 s_n}|F\Big)\cdot P(F) \\
&\geq 1 - 2\exp\Big(-\frac{\varphi(n,M)\epsilon_n}{156\frac{\xi}{1-\xi}+28\log 2}\Big) - n^{-2},
\end{aligned}
$$

and thus the desired consistency result follows immediately. $\square$

References

[1] Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

[2] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *Ann. Prob.*, 33(3):1060–1077, 2005.

[3] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[4] Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45):12679–12684, 2016.