Supplementary Material for "On Variational Inference in Biclustering Models"

Guanhua Fang, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St Bellevue WA 98004 USA
{guanhuafang, liping11}@baidu.com

In this supplementary, we provide the proofs for Theorems 1-11 stated in the main paper. We first recall the following notation. Letter θ denotes the generic model parameter; π_1, π_2 are latent class probabilities and ϕ_{1i}, ϕ_{2j}'s are variational parameters. J_1 is the number of latent classes for the first mode and J_2 is the number of latent classes for the second mode. For positive integer m, we use $[m]$ to denote set $\{1, \ldots, m\}$ and $[m_1] \times [m_2]$ to denote set $\{(i,j) : i \in [m_1], j \in [m_2]\}$. For two positive sequences $\{a_n\}$, $\{b_n\}$, $a_n \lesssim b_n$ means that $a_n \leq C b_n$ for some large constant C independent of n, and $a_n \asymp b_n$ means that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The symbols $\mathbb{E}$ and $P(\cdot)$ denote generic expectation and probability whose distribution may be determined from the context. Additionally, $\|x\|$ / $\|x\|_1$ is used to denote ℓ_2- / ℓ_1- norm of vector x and $\|X\|_F$ is used to denote Frobenius norm of matrix X. We use $x[i]$ to represent the i-th entry of vector x and use $X[i,j]$ to represent the entry of matrix X on i-th row and j-th column. We use ∇f to represent the derivative of function f with respect to θ. For random sequences a_n and b_n, $a_n = O_p(b_n)$ represents that a_n is stochastically bounded by Kb_n for a sufficiently large constant K; $a_n = o_p(b_n)$ represents a_n/b_n converges to 0 with probability tending to 1. Constants c, C may vary from the place to place.

1. Comment on Algorithm 1

We recall that the evidence lower bound has the following form,

$$\begin{aligned}\text{ELBO} &= \sum_{(i,j)} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l] \log f_\theta(y_{ij}|k,l) \\ &\quad + \sum_i \sum_k \phi_{1i}[k] \log(\pi_1[k]/\phi_{1i}[k]) \\ &\quad + \sum_j \sum_l \phi_{2j}[l] \log(\pi_2[l]/\phi_{2j}[l]),\end{aligned}$$

where ϕ_1's and ϕ_2's are variational parameters in multinomial distributions. Then update rule can be obtained as follows.

Update ϕ: by the conjugacy property, we easily know

$$\begin{aligned}\phi_{1i}[k] &\propto \exp\{\sum_{j:(i,j)\in\Omega} \mathbb{E}_{\phi_{2j}} \log f_{\theta_{kz_{2j}}}(y_{ij}) + \log \pi_1[k]\} \\ &= \exp\{\sum_{j:(i,j)\in\Omega} \sum_l \phi_{2j}[l] \log f_{\theta_{kl}}(y_{ij}) + \log \pi_1[k]\} \\ &= \exp\{\sum_{j'=1}^{J_2} \sum_{j:(i,j)\in\Omega; z_{2j}^*=j'} \sum_l \phi_{2j}[l] \log f_{\theta_{kl}}(y_{ij}) + \log \pi_1[k]\}.\end{aligned}$$

and similarly

$$\phi_{2j}[l] \quad \propto \quad \exp\{\sum_{j'=1}^{J_1} \sum_{i:(i,j)\in\Omega; z_{1i}^*=j'} \sum_k \phi_{1i}[k] \log f_{\theta_{kl}}(y_{ij}) + \log \pi_2[l]\}.$$

Update θ: for $k \in [J_1]$ and $l \in [J_2]$, we have

$$\theta_{kl} \quad = \quad \arg\max\{ \sum_{(i,j)\in\Omega} \phi_{1i}[k]\phi_{2j}[l] \log f_\theta(y_{ij})\}.$$

Here we suppress superscript index t here for simplicity.

Some Special Families

- Homogeneous Poisson Process Case: the likelihood is

$$f_\theta(y) = \exp\{n \log\theta - t\theta\},$$

with θ being the intensity parameter and $y = (y_1, \ldots, y_t, \ldots, y_n)$. Then update formula for θ_{kl} can be reduced to $\theta_{kl} = \frac{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]n_{ij}}{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]t_{ij}}$. Its special case, Poisson model, assumes the density,

$$f_\theta(y) = \frac{\exp\{-\theta\}\theta^y}{y!}.$$

The update formula for θ_{kl} can be reduced to $\theta_{kl} = \frac{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]y_{ij}}{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]}$.

- Bernoulli case: the density function is

$$f_\theta(y) = \exp\{y\log(\theta) + (1-y)\log(1-\theta)\},$$

with $y \in \{0, 1\}$. Then the update formula for θ_{kl} is simplified as $\theta_{kl} = \frac{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]y_{ij}}{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]}$.

- Multi-categorical case: the density function becomes

$$f_\theta(y) = \exp\{\sum_{c=1}^{C} \mathbf{1}\{y = c\} \log\theta[c]\},$$

with $y \in \{1, \ldots, C\}$ and $\sum \theta_c = 1$. Then the update formula for θ_{kl} can be written as

$$\theta_{kl}[c] \propto \frac{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]\mathbf{1}\{y_{ij} = c\}}{\sum_{(i,j)\in A} \phi_{1i}[k]\phi_{2j}[l]}$$

and $\sum_c \theta_{kl}[c] = 1$ for all k, l.

2. Proof of Results in Section 3

In this section, we prove Theorem1 - Theorem 5 in Section 3 of the main paper. We first study the consistency property of the variational estimator.

Consistency Let $\hat{\theta}, \hat{\pi}$ be the estimated model parameter and $\hat{\phi}$ be the estimated variational parameter. For any fixed $a \in [J_1]$ and $b \in [J_2]$ and any fixed ϕ, we first compare the difference between

$$Q_1(a,b) := \sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta^*_{ab}}(y_{ij}) \tag{15}$$

and

$$Q_2(\theta,a,b) := \sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta_{kl}}(y_{ij}). \tag{16}$$

We are going to show that there exist constant $c(\delta)$ and δ such that

$$Q_1(a,b) - Q_2(\theta,a,b) \geq c(\delta) m_1 m_2 \tag{17}$$

for any θ satisfying $\min_{k,l} \|\theta_{kl} - \theta^*_{ab}\| \geq \delta$ with high probability.

Proof of (17): For any sequences $\{\psi_{1i}\}$ and $\{\psi_{2i}\}$ satisfying that

C_ψ: $\sum_{i,j} \psi_{1i}\psi_{2j} \geq n_0$ and each element is bounded by 1,

we can show that

$$P(\frac{1}{n_0}|\sum_{i,j} \psi_{1i}\psi_{2j} \log f_\theta(y_{ij}) - \sum_{i,j} \psi_{1i}\psi_{2j} \mathbb{E} \log f_\theta(y_{ij})| \geq t) \leq \exp\{-\frac{n_0^2 t^2/2}{n_0 V + n_0 M t/3}\}, \tag{18}$$

where M is the upper bound of $|\log f_\theta(y_{ij})|$ and V is the upper bound of $\mathbb{E}(\log f_\theta(y_{ij}))^2$. Inequality (18) holds since that $|\psi_{1i}\psi_{2j} \log f_\theta(y_{ij}) - \psi_{1i}\psi_{2j}\mathbb{E}\log f_\theta(y_{ij})|$ is bounded by $|\log f_\theta(y_{ij}) - \mathbb{E}\log f_\theta(y_{ij})|$ and Bernstein inequality. By union bound, we can further have the following uniform concentration inequality,

$$\begin{aligned} & P(\sup_{\psi \in C_\psi, \theta} \frac{1}{n_0}|\sum_{i,j} \psi_{1i}\psi_{2j} \log f_\theta(y_{ij}) - \sum_{i,j} \psi_{1i}\psi_{2j} \mathbb{E} \log f_\theta(y_{ij})| \geq t) \\ \leq \quad & \mathcal{N}(\psi)\mathcal{N}(\theta) \exp\{-\frac{n_0^2 t^2/2}{n_0 V + n_0 M t/3}\}, \end{aligned} \tag{19}$$

where $\mathcal{N}(\psi)$ is the covering number of $t/4$-ball for ψ and $\mathcal{N}(\theta)$ is the covering number of $t/4$-ball for θ. By straightforward calculation, $\mathcal{N}(\psi)$ is bounded by $C^{m_1 \log J_1 + m_2 \log J_2}$ and $\mathcal{N}(\theta)$ is bounded by $C^{J_1 J_2}$ for some sufficiently large constant C.

Note that $\sum_{k,l}\sum_{i,j} \phi_{1i}[k]\phi_{2i}[l] = m_a m_b$, where $m_a := |\{i : z^*_{1i} = a\}|$ and $m_b := |\{i : z^*_{2j} = b\}|$. We consider those pairs (k,l)'s such that $\sum_{i,j} \phi_{1i}[k]\phi_{2i}[l] = \Theta(m_a m_b)$. We then call these pairs (k,l)'s satisfy relation R_m.

By (19), we know that

$$\sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta^*_{ab}}(y_{ij}) = \sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta^*_{ab}}(y_{ij}) + O_p(\sqrt{m_a m_b \mathcal{N}(\psi)\mathcal{N}(\theta)})$$

and

$$\sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta_{kl}}(y_{ij}) = \sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \log f_{\theta_{kl}}(y_{ij}) + O_p(\sqrt{m_a m_b \mathcal{N}(\psi)\mathcal{N}(\theta)}).$$

By the optimality of θ^*_{ab}, we have that

$$\sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \mathbb{E} \log f_{\theta^*_{ab}}(y_{ij}) - \sum_{i,j: z^*_{1i}=a, z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l] \mathbb{E} \log f_{\theta_{kl}}(y_{ij}) \geq c m_a m_b \delta,$$

for any $\|\theta_{kl} - \theta^*_{ab}\| \geq \delta$. Therefore, if there is no θ_{kl} such that $|\theta_{kl} - \theta^*_{ab}| < \delta$,

$$\begin{aligned}
& \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta^*_{ab}}(y_{ij}) - \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta_{kl}}(y_{ij}) \\
\geq \quad & cm_a m_b \delta - O_p(\sqrt{m_a m_b \mathcal{N}(\psi)\mathcal{N}(\theta)}) \qquad (20) \\
\geq \quad & c' m_a m_b \delta.
\end{aligned}$$

Thus, it holds that

$$\begin{aligned}
& Q_1(a,b) - Q_2(\theta,a,b) \\
= \quad & \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta^*_{ab}}(y_{ij}) - \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta_{kl}}(y_{ij}) \\
= \quad & \sum_{k,l \text{ satisfy } R_m} \left\{ \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta^*_{ab}}(y_{ij}) - \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta_{kl}}(y_{ij}) \right\} \\
+ \quad & \sum_{k,l \text{ not satisfy } R_m} \left\{ \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta^*_{ab}}(y_{ij}) - \sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \phi_{1i}[k]\phi_{2j}[l]\log f_{\theta_{kl}}(y_{ij}) \right\} \\
\geq \quad & c' m_a m_b \delta - o_p(m_a m_b 2M) = c(\delta) m_a m_b, \qquad (21)
\end{aligned}$$

by adjusting the constant. This completes the proof of (17). This also implies that $Q_2(\theta, a, b)$ achieves its maximum at the local neighborhood of θ^*_{ab}.

By these, we can claim that $\max_{a,b}\min_{k,l}\|\theta^*_{ab} - \hat{\theta}_{kl}\| \leq \delta$. If not, there must exist a_0, b_0 such that $\min_{k,l}\|\theta^*_{a_0 b_0} - \hat{\theta}_{kl}\| > \delta$. It gives us that

$$\begin{aligned}
& \sum_{a,b} Q_1(a,b) - \sum_{a,b} Q_2(\hat{\theta},a,b) \\
= \quad & \sum_{(a,b)\neq(a_0,b_0)} \{Q_1(a,b) - Q_2(\hat{\theta},a,b)\} + Q_1(a_0,b_0) - Q_2(\hat{\theta},a_0,b_0) \\
\geq \quad & -J_1 J_2 O_p(\sqrt{m_a m_b \mathcal{N}(\psi)\mathcal{N}(\theta)}) + c(\delta) m_a m_b \\
> \quad & 0.
\end{aligned}$$

This contradicts with the definition that $(\hat{\theta}, \hat{\phi})$ is the maximizer. Therefore, we conclude that $\max_{a,b}\min_{k,l}\|\theta^*_{ab} - \hat{\theta}_{kl}\| \leq \delta$ holds with high probability. By the Assumption A1 that rows/columns of θ^* are different from each other, we then can permute rows/columns of $\hat{\theta}$ so that $\|\theta^*_{kl} - \hat{\theta}_{kl}\| \leq \delta$. Thus we conclude that $\|\theta^* - \hat{\theta}\|_F^2 \leq J_1 J_2 \delta^2$. This gives the estimation consistency of the model parameter θ. Next, we are able to show Theorem 1 and give the characterization of $\hat{\phi}$.

Proof of Theorem 1 By above displays, we have that $\frac{1}{\sqrt{J_1 J_2}}\|\theta^* - \hat{\theta}\|_F$ is $o_p(1)$. Indeed, (21) actually gives more information. That is, for any fixed pair of a and b, we can find $k(a) \in [J_1], l(b) \in [J_2]$ such that $\sum_{i,j:z^*_{1i}=a,z^*_{2j}=b} \hat{\phi}_{1i}[k]\hat{\phi}_{2j}[l] \geq m_a m_b(1 - o_p(1))$. This gives us that $\hat{\phi}_{1i}[k] = 1 - o_p(1)$ for any i with $z^*_{1i} = a$ and $\hat{\phi}_{2j}[l] = 1 - o_p(1)$ for any j with $z^*_{2j} = b$. This gives the consistency of latent membership estimation.

Moreover, given fixed i, for any $k' \neq k(= z^*_{1i})$, it holds that,

$$\sum_j \log f_{\theta^*_{kz^*_{2j}}}(y_{ij}) - \sum_j \log f_{\theta^*_{k'z^*_{2j}}}(y_{ij}) \geq cm_2 \qquad (22)$$

for some constant c by law of large number with probability at least $1 - \exp\{-Cm_2\}$. By the optimality conditions for $\hat{\phi}_{1i}$, we can compute that

$$\hat{\phi}_{1i}[k] \propto \exp\{\sum_j \sum_l \hat{\phi}_{2j}[l]\log f_{\hat{\theta}_{kl}}(y_{ij})\}.$$

Thus,

$$\begin{aligned}\log\hat{\phi}_{1i}[k]-\log\hat{\phi}_{1i}[k'] &= \sum_j\sum_l\hat{\phi}_{2j}[l]\log f_{\hat{\theta}_{kl}}(y_{ij})-\sum_j\sum_l\hat{\phi}_{2j}[l]\log f_{\hat{\theta}_{k'l}}(y_{ij})\\ &\geq \sum_j\log f_{\theta^*_{kz^*_{2j}}}(y_{ij})-\sum_j\log f_{\theta^*_{k'z^*_{2j}}}(y_{ij})-o_p(m_2)\\ &\geq c'm_2.\end{aligned}$$

It gives us that

$$\hat{\phi}_{1i}[z^*_{1i}]=\hat{\phi}_{1i}[k]\geq 1-J_1\exp\{-c'm_2\}. \tag{23}$$

Therefore, by the definition of total variation distance, we get

$$d_{TV}(\hat{q}_{1i},\delta_{z^*_{1i}})=\sum_{k'\neq z^*_{1i}}\hat{\phi}_{1i}[k']+(1-\hat{\phi}_{1i}[z^*_{1i}])\leq 2J_1\exp\{-c'm_2\}.$$

Similarly, we have that

$$d_{TV}(\hat{q}_{2j},\delta_{z^*_{2j}})=\sum_{l'\neq z^*_{2j}}\hat{\phi}_{2j}[l']+(1-\hat{\phi}_{2j}[z^*_{2j}])\leq 2J_2\exp\{-c'm_1\}.$$

By union bound, we know that $d_{TV}(\hat{q}_{1i},\delta_{z^*_{1i}})\leq 2J_1\exp\{-c'm_2\}$ and $d_{TV}(\hat{q}_{2j},\delta_{z^*_{2j}})\leq 2J_2\exp\{-c'm_1\}$ hold for $i\in[m_1]$ and $j\in[m_2]$ with probability at least $1-(m_1+m_2)\max\{J_1,J_2\}\exp\{-C\min\{m_1,m_2\}\}$. This completes the proof.

Proof of Theorem 2 By the optimality condition for $\hat{\pi}_1$ and $\hat{\pi}_2$, we can find that

$$\begin{aligned}\hat{\pi}_1[k] &= \frac{\sum_{i\in[m_1]}\hat{\phi}_{1i}[k]}{m_1}\ \text{ for } k\in[J_1],\\ \hat{\pi}_2[l] &= \frac{\sum_{j\in[m_2]}\hat{\phi}_{2j}[l]}{m_2}\ \text{ for } l\in[J_2].\end{aligned}$$

Thus, we have

$$\begin{aligned}|\hat{\pi}_1[k]-\pi^*_1[k]| &= |\frac{\sum_{i\in[m_1]}\hat{\phi}_{1i}[k]}{m_1}-\frac{\sum_{i\in[m_1]}\mathbf{1}\{z^*_{1i}=k\}}{m_1}+\frac{\sum_{i\in[m_1]}\mathbf{1}\{z^*_{1i}=k\}}{m_1}-\pi^*_1[k]|\\ &\leq 2J_1\exp\{-c'm_2\}+|\frac{\sum_{i\in[m_1]}\mathbf{1}\{z^*_{1i}=k\}}{m_1}-\pi^*_1[k]|\\ &\leq 2J_1\exp\{-c'm_2\}+\frac{c_1}{\sqrt{m_1}}\end{aligned}$$

held with probability at least $1-\exp\{-Cm_1\}-(m_1+m_2)\max\{J_1,J_2\}\exp\{-C\min\{m_1,m_2\}\}$ for some constants c_1 and C by Hoeffding's inequality. By repeating the same procedure, we can obtain that

$$|\hat{\pi}_2[l]-\pi^*_2[l]|\leq 2J_2\exp\{-c'm_1\}+\frac{c_2}{\sqrt{m_2}}$$

holds with probability at least $1-\exp\{-Cm_2\}-(m_1+m_2)\max\{J_1,J_2\}\exp\{-C\min\{m_1,m_2\}\}$. This leads to the desired result.

Proof of Theorem 3 Furthermore, we can reduce our problem to the usual maximum likelihood estimation problem. This is because we have that

$$\begin{aligned}
& ELBO(\hat{\theta}, \hat{\pi}_1, \hat{\pi}_2, \hat{\phi}) \\
= \; & \sum_{i,j}\sum_{k,l} \hat{\phi}_{1i}[k]\hat{\phi}_{2j}[l] \log f_{\hat{\theta}_{kl}}(y_{ij}) - \sum_i \sum_k \hat{\phi}_{1i}[k] \log(\hat{\phi}_{1i}[k]) - \sum_j \sum_l \hat{\phi}_{2j}[l] \log(\hat{\phi}_{2j}[l]) \\
& + \sum_i \sum_k \hat{\phi}_{1i}[k] \log(\hat{\pi}_1[k]) + \sum_j \sum_l \hat{\phi}_{2j}[l] \log(\hat{\pi}_2[l]) \\
= \; & \sum_{ij} \log f_{\hat{\theta}_{\hat{z}_{1i}\hat{z}_{2j}}}(y_{ij}) + \sum_i \log(\hat{\pi}_1[\hat{z}_{1i}]) + \sum_j \log(\hat{\pi}_2[\hat{z}_{2j}]) \\
& + O_p(m_1 m_2 \max\{J_1, J_2\}(\exp\{-c' m_1\} + \exp\{-c' m_2\})),
\end{aligned}$$

where $\hat{z}_{1i} = \arg\max_k \hat{\phi}_{1i}[k]$ and $\hat{z}_{2j} = \arg\max_l \hat{\phi}_{2j}[l]$. Since the variational estimator is shown to be consistent, we only need to consider the local neighborhood of θ^*, $B(\theta^*, \delta_0)$ with some sufficiently small radius δ_0. For any $\theta \in B(\theta^*, \delta_0)$, we can find

$$|ELBO(\theta, \phi(\theta)) - \sum_{ij} \log f_\theta(y_{ij})| \leq O_p(m_1 m_2 \max\{J_1, J_2\}(\exp\{-c' m_1\} + \exp\{-c' m_2\})),$$

by continuity of ELBO function and likelihood function. Here $\phi(\theta) := \arg\max_{\phi, \pi_1, \pi_2} ELBO$ given fixed θ. Notice the fact that the EBLO is always upper bounded by the log-likelihood function. By Assumption A4 that the expectation of likelihood function is strictly concave, then there exist ρ such that for any $\mathbb{E} \log F(\Theta) \leq \mathbb{E} \log F(\Theta^*) - 2\rho \|\Theta - \Theta^*\|_F^2$. Here we write $\sum_{ij} \log f_\theta(y_{ij})$ as $\log F(\Theta)$ for notational simplicity.

Then the maximal likelihood estimator should be a consistent estimator as well. Otherwise the log-likelihood achieves larger value at $\hat{\theta}$, this contradicts the definition of MLE. Furthermore, we can obtain the relationship between variational estimator and MLE, that is,

$$\begin{aligned}
\rho\|\hat{\Theta} - \Theta^{MLE}\|_F^2 & \leq \log F(\Theta^{MLE}) - \log F(\hat{\Theta}) \\
& = \log F(\Theta^{MLE}) - \log F(\hat{\Theta}) + ELBO(\Theta^{MLE}, \phi(\Theta^{MLE})) \\
& \quad - ELBO(\Theta^{MLE}, \phi(\Theta^{MLE})) + ELBO(\hat{\Theta}, \hat{\phi}) - ELBO(\hat{\Theta}, \hat{\phi}) \\
& \leq 2C(m_1 m_2 \max\{J_1, J_2\}(\exp\{-c' m_1\} + \exp\{-c' m_2\})) \qquad (24)
\end{aligned}$$

Here, we write $\Theta = (\theta_{ij})$ with $\hat{\Theta}_{ij} = \hat{\theta}_{\hat{z}_{1i}, \hat{z}_{2j}}$. In other words,

$$\|\hat{\Theta} - \Theta^{MLE}\|_F^2 \leq O_p(m_1 m_2 \max\{J_1, J_2\} \exp\{-c^{''} \min\{m_1, m_2\}\})$$

by adjusting the constant c''. In the following, we only need to compute the upper bound of MLE in parameter space $\mathcal{B}_\Theta$.

Work on MLE. In the following, for the sake of notational simplicity, we abuse notation $\hat{\theta}$ / $\hat{\Theta}$ by treating $\hat{\theta}$ / $\hat{\Theta}$ as MLE (θ^{MLE} / Θ^{MLE}) in the rest of this section. By the definition of MLE, we have that

$$\sum_{ij} \log f_{\hat{\theta}_{ij}}(y_{ij}) \geq \sum_{ij} \log f_{\theta_{ij}}(y_{ij}) \qquad (25)$$

for any θ. Since $-\log F$ is strongly convex function with respect to Θ, we then have

$$-\log F(\Theta^*) - \langle \nabla \log F(\Theta^*), \hat{\Theta} - \Theta^* \rangle + \mu \|\hat{\Theta} - \Theta^*\|^2 \leq -\log F(\hat{\Theta}) \leq -\log F(\Theta^*)$$

for some constant μ. Therefore,

$$\mu\|\hat{\Theta} - \Theta^*\|^2 \leq \langle \nabla \log F(\Theta^*), \hat{\Theta} - \Theta^* \rangle. \qquad (26)$$

Define $\bar{\theta}_{ab} = \arg\max \sum_{\hat{z}_{1i}=a, \hat{z}_{2j}=b} \mathbb{E} \log F(y_{ij}, \theta)$ and let $\bar{\Theta}_{ij} = \bar{\theta}_{\hat{z}_{1i}\hat{z}_{2j}}$. We consider to bound the difference between $\hat{\theta}_{ab}$ and $\bar{\theta}_{ab}$. By the definition of $\hat{\theta}$,

$$\hat{\theta}_{ab} = \arg\max \sum_{\hat{z}_{1i}=a, \hat{z}_{2j}=b} \log f_\theta(y_{ij}).$$

Let $n_1(a) = |\{i : \hat{z}_{1i} = a\}|$ and $n_2(b) = |\{j : \hat{z}_{2j} = b\}|$. We then know that $\Delta_{ab} := \sqrt{n_1(a)n_2(b)}(\hat{\theta}_{ab} - \bar{\theta}_{ab})$ is $O_p(1)$, that is,

$$\mathbb{E}\exp\{\Delta_{ab}^2\} \leq \exp\{C_1\}$$

for some constant C_1 and any fixed latent assignment z. This is because that

$$\begin{aligned} 0 = \frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla \log f_{\hat{\theta}_{ab}}(y_{ij}) \quad &= \quad \frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla \log f_{\bar{\theta}_{ab}}(y_{ij}) \\ &\quad + \frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij})(\hat{\theta}_{ab} - \bar{\theta}_{ab}) \end{aligned}$$

by Taylor expansion at $\bar{\theta}_{ab}$. Therefore, we have

$$\begin{aligned} \sqrt{n_1(a)n_2(b)}(\hat{\theta}_{ab} - \tilde{\theta}_{ab}) \quad &= \quad (\frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij}))^{-1} \\ &\quad \cdot \frac{1}{\sqrt{n_1(a)n_2(b)}} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla \log f_{\tilde{\theta}_{ab}}(y_{ij}). \end{aligned} \tag{27}$$

Define an event $\Omega_e = \{|\frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij}) - \frac{1}{n_1(a)n_2(b)}\mathbb{E} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij})| > \epsilon\}$, we know that

$$P(\Omega_e) \leq \exp\{-\frac{1/2n_a n_b \epsilon^2}{v^2}\}$$

by Bernstein inequality, where $v^2 := \max_{ij} \max_\theta \mathbb{E}\nabla^2 \log f_\theta(y_{ij})$.

Let ϵ be $\kappa/2$ with $\kappa := \min_{ij} \min_\theta \mathbb{E} \sum_{z_{1i}=a,z_{2j}=b} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij})$, which is a positive constant. Take small t such that $t \leq \kappa^2/(8M^2v^2)$ we have

$$\begin{aligned} \mathbb{E}\exp\{t\Delta_{ab}^2\} \quad &= \quad \mathbb{E}\{\exp\{t\Delta_{ab}^2\}\mathbf{1}_{\Omega_e}\} + \mathbb{E}\{\exp\{t\Delta_{ab}^2\}\mathbf{1}_{\Omega_e^c}\} \\ &\leq \quad \exp\{-\frac{1/8n_a n_b \epsilon^2}{v^2}\}\exp\{tn_a n_b M^2\} + \mathbb{E}\{\exp\{t\Delta_{ab}^2\}\mathbf{1}_{\Omega_e^c}\} \\ &\leq \quad 1 + \mathbb{E}\{\exp\{t\Delta_{ab}^2\}\mathbf{1}_{\Omega_e^c}\} \\ &\leq \quad 1 + \mathbb{E}\{\exp\{t\frac{2}{\kappa}(\frac{1}{\sqrt{n_1(a)n_2(b)}} \sum_{z_{1i}=a,z_{2j}=b} \nabla \log f_{\bar{\theta}_{ab}}(y_{ij}))^2\} && (28) \\ &\leq \quad 1 + \exp\{C\} \leq \exp C_1, && (29) \end{aligned}$$

by adjusting the constant C_1 and noticing that $\nabla \log f_{\bar{\theta}_{ab}}(y_{ij})$'s are conditionally independent for different i and j given latent memberships. Here (28) uses tower property and the fact that there exist t and C such that

$$\mathbb{E}\exp\{t\frac{1}{m}(\sum_{l=1}^{m} X_l)^2\} \leq \exp C,$$

where X_l's are m independent random variables with mean 0 and bounded variance. Therefore,

$$\|\hat{\Theta} - \bar{\Theta}\|^2 = \sum_{a,b} \sum_{i,j:\hat{z}_{1i}=a,\hat{z}_{2j}=b} \Delta_{ab}^2 = \sum_{a,b} n_1(a)n_2(b)\Delta_{ab}^2.$$

Thus,

$$\begin{aligned} & P(|\|\hat{\Theta} - \bar{\Theta}\|_F^2| > C(J_1J_2 + m_1 \log J_1 + m2 \log J_2)) \\ \leq \quad & \exp\{-tC(J_1J_2 + m_1 \log J_1 + m2 \log J_2)\} \exp\{C_1 J_1 J_2\} J_1^{m_1} J_2^{m_2} \\ \leq \quad & \exp\{-(tC - C_1)(J_1J_2 + m_1 \log J_1 + m2 \log J_2)\}, \end{aligned} \tag{30}$$

which goes to zero by choosing sufficiently large C.

Next, we consider to bound $|\langle \hat\Theta - \bar\Theta, \nabla \log F(\Theta^*)\rangle|$,

$$\begin{aligned} & |\langle \hat\Theta - \bar\Theta, \nabla \log F(\Theta^*)\rangle| \\ = \;& |\sum_{a,b} (\hat\theta_{ab} - \bar\theta_{ab})(\sum_{i,j:\hat z_{1i}=a,\hat z_{2j}=b} \nabla \log f_{\theta^*_{z^*_{1i}z^*_{2j}}}(y_{ij}))|. \end{aligned}$$

Similarly, we know that $H_{ab} := (\hat\theta_{ab} - \bar\theta_{ab})(\sum_{i,j:\hat z_{1i}=a,\hat z_{2j}=b} \nabla \log f_{\theta^*_{z^*_{1i}z^*_{2j}}}(y_{ij}))$ is $O_p(1)$, that is,

$$\mathbb{E}\exp\{|H_{ab}|\} \leq \exp\{C_2\}. \tag{31}$$

This is because $H_{ab} = \sqrt{n_1(a)n_2(b)}(\hat\theta_{ab} - \bar\theta_{ab})(\frac{1}{\sqrt{n_1(a)n_2(b)}}\sum_{i,j:\hat z_{1i}=a,\hat z_{2j}=b} \nabla \log f_{\theta^*_{z^*_{1i}z^*_{2j}}}(y_{ij}))$, which is upper bounded by $\frac{1}{2}\{(\sqrt{n_1(a)n_2(b)}(\hat\theta_{ab} - \bar\theta_{ab}))^2 + (\frac{1}{\sqrt{n_1(a)n_2(b)}}\sum_{i,j:\hat z_{1i}=a,\hat z_{2j}=b} \nabla \log f_{\theta^*_{z^*_{1i}z^*_{2j}}}(y_{ij}))^2\}$. Then (31) holds by using the same technique in proving Δ_{ab}.

Therefore we have

$$\begin{aligned} & P(|\langle \hat\Theta - \bar\Theta, \nabla \log F(\Theta^*)\rangle| > C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)) \\ \leq \;& \exp\{-C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)\}\exp\{C_2J_1J_2\}J_1^{m_1}J_2^{m_2} \\ \leq \;& \exp\{-(C - C_2)(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)\} \end{aligned}$$

which goes to zero for large C.

In addition, we consider to bound the quantity $\langle (\bar\Theta - \Theta^*)/\|\bar\Theta - \Theta^*\|, \nabla \log F(\Theta^*)\rangle$ in the case when $\|\bar\Theta - \Theta^*\|^2 \geq C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)$. Then, each entry of $(\bar\Theta - \Theta^*)/\|\bar\Theta - \Theta^*\|$ is bounded by $M/\sqrt{C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)}$. By the fact that

$$P(|\sum c_iX_i| \geq t) \leq C\exp\{-\min\{\frac{t^2}{B_1}, \frac{t}{B_2\|c\|_\infty}\}\},$$

for any sequence $\{c_i\}$ with $\sum_i c_i^2 = 1$, we arrive at

$$\begin{aligned} & P(|\langle (\bar\Theta - \Theta^*)/\|\bar\Theta - \Theta^*\|, \nabla \log F(y, \Theta^*)\rangle| \geq C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)^{1/2}) \\ \leq \;& \exp\{-\frac{C^2}{M\max\{B_1, B_2\}}(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)\}J_1^{m_1}J_2^{m_2} \\ \leq \;& \exp\{-\tilde C(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)\}, \end{aligned}$$

by adjusting constant $\tilde C$. The right hand side of above inequality goes to zero for large constant C.

Combining all above facts, from inequality (26), we know

$$\begin{aligned} \mu\|\Theta^{MLE} - \Theta^*\|^2 \;&\leq\; \langle \nabla \log F(\Theta^*), \Theta^{MLE} - \Theta^*\rangle \\ &\leq\; \langle \nabla \log F(\Theta^*), \Theta^{MLE} - \bar\Theta\rangle + \langle \nabla \log F(\Theta^*), \bar\Theta - \Theta^*\rangle \\ &\leq\; \langle \nabla \log F(\Theta^*), \Theta^{MLE} - \bar\Theta\rangle + |(\|\Theta^{MLE} - \Theta^*\| + \|\bar\Theta - \Theta^{MLE}\|) \\ &\quad\; \langle \nabla \log F(\Theta^*), (\bar\Theta - \Theta^*)/\|\bar\Theta - \Theta^*\|\rangle| \\ &\leq\; C_1(J_1J_2 + m_1 \log J_1 + m_2 \log J_2) + \frac{\mu}{2}\|\Theta^{MLE} - \Theta^*\|^2 + \frac{2}{\mu}(C_1(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)) \end{aligned}$$

when $\|\bar\Theta - \Theta^*\| \geq C\sqrt{J_1J_2 + m_1 \log J_1 + m_2 \log J_2}$. Additionally,

$$\|\Theta^{MLE} - \Theta^*\| \leq \|\Theta^{MLE} - \bar\Theta\| + \|\bar\Theta - \Theta^*\| \leq 2C\sqrt{J_1J_2 + m_1 \log J_1 + m_2 \log J_2} \tag{32}$$

when $\|\bar\Theta - \Theta^*\| \leq C\sqrt{J_1J_2 + m_1 \log J_1 + m_2 \log J_2}$.

Finally, we arrive at

$$\|\Theta^{MLE} - \Theta^*\|^2 \leq C'(J_1J_2 + m_1 \log J_1 + m_2 \log J_2) \tag{33}$$

for some constant C'. Combining (24) and (33), we then have

$$\|\hat{\Theta} - \Theta^{MLE}\|_F^2 \le Cm_1m_2 \max\{J_1, J_2\}(\exp\{-c'm_1\} + \exp\{-c'm_2\}) + C'(J_1J_2 + m_1 \log J_1 + m_2 \log J_2).$$

This completes the proof.

Proof of Theorem 5 In the partial observation case, we let $\Omega_{1i} = \{j : (i, j) \in \Omega\}$ and $\Omega_{2j} = \{i : (i, j) \in \Omega\}$. Therefore, by Hoeffding's inequality, we know that

$$\frac{pm_2}{2} \le |\Omega_{1i}| \le \frac{3pm_2}{2}, \quad \frac{pm_1}{2} \le |\Omega_{2j}| \le \frac{3pm_1}{2} \tag{34}$$

hold for all $i \in [m_1], j \in [m_2]$ with probability at least $1 - m_1 \exp\{-p(1-p)m_2/2\} - m_2 \exp\{-p(1-p)m_1/2\}$. Then the proofs for classification consistency, population consistency and variational estimator consistency are all the same.

For the upper bound of variational parameter, we only need to derive the parallel formula for (27) in estimating the upper bound of MLE. By calculation, we have

$$\begin{aligned}
\sqrt{n_1(a)n_2(b)}(\theta_{ab}^{MLE} - \bar{\theta}_{ab}) &= (\frac{1}{n_1(a)n_2(b)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij}))^{-1} \\
&\quad \cdot \frac{1}{\sqrt{n_1(a)n_2(b)}} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla \log f_{\bar{\theta}_{ab}}(y_{ij}) \\
&= \sqrt{\frac{n_1(a)n_2(b)}{n(a,b,\Omega)}} \left(\frac{1}{n(a,b,\Omega)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla^2 \log f_{\breve{\theta}_{ab}}(y_{ij})\right)^{-1} \\
&\quad \cdot \frac{1}{\sqrt{n(a,b,\Omega)}} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla \log f_{\tilde{\theta}_{ab}}(y_{ij}),
\end{aligned}$$

where $n(a, b, \Omega) := |(i, j) \in \Omega : \hat{z}_{1i} = a, \hat{z}_{2j} = b|$. Define events

$$\Omega_e = \{|\frac{1}{n(a,b,\Omega)} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla^2 \log f_{\breve{\theta}_{ij}}(y_{ij}) - \frac{1}{n(a,b,\Omega)} \mathbb{E} \sum_{\hat{z}_{1i}=a,\hat{z}_{2j}=b,(i,j)\in\Omega} \nabla^2 \log f_{\breve{\theta}_{ij}}(y_{ij})| > \epsilon\}$$

and $\Omega_n = \{|n(a, b, \Omega) - n_1(a)n_2(b)p| > 1/2n_1(a)n_2(b)p\}$. We then know that

$$P(\Omega_n) \le \exp\{-\frac{1}{16}\frac{n_1(a)n_2(b)p^2}{p(1-p)}\} \le \exp\{-\frac{1}{16}n_1(a)n_2(b)p\}$$

and hence get

$$\begin{aligned}
\mathbb{E}\exp\{pt\Delta_{ij}^2\} &= \mathbb{E}\{\exp\{pt\Delta_{ij}^2\}\mathbf{1}_{\Omega_n}\} + \mathbb{E}\{\exp\{pt\Delta_{ij}^2\}\mathbf{1}_{\Omega_e\cap\Omega_n^c}\} + +\mathbb{E}\{\exp\{pt\Delta_{ij}^2\}\mathbf{1}_{\Omega_e^c\cap\Omega_n^c}\} \\
&\le \exp\{-\frac{1}{16}n_1(a)n_2(b)p\} \exp\{ptn_1(a)n_2(b)M^2\} \\
&\quad + \exp\{-\frac{\kappa^2}{8v^2}n_1(a)n_2(b)p\} \exp\{ptn_1(a)n_2(b)M^2\} + \exp\{C\} \\
&\le \exp\{C_1'\}
\end{aligned} \tag{35}$$

by adjusting constant t and C_1'. Then, by repeating the same procedure in (30) - (32), we get

$$\|\Theta^* - \Theta^{MLE}\|_F^2 \le C'(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)/p \tag{36}$$

held with probability at least $1 - \exp\{C^{''}(J_1J_2 + m_1 \log J_1 + m_2 \log J_2)\}$. By repeating (22) - (24), we get

$$\|\hat{\Theta} - \Theta^{MLE}\|_F^2 \le 2C(m_2m_2(\exp\{-c'pm_1\} + \exp\{-c'pm_2\})). \tag{37}$$

Therefore, we have

$$\begin{array}{ll} & \|\hat{\Theta}-\Theta^*\|_F^2 \\ \leq & 2C\max\{J_1,J_2\}(m_2m_2(\exp\{-c'pm_1\}+\exp\{-c'pm_2\}))+C'(J_1J_2+m_1\log J_1+m_2\log J_2)/p \end{array}$$

with probability at least $1-\exp\{C^{''}(J_1J_2+m_1\log J_1+m_2\log J_2)\}-m_1\exp\{-p^2m_2/2\}-m_2\exp\{-p^2m_1/2\}$.

Proof of Theorem 4 From Yu, 1997 and Guntuboyina, 2011, we have the following generalized Fano's lemma.

Lemma 1 *Let $(\mathbf{\Theta},\ell)$ be a metric space and $\{\mathcal{P}_\theta:\theta\in\mathbf{\Theta}\}$. For any totally bounded set $\mathcal{T}\subset\mathbf{\Theta}$, define the Kullback-Leibler diameter and the chi-squared diameter of $\mathcal{T}$ by*

$$d_{KL}(\mathcal{T}):=\sup_{\theta_1,\theta_2\in\mathcal{T}}KL(P_{\theta_1}\|P_{\theta_2}),\ d_{\chi^2}(\mathcal{T}):=\sup_{\theta_1,\theta_2\in\mathcal{T}}\chi^2(P_{\theta_1}\|P_{\theta_2}).$$

Then, it holds

$$\inf_{\hat{\theta}}\sup_{\theta\in\mathbf{\Theta}}P(\ell^2(\hat{\theta},\theta)\geq\frac{\epsilon^2}{4})\geq 1-\frac{d_{KL}(\mathcal{T})+\log 2}{\log(M(\epsilon,\mathcal{T},\ell))},\tag{38}$$

$$\inf_{\hat{\theta}}\sup_{\theta\in\mathbf{\Theta}}P(\ell^2(\hat{\theta},\theta)\geq\frac{\epsilon^2}{4})\geq 1-\frac{1}{M(\epsilon,\mathcal{T},\ell)}-\sqrt{\frac{d_{\chi^2}(\mathcal{T})}{M(\epsilon,\mathcal{T},\ell)}}.\tag{39}$$

Here packing number $M(\epsilon,\mathcal{T},\ell)$ is the largest number of points in $\mathcal{T}$ such that they are at least ϵ away from each other.

In addition, by Lemma 4.7 in Massart, 2007, we have the following results.

Lemma 2 *There exists a subset $\{w_!,\ldots,w_N\}\subset\{0,1\}^d$ such that*

$$H(w_i,w_j):=\|w_i-w_j\|^2\geq\frac{d}{4},\quad \textit{for any } i\neq j\in[N],\tag{40}$$

for some $N\geq\exp\{d/8\}$.

By Assumption A4, we know that $KL(P_{\theta_1}\|P_{\theta_2})\leq C\|\theta_1-\theta_2\|^2$ holds for certain constant C. For example, we can compute the KL-divergence explicitly for special cases. For Poisson model, we have that

$$\begin{aligned} KL(P_{\theta_1}\|P_{\theta_2}) &= \theta_1\log(\theta_1/\theta_2)+\theta_2-\theta_1 \\ &= \frac{\theta_1}{2\xi^2}(\theta_1-\theta_2)^2 \\ &\leq \frac{c_u}{2c_l^2}(\theta_1-\theta_2)^2, \end{aligned}$$

where ξ is between θ_1 and θ_2. For Bernoulli distributions,

$$\begin{aligned} KL(P_{\theta_1}\|P_{\theta_2}) &= p_1\log\frac{p_1}{p_2}+(1-p_1)\log\frac{1-p_1}{1-p_2} \\ &= 1/2(\frac{p_1}{\xi^2}+\frac{1-p_1}{(1-\xi)^2})(p_1-p_2)^2 \\ &\leq \frac{1-\delta}{\delta^2}(p_1-p_2)^2 \\ &\leq \frac{1-\delta}{\delta^2}p'(1-p')(\theta_1-\theta_2)^2 \\ &\leq 4\frac{1-\delta}{\delta^2}(\theta_1-\theta_2)^2. \end{aligned}$$

We next consider to construct subspace of $\mathcal{B}_\Theta$ as follows.

Bound 1. Without loss of generality, we assume both m_1/J_1 and m_2/J_2 are integers. Consider the latent configuration, $z_{1i} = \lceil iJ_1/m_1 \rceil$ and $z_{2j} = \lceil jJ_2/m_2 \rceil$. For any $\omega \in \{0,1\}^{J_1 \times J_2}$, define parameter θ by letting

$$\theta^{\omega}_{ab} = 1 + \sqrt{\frac{J_1 J_2}{m_1 m_2}} \omega_{ab}. \tag{41}$$

There exists a subset $\mathcal{T} \in \{0,1\}^{J_1 \times J_2}$ such that $|\mathcal{T}| \geq \exp\{J_1 J_2/8\}$ and $H(\omega_1, \omega_2) \geq J_1 J_2/4$ for any $\omega_1 \neq \omega_2 \in \mathcal{T}$. We then construct

$$\boldsymbol{\Theta}(\mathcal{T}) := \{\Theta \mid \Theta_{ij} = \theta^{\omega}_{z_{1i} z_{2j}}, \omega \in \mathcal{T}\}.$$

Therefore, for any two different Θ and Θ' (associated with ω and ω') in $\boldsymbol{\Theta}(\mathcal{T})$, we have

$$\|\Theta - \Theta'\|_F^2 \geq H(\omega, \omega') \geq J_1 J_2/4.$$

Then, $M(\epsilon, \boldsymbol{\Theta}(\mathcal{T}), \|\cdot\|_F) \geq \exp\{J_1 J_2/8\}$ with $\epsilon = \sqrt{J_1 J_2/4}$. By using Lemma 1, we then have the minimax rate is at least $cJ_1 J_2$ by adjusting constant c.

Bound 2. We can pick $\omega_1, \ldots, \omega_{J_2} \in \{0,1\}^{J_1}$ such that $H(\omega_a, \omega_b) \geq J_1/4$ for all $a \neq b$. This is possible when $\exp\{J_1/8\} \geq J_2$. We then define

$$\theta^{\omega}_{ka} = 1 + \sqrt{\frac{m_2 \log J_2}{m_1 m_2}} \omega_{ka} \tag{42}$$

for all $k \in [J_1]$. Consider configuration $z_{1i} = \lceil iJ_1/m_1 \rceil$. Next, we fix $\mathbf{z}_1$ and θ. We can choose $\mathcal{Z} \subset [J_2]^{m_2}$ such that $|\mathcal{Z}| \geq \exp\{Cm_2 \log J_2\}$ and $H(\mathbf{z}_a, \mathbf{z}_b) \geq m_2/6$ for any $\mathbf{z}_a \neq \mathbf{z}_b \in \mathcal{Z}$. Then, the subspace is constructed as

$$\boldsymbol{\Theta}(\mathcal{T}) := \{\Theta \mid \Theta_{ij} = \theta^{\omega}_{z_{1i} z_{2j}}, \mathbf{z}_2 \in \mathcal{Z}\}.$$

For any Θ and $\Theta^{'}$ in $\boldsymbol{\Theta}(\mathcal{T})$, we can see that Θ and $\Theta^{'}$ have at least $m_2/6$ different columns. For two different columns, there are at least $J_1/4 \cdot m_1/J_1$ elements differ. Thus

$$\|\Theta - \Theta^{'}\|_F^2 \geq \frac{m_2}{6} \frac{m_1}{4} \frac{m_2 \log J_2}{m_1 m_2} \geq m_2 \log J_2/24$$

for any $\Theta, \Theta' \in \boldsymbol{\Theta}(\mathcal{T})$. By using Lemma 1 again, we then have the lower bound is at least $c(m_2 \log J_2)$ by adjusting constant c. We can repeat the same procedure to get another lower bound $c(m_1 \log J_1)$. Combining all results, we know that

$$\inf_{\check{\Theta}} \sup_{\Theta^* \in \mathcal{B}_\Theta} P(\|\check{\Theta} - \Theta^*\|_F^2 \geq c(m_2 \log J_2 + m_1 \log J_1 + J_1 J_2)) \geq c'$$

by adjusting the constant c and c'. This concludes the proof.

3. Proof for Convergence Results

3.1. Local Convergence

Proof of Theorem 6 By the update rule, we know that

$$\begin{aligned}\theta_{kl}^{(1)} &= \arg\max_{\theta} \sum_{i,j} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}) \\ &= \arg\max_{\theta} \sum_{k',l'} \Big\{ \sum_{i,j:z_{1i}^*=k', z_{2j}^*=l'} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}) \Big\}.\end{aligned}$$

By Condition I1 that

$$\frac{\sum_{i:z_{1i}^*=k} \phi_{1i}^0[k]}{\sum_{k'\neq k}\sum_{i:z_{1i}^*=k} \phi_{1i}^0[k']} > D_1 \text{ and } \frac{\sum_{j:z_{2j}^*=l} \phi_{2j}^0[l]}{\sum_{l'\neq l}\sum_{i:z_{1i}^*=l} \phi_{2j}^0[l']} > D_2,$$

we get that

$$\sum_{i,j:z_{1i}^*=k, z_{2j}^*=l} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}) \geq c\min\{D_1, D_2\} \sum_{i,j:z_{1i}^*\neq k \text{ or } z_{2j}^*\neq l} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}). \tag{43}$$

for some constant c. Note that the left hand side of above equation is bounded by $m_1 m_2 C_f$ (C_f is defined to be the upper bound of $\log f_\theta(y_{ij})$), thus

$$\frac{1}{m_1 n_1} \sum_{i,j:z_{1i}^*\neq k \text{ or } z_{2j}^*\neq l} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}) \leq \frac{C_f}{c\min\{D_1, D_2\}}.$$

For the notational simplicity, we write

$$W(\theta) = \frac{1}{m_1 m_2} \sum_{k',l'} \Big\{ \sum_{i,j:z_{1i}^*=k', z_{2j}^*=l'} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij}) \Big\}$$

and

$$W_1(\theta) = \frac{1}{m_1 m_2} \sum_{i,j:z_{1i}^*=k, z_{2j}^*=l} \phi_{1i}^{(0)}[k]\phi_{2j}^{(0)}[l] \log f_\theta(y_{ij})$$

in the rest of the proof. Therefore, $|W_1(\theta) - W(\theta)| \leq \frac{C_f}{c\min\{D_1,D_2\}}$. We next define $\tilde{\theta}_{kl}$ as $\arg\max_\theta W_1(\theta)$. By algebraic calculation and concavity of $W_1(\theta)$, we know that there exists constant ρ' such that

$$\begin{aligned}\rho'(\theta_{kl}^{(1)} - \tilde{\theta}_{kl})^2 &\leq W_1(\tilde{\theta}_{kl}) - W_1(\theta_{kl}^{(1)}) \\ &= W_1(\tilde{\theta}_{kl}) - W_1(\theta_{kl}^{(1)}) + W(\tilde{\theta}_{kl}) - W(\tilde{\theta}_{kl}) + W(\theta_{kl}^{(1)}) - W(\theta_{kl}^{(1)}) \\ &\leq W_1(\tilde{\theta}_{kl}) - W_1(\theta_{kl}^{(1)}) - W(\tilde{\theta}_{kl}) + W(\theta_{kl}^{(1)}) \\ &\leq |W_1(\tilde{\theta}_{kl}) - W(\tilde{\theta}_{kl})| + |W_1(\theta_{kl}^{(1)}) - W(\theta_{kl}^{(1)})| \\ &\leq \frac{2C_f}{c\min\{D_1, D_2\}}.\end{aligned}$$

Next we consider to bound the difference between $\tilde{\theta}_{kl}$ and θ_{kl}^*. It is easy to see that θ_{kl}^* is the maximizer of $\mathbb{E}W_1(\theta)$. Again, we can find that

$$\rho^{''}(\tilde{\theta}_{kl} - \theta_{kl}^*)^2 \leq \mathbb{E}W_1(\theta_{kl}^*) - \mathbb{E}W_1(\tilde{\theta}_{kl}) \leq 2\sup_\theta |W_1(\theta) - \mathbb{E}W_1(\theta)|,$$

where $\sup_\theta |W_1(\theta) - \mathbb{E}W_1(\theta)| = O_p(\frac{J_1 J_2 + m_1 \log J_1 + m_2 \log J_2}{\sqrt{m_1 m_2}})$ by concentration inequality (19). Therefore, $|\theta_{kl}^{(1)} - \theta_{kl}^*| \leq \frac{2C_f}{c\min\{D_1,D_2\}} + o_p(1)$.

By update rule, we have that

$$\phi_{1i}^{(1)}[k] \propto \exp\{\{\sum_{l\in[J_2]}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[l]\log f_{\theta_{kl}^{(1)}}(y_{ij})\} + \pi_1^{(1)}[k]\}$$

for each fixed k. We then compare $\phi_{1i}^{(1)}[k]$ and $\phi_{1i}^{(1)}[z_{1i}^*]$ ($k \neq z_{1i}^*$), that is,

$$\begin{aligned}
& \log(\phi_{1i}^{(1)}[z_{1i}^*]) - \log(\phi_{1i}^{(1)}[k]) \\
= \; & \sum_{l\in[J_2]}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[l]\log f_{\theta_{z_{1i}^* l}^{(1)}}(y_{ij}) + \pi_1^{(1)}[z_{1i}^*] - \sum_{l\in[J_2]}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[l]\log f_{\theta_{kl}^{(1)}}(y_{ij})\} - \pi_1^{(1)}[k] \\
\geq \; & \sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{z_{1i}^* z_{2j}^*}^{(1)}}(y_{ij}) - \sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{k z_{2j}^*}^{(1)}}(y_{ij}) - 2m_2\frac{C_f}{\min\{D_1, D_2\}} - 2 \\
\geq \; & c_0 m_2 - 2m_2\frac{C_f}{\min\{D_1, D_2\}} - 2. \qquad (44)
\end{aligned}$$

Here (44) uses the fact that

$$\begin{aligned}
& \sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{z_{1i}^* z_{2j}^*}^{(1)}}(y_{ij}) - \sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{k z_{2j}^*}^{(1)}}(y_{ij}) \\
\geq \; & |\mathbb{E}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{z_{1i}^* z_{2j}^*}^{(1)}}(y_{ij}) - \mathbb{E}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_{\theta_{k z_{2j}^*}^{(1)}}(y_{ij})| - 2D \\
\geq \; & 2c_0 m_2 - O_p(\sqrt{m_2}) \geq c_0 m_2
\end{aligned}$$

by using continuity and concentration bound for D, where

$$D := \sup_\theta |\sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_\theta(y_{ij}) - \mathbb{E}\sum_{j\in[m_2]} \phi_{2j}^{(0)}[z_{2j}^*]\log f_\theta(y_{ij})|.$$

Hence, we know that

$$\phi_{1i}[z_{1i}^*] \geq 1 - J_1\exp\{c_0' m_2\},$$

when D_1 and D_2 are large enough. Similarly, we can obtain that

$$\phi_{2j}[z_{2j}^*] \geq 1 - J_2\exp\{c_0' m_1\}$$

for $j \in [m_2]$ as well. For iteration $t \geq 2$, repeating the previous procedure we know that $\theta^{(t)}$ is always in the local neighborhood of θ^*. Thus the estimated latent class memberships $\hat{z}_{1i}$ and $\hat{z}_{2j}$ are consistent. Then the consistency of $\hat{\theta}$ follows as well.

Proof of Theorem 7 is similar to that of Theorem 6. Hence we omit it here.

3.2. Global Convergence

In the next, we consider the global convergence of the algorithm. Recall that the initialization for ϕ_{1i}'s and ϕ_{2j}'s are

$$\phi_{1i} \sim \text{Dir}(\boldsymbol{\alpha}_1) \;\text{ and }\; \phi_{2i} \sim \text{Dir}(\boldsymbol{\alpha}_2), \qquad (45)$$

where $\boldsymbol{\alpha}_1$ is a vector of length J_1 with all entries being 1 and $\boldsymbol{\alpha}_2$ is a vector of length J_2 with all entries being 1. (Remark: (45) can be replaced by other non-informative priors, i.e., $\phi_{1i}[k]$'s ($k \in [J_1]$) have the same marginal distribution.) We first consider the degenerate case, i.e., $J_2 = 1$, then the biclustering model reduces to the latent class model.

Case $J_2 = 1$: We first show that the algorithm can return global optimum with high probability for arbitrary true model parameters when $J_1 = 2$ and $J_2 = 1$ with $m_2 \gg \log m_1$.

Proof of Theorem 8 We first write $\phi^{(0)}_{1i} = \frac{1}{2} + \tilde{\phi}^{(0)}_{1i}$ and $\phi^{(0)}_{2j} = \frac{1}{2} + \tilde{\phi}^{(0)}_{2j}$. Therefore, we know $\mathbb{E}\tilde{\phi}^{(0)}_{1i} = \mathbb{E}\tilde{\phi}^{(0)}_{2j} = 0$. Without loss of generality, we only prove for Gaussian / Bernoulli / Poisson model, then we have

$$\theta^{(1)}_{11} = \frac{\frac{m_1m_2}{2}\bar{y} + \sum_i \tilde{\phi}^{(0)}_{1i} y_{i\cdot}}{\frac{m_1m_2}{2} + m_2\sum_i \tilde{\phi}^{(0)}_{1i}} \text{ and } \theta^{(1)}_{21} = \frac{\frac{m_1m_2}{2}\bar{y} - \sum_i \tilde{\phi}^{(0)}_{1i} y_{i\cdot}}{\frac{m_1m_2}{2} - m_2\sum_i \tilde{\phi}^{(0)}_{1i}}, \tag{46}$$

where $y_{i\cdot} = \sum_j y_{ij}$ and $\bar{y} = \sum_{ij} y_{ij}/(m_1m_2)$. In addition, we let $a = (\sum_i \tilde{\phi}_{1i} y_{i\cdot})/(\frac{m_1m_2}{2})$ and $b = (m_2 \sum_i \tilde{\phi}_{1i})/(\frac{m_1m_2}{2})$. Then, we can compute the difference

$$\begin{aligned}
\theta^{(1)}_{11} - \theta^{(1)}_{21} &= 2\frac{a - b\bar{y}}{1-b^2} \\
&= \frac{4}{1-b^2}\left(\frac{\sum_i \tilde{\phi}^{(0)}_{1i} y_{i\cdot}}{m_1m_2} - \frac{\bar{y}\sum_i \tilde{\phi}^{(0)}_{1i}}{m_1}\right) \\
&= \frac{4}{1-b^2}\left(\frac{\sum_i \tilde{\phi}^{(0)}_{1i}\theta^*_{z^*_{1i}1}}{m_1} - \frac{\bar{\theta}^*\sum_i \tilde{\phi}^{(0)}_{1i}}{m_1}(1 + O_p(\frac{1}{\sqrt{m_2}}))\right)) \\
&= \frac{4}{1-b^2}\left(\frac{\sum_i \tilde{\phi}^{(0)}_{1i}(\theta^*_{z^*_{1i}1} - \bar{\theta}^*)}{m_1} + \frac{1}{\sqrt{m_1m_2}}\right) \\
&= \frac{4}{1-b^2}\left(\frac{\sum_i \tilde{\phi}^{(0)}_{1i}(\theta^*_{z^*_{1i}1} - \bar{\theta}^*)}{m_1} + \frac{1}{\sqrt{m_1m_2}}\right).
\end{aligned}$$

By central limit theorem, we get

$$P(|\theta^{(1)}_{11} - \theta^{(1)}_{21}| \geq \frac{t}{\sqrt{m_1}}) \geq 2 - 2\Phi(ct)$$

for some universal constant c and any $t > 0$. That is, $|\theta^{(1)}_{11} - \theta^{(1)}_{21}| = O(\frac{1}{\sqrt{m_1}})$. In below, we show that $|\theta^{(t)}_{11} - \theta^{(t)}_{21}|$ is strictly increasing as long as $|\theta^{(t)}_{11} - \theta^{(t)}_{21}| = o(1)$. Thus, there exists t_c such that $|\theta^{(t_c)}_{11} - \theta^{(t_c)}_{21}| = \Omega(1)$. Then, at this time $d(q^{(t_c)}_{1i}, z^*_{1i}) = o(1)$. This further implies the parameter $\theta^{(t_c+1)}$ is consistent.

Without loss of generality, we can assume $\theta^{(1)}_{11} > \theta^{(1)}_{21}$. By update rule for $\phi^{(1)}_{1i}$, we can compute

$$\begin{aligned}
& \log(\phi^{(1)}_{1i}[1]) - \log(\phi^{(1)}_{1i}[2]) \\
= \; & \sum_j (\log f_{\theta^{(1)}_{11}}(y_{ij}) - \log f_{\theta^{(1)}_{21}}(y_{ij})) + \log \pi_1[1] - \log \pi_1[2] \\
= \; & m_2(\mathbb{E}_{y\sim\theta^*_{11}} \log f_{\theta^{(1)}_{11}}(y) - \mathbb{E}_{y\sim\theta^*_{11}} \log f_{\theta^{(1)}_{21}}(y) + O_p(\sqrt{\frac{\text{var}_{y\sim\theta^*_{11}}(\log f_{\theta^{(1)}_{11}}(y) - \log f_{\theta^{(1)}_{21}}(y))}{m_2}})) \\
\geq \; & cm_2(\theta^{(1)}_{11} - \theta^{(1)}_{21})
\end{aligned}$$

for any i with $z^*_{1i} = 1$ and some constant c. Similarly, we can get

$$\log(\phi^{(1)}_{1i}[2]) - \log(\phi^{(1)}_{1i}[1]) \geq cm_2(\theta^{(1)}_{11} - \theta^{(1)}_{21})$$

for any i with $z^*_{1i} = 2$. Therefore, we can write $\phi^{(1)}_{1i} = 1/2 + \delta_i$ for $i \in A_1$ and $\phi^{(1)}_{1i} = 1/2 - \delta_i$ for $i \in A_2$ such that δ_i's are all positive with $A_1 := \{i : z^*_{1i} = 1\}$ and $A_2 := \{i : z^*_{1i} = 2\}$.

For the second iteration, the parameters are updated as

$$\theta^{(2)}_{11} = \frac{\sum_{i\in A_1}(1+\delta_i)\bar{y}_{i\cdot} + \sum_{i\in A_2}(1-\delta_i)\bar{y}_{i\cdot}}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i\in A_2}(1-\delta_i)} \tag{47}$$

and

$$\theta^{(2)}_{21} = \frac{\sum_{i\in A_1}(1-\delta_i)\bar{y}_{i\cdot} + \sum_{i\in A_2}(1+\delta_i)\bar{y}_{i\cdot}}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i\in A_2}(1+\delta_i)},$$

where $\bar{y}_{i\cdot} = \sum_j y_{ij}/m_2$. Therefore, we can compute $\theta_{11}^{(2)} - \theta_{21}^{(2)}$ as

$$\begin{aligned}\theta_{11}^{(2)} - \theta_{21}^{(2)} &= \sum_{i \in A_1} \Big(\frac{1+\delta_i}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i \in A_2}(1-\delta_i)} - \frac{1-\delta_i}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i \in A_2}(1+\delta_i)}\Big)\bar{y}_{i\cdot} \\ &\quad + \sum_{i \in A_2} \Big(\frac{1-\delta_i}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i \in A_2}(1-\delta_i)} - \frac{1+\delta_i}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i \in A_2}(1+\delta_i)}\Big)\bar{y}_{i\cdot} \\ &\quad . \end{aligned} \tag{48}$$

To simplify the above equation, we consider the following

$$\begin{aligned} & \frac{1+\delta_i}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i \in A_2}(1-\delta_i)} - \frac{1-\delta_i}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i \in A_2}(1+\delta_i)} \\ = & \frac{1+\delta_i}{n + (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)} - \frac{1-\delta_i}{n - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)} \\ = & \frac{(1+\delta_i)(n - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)) - (1-\delta_i)(n + (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i))}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2} \\ = & 2\frac{n\delta_i - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2}. \end{aligned}$$

As we know that $\bar{y}_{i\cdot} = \theta_{11}^* + O_p(\frac{1}{\sqrt{m_2}})$, we get that

$$\begin{aligned} & \sum_{i \in A_1} \Big(\frac{1+\delta_i}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i \in A_2}(1-\delta_i)} - \frac{1-\delta_i}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i \in A_2}(1+\delta_i)}\Big)\bar{y}_{i\cdot} \\ = & 2\sum_{i\in A_1}\frac{n\delta_i - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2}(\theta_{11}^* + O_p(\frac{1}{\sqrt{m_2}})) \\ = & 2\frac{|A_2|\sum_{i\in A_1}\delta_i + |A_1|\sum_{i\in A_2}\delta_i}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2}(\theta_{11}^* + O_p(\frac{1}{\sqrt{m_2}})). \end{aligned}$$

Similarly, we get

$$\begin{aligned} & \sum_{i \in A_2} \Big(\frac{1-\delta_i}{\sum_{i\in A_1}(1+\delta_i) + \sum_{i \in A_2}(1-\delta_i)} - \frac{1+\delta_i}{\sum_{i\in A_1}(1-\delta_i) + \sum_{i \in A_2}(1+\delta_i)}\Big)\bar{y}_{i\cdot} \\ = & -2\frac{|A_2|\sum_{i\in A_1}\delta_i + |A_1|\sum_{i\in A_2}\delta_i}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2}(\theta_{21}^* + O_p(\frac{1}{\sqrt{m_2}})). \end{aligned}$$

From (48), we arrive at

$$\begin{aligned} \theta_{11}^{(2)} - \theta_{21}^{(2)} &= 2\frac{|A_2|\sum_{i\in A_1}\delta_i + |A_1|\sum_{i\in A_2}\delta_i}{n^2 - (\sum_{i\in A_1}\delta_i - \sum_{i\in A_2}\delta_i)^2}(\theta_{11}^* - \theta_{21}^* + O_p(\frac{1}{\sqrt{m_2}})) \\ &\geq (\theta_{11}^* - \theta_{21}^*)\delta_0\frac{|A_2||A_1|}{n^2} \\ &= (\theta_{11}^* - \theta_{21}^*)\delta_0\pi_1[1]\pi_1[2], \end{aligned}$$

where $\delta_0 = \min_i \delta_i$. Note that $\delta_i \geq \exp\{cm_2(\theta_{11}^{(1)} - \theta_{21}^{(1)})\}/(\exp\{cm_2(\theta_{11}^{(1)} - \theta_{21}^{(1)})\}) - 1/2$. When m_2 is sufficiently large, we have

$$\theta_{11}^{(2)} - \theta_{21}^{(2)} \geq C(\theta_{11}^* - \theta_{21}^*)\delta_0\pi_1[1]\pi_1[2] > \theta_{11}^{(1)} - \theta_{21}^{(1)},$$

where C is a large constant. Thus, by repeating this procedure, we will have that gap $\theta_{11}^{(t)} - \theta_{21}^{(t)}$ strictly increases as t increases. Thus, $q_{1i}^{(t)} \to \delta_{z_{1i}^*}$ for all $i \in [m_1]$. This gives the global convergence of the variational algorithm.

Proof of Theorem 9 When $\arg\max_\theta \sum_{i,j} \phi_{1i}^{(0)}[1] \log f_\theta(y_{ij}) = \arg\max_\theta \phi_{1i}^{(0)}[2] \log f_\theta(y_{ij})$ holds, this tells us that $\theta_{11}^{(1)} = \theta_{21}^{(1)}$. This further gives us that $\phi_{1i}^{(1)}[1] = \phi_{1i}^{(1)}[2] = \frac{1}{2}$. Thus, we have $\pi_1^{(0)}[1] = \pi_1^{(0)}[2] = 1/2$ as well. By induction, we can show that $\theta_{11}^{(t)} = \theta_{21}^{(t)}$ and $\phi_{1i}^{(t)}[1] = \phi_{1i}^{(t)}[2] = \frac{1}{2} = \pi_1^{(t)}[1] = \pi_1^{(t)}[2]$ for any $t > 1$. Thus the algorithm never converges to the global optimum.

On the other hand, when $\arg\max_\theta \sum_{i,j} \phi_{1i}^{(0)}[1] \log f_\theta(y_{ij}) = \arg\max_\theta \phi_{1i}^{(0)}[2] \log f_\theta(y_{ij})$ does not hold, we know that $\theta_{11}^{(1)} \neq \theta_{21}^{(1)}$. By using the same procedure in the proof of Theorem 8, we can show that $|\theta_{11}^{(t)} - \theta_{21}^{(t)}|$ strictly increases as t increases. Thus the algorithm will converge to the global optimal point.

On the other hand, when $J_1 \geq 3$, the global convergence result is different. Specifically, we consider the case that $J_1 = 3$.

Proof of Theorem 10 By update rule, we then can compute that

$$\theta_{11}^{(1)} = \frac{\sum_i \phi_i[1] y_{i\cdot}}{m_2 \sum_i \phi_i[1]} = \frac{\sum_i (1/3 + \tilde{\phi}_{1i}^{(0)}[1]) y_{i\cdot}}{m_2 \sum_i (1/3 + \tilde{\phi}_i[1])} = \frac{\frac{\bar{y}}{3} + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[1] \bar{y}_{i\cdot}}{1/3 + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[1]} := \frac{\bar{y} + a_1}{1 + b_1}, \tag{49}$$

where $\tilde{\phi}_{1i}^{(0)} = \phi_{1i}^{(0)} - 1/3$, $a_1 = \frac{3}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[1]\bar{y}_{i\cdot}$ and $b_1 = \frac{3}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[1]$. Similarly, we get

$$\theta_{21} = \frac{\frac{\bar{y}}{3} + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[2] \bar{y}_{i\cdot}}{1/3 + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[2]} := \frac{\bar{y} + a_2}{1 + b_2} \quad \text{and} \quad \theta_{31} = \frac{\frac{\bar{y}}{3} + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[3] \bar{y}_{i\cdot}}{1/3 + \frac{1}{m_1}\sum_i \tilde{\phi}_{1i}^{(0)}[3]} := \frac{\bar{y} + a_3}{1 + b_3}. \tag{50}$$

Then, we can get the difference

$$\begin{aligned} \theta_{11}^{(1)} - \theta_{21}^{(1)} &= \frac{\bar{y} + a_1}{1 + b_1} - \frac{\bar{y} + a_2}{1 + b_2} = \frac{a_1 - a_2 + (b_2 - b_1)\bar{y} + a_1 b_2 - a_2 b_1}{(1 + b_1)(1 + b_2)} \\ &= \frac{\sum_i (\tilde{\phi}_{1i}^{(0)}[1] - \tilde{\phi}_{1i}^{(0)}[2])(\theta^*_{z^*_{1i}1} - \bar{\theta})}{m_1(1 + b_1)(1 + b_2)} + O_p(\frac{1}{\sqrt{m_1 m_2}}). \end{aligned} \tag{51}$$

Similarly, we can get

$$\begin{aligned} \theta_{21}^{(1)} - \theta_{31}^{(1)} &= \frac{\sum_i (\tilde{\phi}_{1i}^{(0)}[2] - \tilde{\phi}_{1i}^{(0)}[3])(\theta^*_{z^*_{1i}1} - \bar{\theta})}{(1 + b_2)(1 + b_3)} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \\ &= \frac{\sum_i (3\tilde{\phi}_{1i}^{(0)}[2] + (\tilde{\phi}_{1i}^{(0)}[1] - \phi_{1i}^{(0)}[2])(\theta^*_{z^*_{1i}1} - \bar{\theta})}{m_1(1 + b_2)(1 + b_3)} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \end{aligned} \tag{52}$$

by using the relation $0 = \tilde{\phi}_{1i}^{(0)}[1] + \tilde{\phi}_{1i}^{(0)}[2] + \tilde{\phi}_{1i}^{(0)}[3]$. Moreover, we know that $\frac{1}{\sqrt{m_1}}\sum_i (\tilde{\phi}_{1i}^{(0)}[1] - \tilde{\phi}_{1i}^{(0)}[2])(\theta^*_{z^*_{1i}1} - \bar{\theta})$ converges to a normal distribution with mean 0 and $\frac{1}{\sqrt{m_1}}\sum_i \tilde{\phi}_{1i}^{(0)}[2](\theta^*_{z^*_{1i}1} - \bar{\theta})$ also converges to a normal distribution with mean 0. This fact will be used later. We further assume $\theta_{11}^{(1)} > \theta_{21}^{(1)} > \theta_{31}^{(1)}$.

For $\phi_{1i}^{(1)}$, we know that

$$\phi_{1i}^{(1)}[k] \propto \exp\{\sum_j \log f_{\theta_{k1}^{(1)}}(y_{ij}) + \log \pi_1^{(0)}[k]\}, \tag{53}$$

that is,

$$\frac{\phi_{1i}^{(1)}[k]}{\phi_{1i}^{(1)}[k']} = \exp\{\sum_j \{\log f_{\theta_{k1}^{(1)}}(y_{ij}) - \log f_{\theta_{k'1}^{(1)}}(y_{ij})\}\}. \tag{54}$$

Take i from Class 2 and i' from Class 3, we then know

$$
\begin{aligned}
\frac{\phi_{1i}^{(1)}[2]}{\phi_{1i}^{(1)}[1]}/\frac{\phi_{1i'}^{(1)}[2]}{\phi_{1i'}^{(1)}[3]} &= \exp\{\sum_j \{\log f_{\theta_{21}^{(1)}}(y_{ij}) + \log f_{\theta_{31}^{(1)}}(y_{i'j}) - \log f_{\theta_{11}^{(1)}}(y_{ij}) - \log f_{\theta_{21}^{(1)}}(y_{i'j})\}\} \\
&= \exp\{m_2\mathbb{E}[\log f_{\theta_{21}^{(1)}}(y_i) + \log f_{\theta_{31}^{(1)}}(y_{i'}) - \log f_{\theta_{11}^{(1)}}(y_i) - \log f_{\theta_{21}^{(1)}}(y_{i'})] + O_p(\sqrt{m_2 d_\theta})\} \\
&= \exp\{m_2(\nabla\mathbb{E}[\log f_{\tilde{\theta}_{21}^{(1)}}(y_i)](\theta_{21}^{(1)} - \theta_{11}^{(1)}) - \nabla\mathbb{E}[\log f_{\tilde{\theta}_{23}^{(1)}}(y_{i'})](\theta_{21}^{(1)} - \theta_{31}^{(1)})) + O_p(\sqrt{m_2 d_\theta})\} \\
&= \exp\{m_2(\nabla\mathbb{E}[\log f_{\bar{\theta}}(y_i)](\theta_{21}^{(1)} - \theta_{11}^{(1)}) - \nabla\mathbb{E}[\log f_{\bar{\theta}}(y_{i'})](\theta_{21}^{(1)} - \theta_{31}^{(1)})) + O_p(\sqrt{m_2 d_\theta})\}, \\
&= \exp\{m_2(g_2(\theta_{21}^{(1)} - \theta_{11}^{(1)}) - g_3(\theta_{21}^{(1)} - \theta_{31}^{(1)})) + O_p(\sqrt{m_2 d_\theta})\} \qquad (55)
\end{aligned}
$$

where $d_\theta = \max\{\|\theta_{21}^{(1)} - \theta_{31}^{(1)}\|, \|\|\theta_{21}^{(1)} - \theta_{11}^{(1)}\}$. Therefore, (55) depends on g_2 and g_3 (see the definition (27) in the main paper). Notice the fact that Class 3 becomes the dominate group in the second estimated group, if $\frac{\phi_{1i}^{(1)}[2]}{\phi_{1i}^{(1)}[1]}/\frac{\phi_{1i'}^{(1)}[2]}{\phi_{1i'}^{(1)}[3]} < 1$. In other words, the algorithm will converges to a local optimum, i.e.,

$$\hat{\theta}_{11} \to \bar{\theta}_{12},\ \hat{\theta}_{21} \to \theta_{31}^*,\ \hat{\theta}_{31} \to \theta_{31}^*,$$

where $\bar{\theta}_{12} := \arg\max_\theta \pi_1 \mathbb{E}_{y\sim f_{\theta_{11}^*}} \log f_\theta(y) + \pi_2 \mathbb{E}_{y\sim f_{\theta_{11}^*}} \log f_\theta(y)$. By formula (51) and (52), the asymptotic probability of $P(\theta_{11}^{(1)} - \theta_{21}^{(1)} > x(\theta_{21}^{(1)} - \theta_{31}^{(1)}))$ does not depend on true model parameter for any fixed x. To see this, we define

$$Z_1 := \frac{1}{\sqrt{m_1}} \sum_i \tilde{\phi}_{1i}^{(0)}[1](\theta_{z_{1i}^*1}^* - \bar{\theta}) \quad \text{and} \quad Z_1 := \frac{1}{\sqrt{m_1}} \sum_i \tilde{\phi}_{1i}^{(0)}[2](\theta_{z_{1i}^*1}^* - \bar{\theta}) \qquad (56)$$

By straightforward calculation,

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \to N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, N(\begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix})) \qquad (57)$$

in distribution, where $v_1 = v_2 = \mathbb{E}[(\tilde{\phi}_{1i}^{(0)}[1])^2]\mathbb{E}[(\theta_{z_{1i}1}^* - \bar{\theta})^2]$ and $v_{12} = \mathbb{E}[\tilde{\phi}_{1i}^{(0)}[1]\tilde{\phi}_{1i}^{(0)}[2]]\mathbb{E}[(\theta_{z_{1i}1}^* - \bar{\theta})^2]$. Thus, we know that $P(\theta_{11}^{(1)} - \theta_{21}^{(1)} > x(\theta_{21}^{(1)} - \theta_{31}^{(1)}))$ is free of model parameters for any fixed x as $m_1, m_2 \to \infty$.

Under constraint $\mathbb{E}_{y\sim f_{\theta_{21}^*}} \log f_{\bar{\theta}_{12}}(y) > \mathbb{E}_{y\sim f_{\theta_{21}^*}} \log f_{\theta_{31}^*}(y)$, the algorithm can never jump out of the local optimum value, once (55) is smaller than 1. Thus the failure probability is asymptotically at least

$$P(g_2 Z_1 + g_3 Z_2 > 0, Z_1 > 0, Z_2 > 0)$$

This completes the proof.

Proof of Theorem 11 We next consider the scenario when $J_1 = 2$ and $J_2 = 2$. Under this setting, we have

$$\theta_{11}^{(1)} = \frac{\frac{m_1 m_2}{4}\bar{y} + \frac{1}{2}\sum_i \tilde{\phi}_{1i} y_{i\cdot} + \frac{1}{2}\sum_j \tilde{\phi}_{2j} y_{\cdot j} + \sum_{i,j} \tilde{\phi}_{1i}\tilde{\phi}_{2j}}{\frac{m_1 m_2}{4} + \frac{m_2}{2}\sum_i \tilde{\phi}_{1i} + \frac{m_1}{2}\sum_i \tilde{\phi}_{2j} + \sum_{i,j} \tilde{\phi}_{1i}\tilde{\phi}_{2j}} \qquad (58)$$

and

$$\theta_{21}^{(1)} = \frac{\frac{m_1 m_2}{4}\bar{y} - \frac{1}{2}\sum_i \tilde{\phi}_{1i} y_{i\cdot} + \frac{1}{2}\sum_j \tilde{\phi}_{2j} y_{\cdot j} - \sum_{i,j} \tilde{\phi}_{1i}\tilde{\phi}_{2j}}{\frac{m_1 m_2}{4} - \frac{m_2}{2}\sum_i \tilde{\phi}_{1i} + \frac{m_1}{2}\sum_i \tilde{\phi}_{2j} - \sum_{i,j} \tilde{\phi}_{1i}\tilde{\phi}_{2j}}, \qquad (59)$$

where $y_{i\cdot} = \sum_j y_{ij}$ and $y_{\cdot j} = \sum_i y_{ij}$. We write $a_1 = (\sum_i \tilde{\phi}_{1i} y_{i\cdot})/(\frac{m_1 m_2}{2})$, $a_2 = (m_2 \sum_i \tilde{\phi}_{1i})/(\frac{m_1 m_2}{2})$ and also write $b_1 = (\sum_j \tilde{\phi}_{2j} y_{\cdot j})/(\frac{m_1 m_2}{2})$, $b_2 = (m_1 \sum_i \tilde{\phi}_{2j})/(\frac{m_1 m_2}{2})$. Thus, we know that a_1, a_2 are $O_p(\frac{1}{\sqrt{m_1}})$ and b_1, b_2 are $O_p(\frac{1}{\sqrt{m_2}})$.

Then we can compute

$$
\begin{aligned}
\theta_{11}^{(1)} - \theta_{21}^{(1)} &= \frac{\bar{y} + a_1 + b_1}{1 + a_2 + b_2} - \frac{\bar{y} - a_1 + b_1}{1 - a_2 + b_2} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \\
&= 2\frac{a_1(1 + b_2) - a_2(\bar{y} + b_1)}{(1 + b_2)^2 - a_2^2} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \\
&= 2\frac{a_1 - a_2\bar{y}}{(1 + b_2)^2 - a_2^2} + O_p(\frac{1}{\sqrt{m_1 m_2}}). \qquad (60)
\end{aligned}
$$

Similarly, we have

$$\begin{aligned}
\theta_{12}^{(1)} - \theta_{22}^{(1)} &= \frac{\bar{y} + a_1 - b_1}{1 + a_2 - b_2} - \frac{\bar{y} - a_1 - b_1}{1 - a_2 - b_2} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \\
&= 2\frac{a_1(1-b_2) - a_2(\bar{y} - b_1)}{(1-b_2)^2 - a_2^2} + O_p(\frac{1}{\sqrt{m_1 m_2}}) \\
&= 2\frac{a_1 - a_2\bar{y}}{(1-b_2)^2 - a_2^2} + O_p(\frac{1}{\sqrt{m_1 m_2}}). \qquad (61)
\end{aligned}$$

Thus, $\theta_{11}^{(1)} - \theta_{21}^{(1)} = \theta_{12}^{(1)} - \theta_{22}^{(1)} + O_p(\frac{1}{\sqrt{m_1 m_2}})$.

By computation, we see that

$$\begin{aligned}
&\log(\phi_{1i}^{(1)}[1]/\phi_{1i}^{(1)}[2]) \\
=\ & \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{1l}^{(1)}}(y_{ij}) + \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{1l}^{(1)}}(y_{ij}) \\
& -\{ \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{2l}^{(1)}}(y_{ij}) + \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{2l}^{(1)}}(y_{ij}) \\
=\ & \{ \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{1l}^{(1)}}(y_{ij}) - \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{2l}^{(1)}}(y_{ij}) \} \\
& +\{ \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{1l}^{(1)}}(y_{ij}) - \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \log f_{\theta_{2l}^{(1)}}(y_{ij}) \} \\
=\ & \{ \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \big(\log f_{\theta_{1l}^{(1)}}(y_{ij}) - \log f_{\theta_{2l}^{(1)}}(y_{ij}) \big) \} \\
& +\{ \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \big(\log f_{\theta_{1l}^{(1)}}(y_{ij}) - \log f_{\theta_{2l}^{(1)}}(y_{ij}) \big) \}. \qquad (62)
\end{aligned}$$

For any fixed set of parameter $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, we have that

$$\begin{aligned}
& \sum_{j:z_{2j}^*=1} \sum_l \phi_{2j}^{(0)}[l] \big(\log f_{\theta_{1l}}(y_{ij}) - \log f_{\theta_{2l}}(y_{ij}) \big) \\
=\ & g_{z_{1i}^* 1} \delta_{1,2} (m_2 + O_p(\sqrt{m_2})), \qquad (63)
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{j:z_{2j}^*=2} \sum_l \phi_{2j}^{(0)}[l] \big(\log f_{\theta_{1l}}(y_{ij}) - \log f_{\theta_{2l}}(y_{ij}) \big) \\
=\ & g_{z_{1i}^* 2} \delta_{1,2} (m_2 + O_p(\sqrt{m_2})), \qquad (64)
\end{aligned}$$

where $g_{kl} := \nabla \mathbb{E}_{y \sim f_{\theta_{kl}^*}} \log f_{\bar{\theta}}(y)$. Additionally, $\delta_{1,2} := \theta_{11}^{(1)} - \theta_{21}^{(1)}$. Furthermore, noticing that $\nabla \mathbb{E}_{y \sim f_\theta} \log f_{\bar{\theta}}(y)$ is increasing function of θ and $\theta_{11}^* > \theta_{12}^*, \theta_{21}^* > \theta_{22}^*, \theta_{11}^* > \theta_{21}^*, \theta_{12}^* > \theta_{22}^*$, we then know $g_{11} > g_{21}$ and $g_{12} > g_{22}$.

Let $\phi_{1i}^{(1)} = (\frac{1}{2} + \delta_i, \frac{1}{2} - \delta_i)$. We then know that $\delta_i = \exp\{c\delta_{1,2}^{(1)}\}/(1 + \exp\{c\delta_{1,2}^{(1)}\}) - 1/2$ for some positive constant c. (Here $\delta_{1,2}^{(1)} = \theta_{11}^{(1)} - \theta_{21}^{(1)}$.) Then minimum of δ_i for i with $z_{1i}^* = 1$ is larger than maximum of δ_i for i with $z_{1i}^* = 2$. Similarly, we let $\phi_{2j}^{(1)} = (\frac{1}{2} + \delta_j, \frac{1}{2} - \delta_j)$. Then $\delta_j = \exp\{c'\delta_{2,1}^{(1)}\}/(1 + \exp\{c'\delta_{2,1}^{(1)}\}) - 1/2$ where $\delta_{2,1}^{(1)} := \theta_{11}^{(1)} - \theta_{12}^{(1)}$. In addition, the minimum of δ_j for j with $z_{2j}^* = 1$ is larger than the maximum of δ_j for j with $z_{2j}^* = 2$. In other words, for sufficiently large m_1 and m_2, we know $\min_{i \in Z_{11}} \delta_i - \max_{i \in Z_{12}} \delta_i = \min_{j \in Z_{21}} \delta_j - \max_{j \in Z_{22}} \delta_j = \Omega(\max\{\theta_{11}^{(1)} - \theta_{12}^{(1)}, \theta_{11}^{(1)} - \theta_{21}^{(1)}\})$, where $Z_{1l} = \{i : z_{1i}^* = l\}$, $Z_{2l} = \{j : z_{2j}^* = l\}$ $(l = 1, 2)$.

Define sets $A_{11} = \{(i,j) : z_{1i}^* = 1, z_{2j}^* = 1\}$, $A_{12} = \{(i,j) : z_{1i}^* = 1, z_{2j}^* = 2\}$, $A_{21} = \{(i,j) : z_{1i}^* = 2, z_{2j}^* = 1\}$, $A_{22} = \{(i,j) : z_{1i}^* = 2, z_{2j}^* = 2\}$. Then for the second iteration, we can update the parameters as follows.

$$\theta_{11}^{(2)} = \frac{\sum_{(i,j)\in A_{11}} (\frac{1}{2} + \delta_i)(\frac{1}{2} + \delta_j) Y_{ij} + \sum_{(i,j)\in A_{12}} (\frac{1}{2} + \delta_i)(\frac{1}{2} - \delta_j) Y_{ij} + \sum_{(i,j)\in A_{21}} (\frac{1}{2} - \delta_i)(\frac{1}{2} + \delta_j) Y_{ij} + \sum_{(i,j)\in A_{22}} (\frac{1}{2} - \delta_i)(\frac{1}{2} - \delta_j) Y_{ij}}{\sum_{(i,j)\in A_{11}} (\frac{1}{2} + \delta_i)(\frac{1}{2} + \delta_j) + \sum_{(i,j)\in A_{12}} (\frac{1}{2} + \delta_i)(\frac{1}{2} - \delta_j) + \sum_{(i,j)\in A_{21}} (\frac{1}{2} - \delta_i)(\frac{1}{2} + \delta_j) + \sum_{(i,j)\in A_{22}} (\frac{1}{2} - \delta_i)(\frac{1}{2} - \delta_j)};$$

$$\theta_{21}^{(2)} = \frac{\sum_{(i,j)\in A_{11}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}+\sum_{(i,j)\in A_{12}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}+\sum_{(i,j)\in A_{21}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}+\sum_{(i,j)\in A_{22}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}}{\sum_{(i,j)\in A_{11}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{12}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{21}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{22}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)};$$

$$\theta_{12}^{(2)} = \frac{\sum_{(i,j)\in A_{11}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}+\sum_{(i,j)\in A_{12}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}+\sum_{(i,j)\in A_{21}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}+\sum_{(i,j)\in A_{22}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\sum_{(i,j)\in A_{11}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{12}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{21}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{22}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)};$$

$$\theta_{22}^{(2)} = \frac{\sum_{(i,j)\in A_{11}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}+\sum_{(i,j)\in A_{12}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}+\sum_{(i,j)\in A_{21}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)Y_{ij}+\sum_{(i,j)\in A_{22}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\sum_{(i,j)\in A_{11}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{12}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{21}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{22}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)}.$$

Compute

$$\begin{aligned}
& \frac{(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\sum_{(i,j)\in A_{11}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{12}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{21}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{22}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)} \\
- & \frac{(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\sum_{(i,j)\in A_{11}}(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{12}}(\frac{1}{2}-\delta_i)(\frac{1}{2}-\delta_j)+\sum_{(i,j)\in A_{21}}(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)+\sum_{(i,j)\in A_{22}}(\frac{1}{2}+\delta_i)(\frac{1}{2}-\delta_j)} \\
= & \frac{(\frac{1}{2}+\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\frac{|A|}{4}+\sum_{A_{11}}\frac{1}{2}\delta_i+\sum_{A_{12}}\frac{\delta_i}{2}-\sum_{A_{21}}\frac{\delta_i}{2}-\sum_{A_{22}}\frac{\delta_i}{2}+\sum_{A_{11}}\frac{\delta_j}{2}-\sum_{A_{12}}\frac{\delta_j}{2}+\sum_{A_{21}}\frac{\delta_j}{2}-\sum_{A_{22}}\frac{\delta_j}{2}+\sum_{A_{11}}\delta_i\delta_j-\sum_{A_{12}}\delta_i\delta_j-\sum_{A_{21}}\delta_i\delta_j+\sum_{A_{22}}\delta_i\delta_j} \\
- & \frac{(\frac{1}{2}-\delta_i)(\frac{1}{2}+\delta_j)Y_{ij}}{\frac{|A|}{4}-\sum_{A_{11}}\frac{1}{2}\delta_i-\sum_{A_{12}}\frac{\delta_i}{2}+\sum_{A_{21}}\frac{\delta_i}{2}+\sum_{A_{22}}\frac{\delta_i}{2}+\sum_{A_{11}}\frac{\delta_j}{2}-\sum_{A_{12}}\frac{\delta_j}{2}+\sum_{A_{21}}\frac{\delta_j}{2}-\sum_{A_{22}}\frac{\delta_j}{2}-\sum_{A_{11}}\delta_i\delta_j+\sum_{A_{12}}\delta_i\delta_j+\sum_{A_{21}}\delta_i\delta_j-\sum_{A_{22}}\delta_i\delta_j}.
\end{aligned} \tag{65}$$

By taking

$$x = \sum_{A_{11}}\frac{1}{2}\delta_i+\sum_{A_{12}}\frac{1}{2}\delta_i-\sum_{A_{21}}\frac{1}{2}\delta_i-\sum_{A_{22}}\frac{1}{2}\delta_i$$

and

$$y = \sum_{A_{11}}\frac{1}{2}\delta_j-\sum_{A_{12}}\frac{1}{2}\delta_j+\sum_{A_{21}}\frac{1}{2}\delta_j-\sum_{A_{22}}\frac{1}{2}\delta_j+\sum_{A_{11}}\delta_i\delta_j-\sum_{A_{12}}\delta_i\delta_j-\sum_{A_{21}}\delta_i\delta_j+\sum_{A_{22}}\delta_i\delta_j,$$

then (65) becomes

$$\begin{aligned}
& \frac{(\frac{1}{2}\delta_i+\delta_i\delta_j)(\frac{|A|}{4}+y)-(\frac{1}{4}+\frac{1}{2}\delta_j)x}{(\frac{|A|}{4}+y)^2-x^2} \\
= & \frac{1}{8}\frac{(n_1+n_2)(n_3+n_4)\delta_i-\sum_{i'\in Z_{11}}(n_3+n_4)\delta_{i'}+\sum_{i'\in Z_{12}}(n_3+n_4)\delta_{i'}+\text{ higher order}}{(\frac{|A|}{4}+y)^2-x^2}, \\
:= & w_{ij}^{1,1}
\end{aligned}$$

where $n_1=|A_{11}|$, $n_2=|A_{12}|$, $n_3=|A_{21}|$, $n_4=|A_{22}|$, higher order terms incorporate all $\delta_i\delta_j, \delta_i\delta_{j_1}\delta_{j_2}$ terms. Similarly, we can define $w_{ij}^{1,2}$, $w_{ij}^{2,1}$ and $w_{ij}^{2,2}$. Thus,

$$\begin{aligned}
& \theta_{11}^{(2)}-\theta_{21}^{(2)} \\
= & \sum_{i,j\in A_{11}}w_{ij}^{1,1}Y_{ij}+\sum_{i,j\in A_{12}}w_{ij}^{1,2}Y_{ij}+\sum_{i,j\in A_{21}}w_{ij}^{2,1}Y_{ij}+\sum_{i,j\in A_{22}}w_{ij}^{2,2}Y_{ij} \\
= & \frac{1}{8((\frac{|A|}{4}+x)^2-y^2)}((\sum_{i\in Z_{11}}n_2(n_3+n_4)^2\delta_i+\sum_{i\in Z_{12}}n_1(n_3+n_4)^2\delta_i)(\theta_{11}^*+\theta_{12}^*-\theta_{21}^*-\theta_{22}^*)+O_p(\frac{1}{\sqrt{m_2}}+\delta_i\delta_j)) \\
\geq & c\pi_1\pi_2\delta(\theta_{11}^*+\theta_{12}^*-\theta_{21}^*-\theta_{22}^*),
\end{aligned}$$

where $\delta > 0$ is $\min_{i \in Z_{11}} \delta_i - \max_{i \in Z_{12}} \delta_i$ and c is a universal constant. Similarly, we can also compute the difference

$$\theta_{11}^{(2)} - \theta_{12}^{(2)} \geq 2\pi_1\pi_2\delta(\theta_{11}^* + \theta_{21}^* - \theta_{12}^* - \theta_{22}^*) > \theta_{11}^{(1)} - \theta_{12}^{(1)}.$$

Therefore, before $\max\{|\theta_{11}^{(t)} - \theta_{21}^{(t)}|, |\theta_{11}^{(t)} - \theta_{12}^{(t)}|\} = \Omega(1)$, both gaps $\theta_{11}^{(t)} - \theta_{21}^{(t)}$ and $\theta_{11}^{(t)} - \theta_{12}^{(t)}$ increase as iterate t goes on. It eventually gives $q_{1i}^{(t)} \to \delta_{z_{1i}^*}$ and $q_{2j}^{(t)} \to \delta_{z_{2j}^*}$ for all $i \in [m_1]$, $j \in [m_2]$. We then obtain the global convergence.

References

Adityanand Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.