The proof of Theorem 1

Yuchen Zeng

1 Assumptions

1. $\frac{d_k}{d} \to \gamma_k$ where d and γ_k are real constant, $\forall k \in [K]$;
2. $\mathcal{C} \to \mathcal{C}_0$ as $d \to \infty$.

2 Notations

$\boldsymbol{c^{(k)}} \in \mathbb{R}^{d_k}$: unknown mode-k cluster membership vector with element $c_{i_k}^{(k)}$ refers to the true label of i_kth fiber in mode k, $\forall k \in [K],\ i_k \in [d_k]$;

$\boldsymbol{g^{(k)}} \in \mathbb{R}^{d_k}$: mode-k cluster assignment vector with element $g_{i_k}^{(k)}$ refers to the assigned label of i_kth fiber in mode k, $\forall k \in [K],\ i_k \in [d_k]$;

$\boldsymbol{p^{(k)}} \in \mathbb{R}^{R_k}$: mode-k cluster proportion vector with element $p_{r_k}^{(k)} = \frac{\sum_{i_k=1}^{d_k} \mathbb{I}\{c_{i_k}^{(k)}=r_k\}}{d_k}$, $\forall k \in [K],\ r_k \in [R_k]$;

$\boldsymbol{q^{(k)}} \in \mathbb{R}^{R_k}$: mode-k label proportion vector with element $q_{r_k}^{(k)} = \frac{\sum_{i_k=1}^{d_k} \mathbb{I}\{g_{i_k}^{(k)}=r_k\}}{d_k}$, can be seen as a function of $\boldsymbol{g^{(k)}}$, $\forall k \in [K],\ r_k \in [R_k]$;

$\boldsymbol{D^{(k)}} \in \mathbb{R}^{R_k \times R_k}$: mode-k confusion matrix with element $D_{r_k,r_k'}^{(k)} = \frac{1}{d_k}\sum_{i_k=1}^{d_k} \mathbb{I}\{c_{i_k}^{(k)} = r_k, g_{i_k}^{(k)} = r_k'\}$, can be seen as a function of $(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})$, $\forall k \in [K],\ r_k \in [R_k]$;

$\mathcal{J}_\tau = \{(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) : q_{r_1}^{(1)}(\boldsymbol{g^{(1)}}) > \tau, ..., q_{r_K}^{(K)}(\boldsymbol{g^{(K)}}) > \tau, r_k \in [R_k], k \in [K]\}$;

$\mathcal{I}_d \subset 2^{[d_1]} \times \cdots \times 2^{[d_K]}$: is the set of all the blocks that satisfy that $p_{i_k}^{(k)} > \tau,\ \forall i_k \in [d_k],\ \forall k \in [K]$;

$L_d = \inf\{|I| : I \in \mathcal{I}_d\}$;

$\mathcal{D}_{\boldsymbol{t}}^N = \{\mathbf{W} \in \mathbb{R}_+^{N\times N} : \boldsymbol{W}\boldsymbol{1} = \boldsymbol{t}\}$;

$\mathcal{P}_\varepsilon^{(k)} = \{D^{(k)} \in \mathcal{D}_{p^{(k)}}^{R_k} : MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) < \varepsilon\}$, $\forall k \in [K]$;

$\mathcal{M}_0$: the convex hull of the entries of $\mathcal{C}_0$;

$\mathcal{M}$: the neighborhood of $\mathcal{M}_0$;

$||\boldsymbol{A}||_\infty = \max_{r_1,...,r_K} |\boldsymbol{A}_{r_1,...,r_K}|$ for any tensor $\boldsymbol{A} \in \mathbb{R}^{R_1\times\cdots\times R_K}$.

Remark. 1. $\boldsymbol{D^{(k)}}\mathbf{1} = \boldsymbol{p^{(k)}}$, $\boldsymbol{D^{(k)}}^{\boldsymbol{T}}\mathbf{1} = \boldsymbol{q^{(k)}}$. If $\boldsymbol{D^{(k)}}$ is diagonal, then the assigned labels match the true cluster in mode k, $\forall k \in [K]$.
2. For $\boldsymbol{D^{(k)}} \notin \mathcal{P}_{\varepsilon}^{(k)}$ where $k \in [K]$, then there exist $a_k \neq a'_k \in [R_k]$ and $r_k \in [R_k]$ such that $D^{(k)}_{a_k r_k}, D^{(k)}_{a'_k r_k} > 0$, so $D^{(k)}_{a_k r_k} D^{(k)}_{a'_k r_k} > \frac{\varepsilon}{R_k^2}$.

3 Introduction

Theorem 3.1. *Consider a sub-Gaussian tensor block model with variance parameter σ^2 and non-degenerate clusterings, $\delta_{min} = \min\{\min_{r_1 \neq r'_1} \max_{r_2,...,r_K} (c_{r_1,...,r_K} - c_{r'_1,...,r_K})^2, ..., \min_{r_K \neq r'_K} \max_{r_1,...,r_{K-1}} (c_{r_1,...,r_K} - c_{r_1,...,r'_K})^2\}$, for all $k \in [K]$,*

$$\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon) \leq 2^{1+\sum_{k=1}^{K} d_k} exp(-\frac{C_2 \delta_{min}^2 \varepsilon^2 \prod_{k=1}^{K} d_k}{\sigma^2 \prod_{k=1}^{K} R_k^4})$$

To prove the theorem, considering our least-square estimator

$$\begin{aligned}
\hat{\Theta} &= \underset{\Theta \in \mathcal{P}}{\text{argmin}}\{-2 < \mathcal{Y}, \Theta > + ||\Theta||_F^2\} \\
&= \underset{\Theta \in \mathcal{P}}{\text{argmax}}\{< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}\}
\end{aligned}$$

the $< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}$ is the log-likelihood of the data tensor when our model is a Gaussian tensor block model.

Then the profile log-likelihood $F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})$ satisfies

$$\begin{aligned}
F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) &= \sup_{\Theta \in \mathcal{P}}\{< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}\} \\
&= \sup_{\Theta \in \mathcal{P}}\{\sum_{i_1,...,i_K} y_{i_1,...,i_K} c_{r_1(i_1),...,r_K(i_K)} - \frac{1}{2}\sum_{i_1,...,i_K} c^2_{r_1(i_1),...,r_K(i_K)}\} \\
&= \sup_{\Theta \in \mathcal{P}}\{\sum_{i_1,...,i_K} \overline{y_{r_1(i_1),...,r_K(i_K)}} c_{r_1(i_1),...,r_K(i_K)} - \frac{1}{2}\sum_{i_1,...,i_K} c^2_{r_1(i_1),...,r_K(i_K)}\} \\
&= \sum_{i_1,...,i_K} \overline{y_{r_1(i_1),...,r_K(i_K)}}^2 - \frac{1}{2}\sum_{i_1,...,i_K} \overline{y_{r_1(i_1),...,r_K(i_K)}}^2 \\
&= \frac{1}{2}\sum_{i_1,...,i_K} \overline{y_{r_1(i_1),...,r_K(i_K)}}^2 \\
&= \frac{1}{2}\sum_{r_1,...,r_K} \prod_{k=1}^{K} \hat{p}^{(k)}_{r_k} \overline{y_{r_1(i_1),...,r_K(i_K)}}^2 \\
&= \sum_{r_1,...,r_K} \prod_{k=1}^{K} \hat{p}^{(k)}_{r_k} f(\overline{y_{r_1(i_1),...,r_K(i_K)}})
\end{aligned}$$

where $f(x) = \frac{x^2}{2}$. Thus our clustering estimator can be represented as

$$(\widehat{\boldsymbol{g^{(1)}}}, ..., \widehat{\boldsymbol{g^{(K)}}}) = \underset{(\boldsymbol{g^{(1)}},...,\boldsymbol{g^{(K)}}) \in \mathcal{J}_\tau}{\text{argmax}} F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) \quad (1)$$

The error $||\hat{\Theta} - \Theta||_F^2$ comes from two aspects: noise and clustering. To measure the error which is from noise, we define a new function $G(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) : \mathcal{D}^{R_1}_{\boldsymbol{p^{(1)}}} \times \cdots \times \mathcal{D}^{R_K}_{\boldsymbol{p^{(K)}}} \to \mathbb{R}$:

$$G(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) = \sum_{r_1,...,r_K} [\boldsymbol{D^{(1)}}^{\boldsymbol{T}}\mathbf{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}}\mathbf{1}]_{r_K} f(E_{r_1,...,r_K})$$

where $\boldsymbol{E(g^{(1)}, ..., g^{(K)})} \in \mathbb{R}^{R_1 \times R_2 \times \cdots R_K}$,

$$E(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K} = \frac{\sum_{i_1,...,i_K} \sum_{j_1,...,j_K} c_{j_1,...,j_K} \mathbb{I}\{c^{(1)}_{i_1} = j_1, g^{(1)}_{i_1} = r_1\} \cdots \mathbb{I}\{c^{(K)}_{i_K} = j_K, g^{(K)}_{i_K} = r_K\}}{\sum_{i_1,...,i_K} \mathbb{I}\{g^{(1)}_{i_1} = r_1, ..., g^{(K)}_{i_K} = r_K\}}$$

is the average value of $Ey_{i_1,...,i_K}$ over the block defined by labels $r_1, ..., r_K$. Additionally, we define normalized residual matrix $\boldsymbol{R}(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) \in \mathbb{R}^{R_1 \times \cdots \times R_K}$:

$$R(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K} = \overline{Y_{r_1,...,r_K}} - E(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K}$$

4 Proof

Since the event $MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon \ for \ all \ k \in [K]$ is contained in the event $(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}})) \notin \mathcal{P}^{(1)}_\varepsilon \times \cdots \mathcal{P}^{(K)}_\varepsilon$, we convert our goal into obtaining the upper bound of $\mathbb{P}((\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}})) \notin \mathcal{P}^{(1)}_\varepsilon \times \cdots \mathcal{P}^{(K)}_\varepsilon)$. Under the condition of $(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}})) \notin \mathcal{P}^{(1)}_\varepsilon \times \cdots \mathcal{P}^{(K)}_\varepsilon$, the most of the error comes from the noise but not clustering. Because the ε is arbitrary, when ε is very small, we can convert our goal into finding the upper bound of $\mathbb{P}(G(\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}})) - \sum_{r_1,...,r_K} p^{(1)}_{r_1} \cdots p^{(K)}_{r_K} f(c_{r_1,...,r_K}) \leq h(\varepsilon)$ where $h(\varepsilon)$ is a function of ε. Here is rigorous proof:

Lemma 4.1. *For all $\tau > 0$, for $(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) \in \mathcal{J}_\tau$ and $(\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}}) \notin \mathcal{P}^{(1)}_\varepsilon \times \cdots \mathcal{P}^{(K)}_\varepsilon$,*

$$G(\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}}) - \sum_{r_1,...,r_K} p^{(1)}_{r_1} \cdots p^{(K)}_{r_K} f(c_{r_1,...,r_K}) \leq -\frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^K R_k^2}$$

Proof. If $\boldsymbol{D^{(1)}} \notin \mathcal{P}^{(1)}_\varepsilon$, then for some r_1 and some $a_1 \neq a_1'$, $D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} \geq \frac{\varepsilon}{R_1^2}$. Since the core tensor is irreducible according to our basic assumption in paper, there exist $a_2, ..., a_K$ such that $c_{a_1,...,a_K} \neq c_{a_1',...,a_K}$. Select the $a_2, ..., a_K$ such that $(c_{a_1,...,a_K} - c_{a_1',...,a_K})^2 = \min_{a_1 \neq a_1'} \max_{a_2,...,a_K} (c_{a_1,...,a_K} - c_{a_1',...,a_K})^2)$. Let r_j be the index of the largest element in row a_j of matrix $\boldsymbol{D^{(j)}}$, $j = 2, ..., K$. Those elements must be at least as large as the mean:

$$D^{(j)}_{a_j r_j} \geq \frac{[\boldsymbol{D^{(j)}}\mathbf{1}]_{a_j}}{d_j} \geq \frac{\tau}{d_j} \ j = 2, ..., K$$

Let $W = [\boldsymbol{D^{(1)}}^{\boldsymbol{T}}\mathbf{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}}\mathbf{1}]_{r_K}$, this is nonzero according to the selection of $r_1, ..., r_K$. Now, there exists $c_* \in \mathcal{M}$ such that

$$\begin{aligned}[C \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,...,r_K} &= D^{(1)}_{a_1 r_1} \cdots D^{(K)}_{a_K r_K} c_{a_1,....,a_K} + D^{(1)}_{a_1' r_1} \cdots D^{(K)}_{a_K r_K} c_{a_1',...,a_K} \\ &\quad + (W - D^{(1)}_{a_1 r_1} \cdots D^{(K)}_{a_K r_K} - D^{(1)}_{a_1' r_1} \cdots D^{(K)}_{a_K r_K}) c_*\end{aligned}$$

Let $z = \frac{[C_0 \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K}}{W}$ and define $N = [\nu_{a_1,\ldots,a_K}] \in \mathbb{R}^{R_1 \times \cdots \times R_K}$ with $\nu_{a_1,\ldots,a_K} = f(c_{a_1,\ldots,a_K})$, then,

$$
\begin{aligned}
& \frac{[N \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K}}{W} - f(z) \\
\geq \quad & \frac{1}{2}(c_{a_1,\ldots,a_K} - z)^2 \ (Taylor\ Series) \\
\geq \quad & \frac{1}{2}[\frac{D^{(1)}_{a_1 r_1} \cdots D^{(K)}_{a_K r_K}}{W}(c_{a_1,\ldots,a_K} - z)^2 + \frac{D^{(1)}_{a_1' r_1} \cdots D^{(K)}_{a_K r_K}}{W}(c_{a_1' \cdots a_K} - z)^2 \\
& + \frac{W - D^{(1)}_{a_1 r_1} \cdots D^{(K)}_{a_K r_K} - D^{(1)}_{a_1' r_1} \cdots D^{(K)}_{a_K r_K}}{W}(c_* - z)^2] \\
\geq \quad & \frac{1}{2}[\frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{W^2}(c_{a_1,\ldots,a_K} - z)^2 + \frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{W^2}(c_{a_1' \cdots a_K} - z)^2] \\
= \quad & \frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{W}[\frac{1}{2}(c_{a_1,\ldots,a_K} - z)^2 + \frac{1}{2}(z - c_{a_1',\ldots,a_K})^2] \\
\geq \quad & \frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{W}[\frac{1}{2}(c_{a_1,\ldots,a_K} - z) + \frac{1}{2}(z - c_{a_1',\ldots,a_K})]^2 \\
= \quad & \frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{4W^2}(c_{a_1,\ldots,a_K} - c_{a_1',\ldots,a_K})^2
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& [\boldsymbol{D^{(1)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_K} f(\frac{[\mathcal{C}_0 \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K}}{[\boldsymbol{D^{(1)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_K}}) \\
& -[N \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K} \\
= \quad & W f(z) - [N \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K} \\
\leq \quad & -\frac{D^{(1)}_{a_1 r_1} D^{(1)}_{a_1' r_1} {D^{(2)}_{a_2 r_2}}^2 \cdots {D^{(K)}_{a_K r_K}}^2}{4W}(c_{a_1,\ldots,a_K} - c_{a_1',\ldots,a_K})^2 \\
\leq \quad & -\frac{\varepsilon \tau^{2(K-1)}}{4R_1^2 \cdots R_K^2}(c_{a_1,\ldots,a_K} - c_{a_1',\ldots,a_K})^2
\end{aligned}
$$

This inequality only holds for one particular choice for $r_1, \ldots, r_K$. For other choices, the left hand side is non-positive by Jensen's inequality:

$$
f(\frac{[C_0 \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K}}{[\boldsymbol{D^{(1)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_K}}) - \frac{[N \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,\ldots,r_K}}{[\boldsymbol{D^{(1)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}} \boldsymbol{1}]_{r_K}} \geq 0
$$

It follows that,

$$\begin{aligned}
& \sum_{r_1,...,r_K} [\boldsymbol{D^{(1)}}^{\boldsymbol{T}}\mathbf{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}}\mathbf{1}]_{r_K} f(\frac{[\mathcal{C}_0 \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,...,r_K}}{[\boldsymbol{D^{(1)}}^{\boldsymbol{T}}\mathbf{1}]_{r_1} \cdots [\boldsymbol{D^{(K)}}^{\boldsymbol{T}}\mathbf{1}]_{r_K}}) \\
& - \sum_{r_1,...,r_K} [N \times_1 \boldsymbol{D^{(1)}}^{\boldsymbol{T}} \times_2 \cdots \times_K \boldsymbol{D^{(K)}}^{\boldsymbol{T}}]_{r_1,...,r_K} \\
= \;& G(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}})) - \sum_{r_1,...,r_K} \prod_{k=1}^{K} p_{r_k}^{(k)} f(c_{r_1,...,r_K}) \\
\leq \;& -\frac{\varepsilon\tau^{2(K-1)}}{4R_1^2 \cdots R_K^2}(c_{a_1,...,a_K} - c_{a_1',...,a_K})^2
\end{aligned}$$

Similarly, if $\boldsymbol{D^{(k)}} \notin \mathcal{P}_\varepsilon^{(k)}$, then the left hand side would be bounded by $-\frac{\varepsilon\tau^{2(K-1)}}{4\prod_{k=2}^{K} R_k^2}(c_{a_1,...,a_K} - c_{a_1',...,a_K})^2$. Thus,

$$G(\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}}) - \sum_{r_1,...,r_K} p_{r_1}^{(1)} \cdots p_{r_K}^{(K)} f(c_{r_1,...,r_K}) \leq -\frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^{K} R_k^2}$$

□

By lemma 4.1, we obtained

$$\begin{aligned}
& \mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon) \\
\leq \;& \mathbb{P}((\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}}) \notin \mathcal{P}_\varepsilon^{(1)} \times \cdots \times \mathcal{P}_\varepsilon^{(K)}) \\
\leq \;& \mathbb{P}(G(\boldsymbol{D^{(1)}}, ..., \boldsymbol{D^{(K)}}) - \sum_{r_1,...,r_K} p_{r_1}^{(1)} \cdots p_{r_K}^{(K)} f(c_{r_1,...,r_K}) \leq -\frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^{K} R_k^2}) \\
= \;& \mathbb{P}(G(\boldsymbol{D^{(1)}}(\widehat{\boldsymbol{g^{(1)}}}), ..., \boldsymbol{D^{(K)}}(\widehat{\boldsymbol{g^{(K)}}})) - F(\boldsymbol{c^{(1)}}, ..., \boldsymbol{c^{(K)}}) \leq -\frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^{K} R_k^2})
\end{aligned}$$

Additionally, we let $r_d = \sup_{\mathcal{J}_\tau} |F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) - G(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}}))|$,

$$\begin{aligned}
& F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) - F(\boldsymbol{c^{(1)}}, ..., \boldsymbol{c^{(K)}}) \\
\leq \;& |F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) - G(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}}))| \\
& + |F(\boldsymbol{c^{(1)}}, ..., \boldsymbol{c^{(K)}}) - G(\boldsymbol{D^{(1)}}(\boldsymbol{c^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{c^{(K)}}))| \\
& + [G(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}})) - G(\boldsymbol{D^{(1)}}(\boldsymbol{c^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{c^{(K)}}))] \\
\leq \;& 2r_d - \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^{K} R_k^2}
\end{aligned}$$

Thus,

$$\begin{aligned}
& \mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon) \\
\leq \;& \mathbb{P}(F(\widehat{\boldsymbol{g^{(1)}}}, ..., \widehat{\boldsymbol{g^{(K)}}}) - F(\boldsymbol{c^{(1)}}, ..., \boldsymbol{c^{(K)}}) \leq 2r_d - \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{4\prod_{k=1}^{K} R_k^2}) \\
\leq \;& \mathbb{P}(r_d \geq \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{8\prod_{k=1}^{K} R_k^2})
\end{aligned}$$

Now we convert our problem into find the upper bound of $\mathbb{P}(r_d \geq \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{8\prod_{k=1}^{K} R_k^2})$. Consider $\mathbb{P}(r_d \leq t)$, because f is locally lipschitz continuous with lipschitz constant $c = \sup|f'(\mu)|$ for μ in a neighborhood of $\mathcal{M}$

$$\begin{aligned}
& |F(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}}) - G(\boldsymbol{D^{(1)}}(\boldsymbol{g^{(1)}}), ..., \boldsymbol{D^{(K)}}(\boldsymbol{g^{(K)}}))| \\
= & |\sum_{r_1,...,r_K} \hat{p}_{r_1}^{(1)}\hat{p}_{r_2}^{(2)}\cdots\hat{p}_{r_K}^{(K)}[f(\overline{Y_{r_1,...,r_K}}) - f(E_{r_1,...,r_K})]| \\
\leq & \sum_{r_1,...,r_K} \hat{p}_{r_1}^{(1)}\hat{p}_{r_2}^{(2)}\cdots\hat{p}_{r_K}^{(K)}|f(\overline{Y_{r_1,...,r_K}}) - f(E_{r_1,...,r_K})| \\
\leq & c\sum_{r_1,...,r_K} \hat{p}_{r_1}^{(1)}\hat{p}_{r_2}^{(2)}\cdots\hat{p}_{r_K}^{(K)}|R(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K}| \\
\leq & c||\boldsymbol{R}(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})||_\infty
\end{aligned}$$

Therefore,

$$\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon) \leq \mathbb{P}(||\boldsymbol{R}(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})||_\infty \geq \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{8c\prod_{k=1}^{K} R_k^2}) \tag{2}$$

According to Hoeffding's inequality,

$$\begin{aligned}
& \mathbb{P}(|R(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K}| \geq \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{8c\prod_{k=1}^{K} R_k^2}) \\
= & \mathbb{P}(|\overline{Y_{r_1,...,r_K}} - E(\boldsymbol{g^{(1)}}, ..., \boldsymbol{g^{(K)}})_{r_1,...,r_K}| \geq \frac{\delta_{min}\tau^{2(K-1)}\varepsilon}{8c\prod_{k=1}^{K} R_k^2}) \\
\leq & 2exp(-\frac{\delta_{min}^2\tau^{4(K-1)}\varepsilon^2\mathcal{L}_d}{128c^2\sigma^2\prod_{k=1}^{K} R_k^4})
\end{aligned}$$

Combine the result with (3) and $\mathcal{L}_d \geq \tau^K \prod_{k=1}^{K} d_k$

$$\begin{aligned}
\mathbb{P}(MCR(\hat{\mathbf{M}}_k, \mathbf{P}_k\mathbf{M}_{k,true}) \geq \varepsilon) & \leq 2^{1+\sum_{k=1}^{K} d_k} exp(-\frac{\delta_{min}^2\tau^{4(K-1)}\varepsilon^2\mathcal{L}_d}{128c^2\sigma^2\prod_{k=1}^{K} R_k^4}) \\
& \leq 2^{1+\sum_{k=1}^{K} d_k} exp(-\frac{\delta_{min}^2\tau^{5K-4}\varepsilon^2\prod_{k=1}^{K} d_k}{128c^2\sigma^2\prod_{k=1}^{K} R_k^4})
\end{aligned}$$

Letting $C_2 = \frac{\tau^{5K-4}}{128c^2}$ yields the result.