

# Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu and Miaoyan Wang\*

Department of Statistics, University of Wisconsin - Madison

## Abstract

We consider the problem of multiway clustering in the presence of unknown degree heterogeneity. Such data problems arise commonly in applications such as recommendation system, neuroimaging, community detection, and hypergraph partitions in social networks. The allowance of degree heterogeneity provides great flexibility in clustering models, but the extra complexity poses significant challenges in both statistics and computation. Here, we develop a degree-corrected tensor block model with estimation accuracy guarantees. We present the phase transition of clustering performance based on the notion of angle separability, and we characterize three signal-to-noise regimes corresponding to different statistical-computational behaviors. In particular, we demonstrate that an intrinsic statistical-to-computational gap emerges only for tensors of order three or greater. Further, we develop an efficient polynomial-time algorithm that provably achieves exact clustering under mild signal conditions. The efficacy of our procedure is demonstrated through two data applications, one on human brain connectome project, and another on Peru Legislation network dataset.

## Index Terms

tensor clustering, degree correction, statistical-computational efficiency, human brain connectome networks

## I. INTRODUCTION

MULTIWAY arrays have been widely collected in various fields including social networks (Anandkumar et al., 2014), neuroscience (Wang et al., 2017), and computer science (Koniusz and Cherian, 2016). Tensors effectively represent the multiway data and serve as the foundation in higher-order data analysis. One data example is from multi-tissue multi-individual gene expression study (Hore et al., 2016; Wang et al., 2019), where the data tensor consists of expression measurements indexed by (gene, individual, tissue) triplets. Another example is *hypergraph* network (Ahn et al., 2019; Ghoshdastidar and Dukkipati, 2017; Ghoshdastidar et al., 2017; Ke et al., 2019) in social science. A  $K$ -uniform hypergraph can be naturally represented as an order- $K$  tensor, where each entry indicates the presence of  $K$ -way hyperedge among nodes (a.k.a. entities). In both examples, identifying the similarity among tensor entities is important for scientific discovery.

We study the problem of multiway clustering based on a data tensor. The goal of multiway clustering is to identify a checkerboard structure from a noisy data tensor. Figure 1 illustrates the noisy tensor and the underlying checkerboard structures discovered by multiway clustering methods. In the hypergraph example, the multiway clustering aims to identify the underlying block partition of nodes based on their higher-order connectivities; therefore, we also refer to the clustering as *higher-order clustering*. The most common model for higher-order clustering is called *tensor block model* (TBM) (Wang and Zeng, 2019), which extends the usual matrix stochastic block model (Abbe, 2017) to tensors. The matrix analysis tools, however, are sub-optimal for higher-order clustering. Developing tensor tools for solving block models has received increased interest recently (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019).

The classical tensor block model suffers from drawbacks to model real world data in spite of the popularity. The key underlying assumption of block model is that all nodes in the same community are exchangeable; i.e., the nodes have no individual-specific parameters apart from the community-specific parameters. However, the exchangeability assumption is often non-realistic. Each node may contribute to the data variation by its own multiplicative effect. We call the unequal node-specific effects the *degree heterogeneity*. Such degree heterogeneity appears commonly

\*Part of this paper was presented at 25th International Conference on Artificial Intelligence and Statistics (AISTATS).

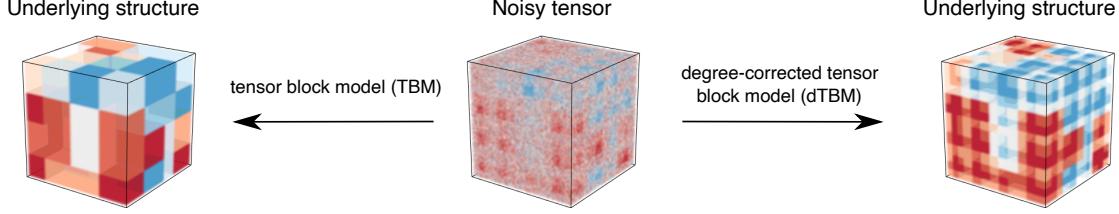


Fig. 1: Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

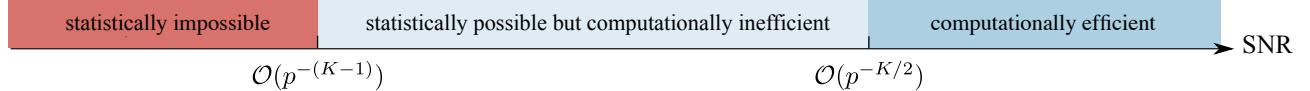


Fig. 2: SNR thresholds for statistical and computational limits in order- $K$  dTBM with dimension  $(p, \dots, p)$  and  $K \geq 2$ . The SNR gap between statistical possibility and computational efficiency exists only for tensors with  $K \geq 3$ .

in social networks. Ignoring the degree heterogeneity may seriously mislead the clustering results. For example, the regular block model fails to model the member affiliation in the Karate Club network (Bickel and Chen, 2009) without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity (Ke et al., 2019). The dTBM combines a higher-order checkerboard structure with degree parameter  $\theta = (\theta(1), \dots, \theta(p))^T$  to allow heterogeneity among  $p$  nodes. Figure 1 compares the underlying structures of TBM and dTBM with the same number of communities. The dTBM allows varying values within the same community, thereby allowing a richer structure. To solve dTBM, we project clustering objects to a unit sphere and perform iterative clustering based on angle similarity. We refer to the algorithm as the *spherical* clustering; detailed procedures are in Section IV. The spherical clustering avoids the estimation of nuisance degree heterogeneity. The usage of angle similarity brings new challenges to the theoretical results, and we develop new polar-coordinate based techniques in the proofs.

**Our contributions.** The primary goal of this paper is to provide both statistical and computational guarantees for dTBM. Our main contributions are summarized below.

- We develop a general dTBM and establish the identifiability for the uniqueness of clustering using the notion of angle separability.
- We present the phase transition of clustering performance with respect to three different statistical and computational behaviors. We characterize, for the first time, the critical signal-to-noise (SNR) thresholds in dTBMs, revealing the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering. Specific SNR thresholds and algorithm behaviors are depicted in Figure 2.
- We provide an angle-based algorithm that achieves exact clustering *in polynomial time* under mild conditions. Simulation and data studies demonstrate that our algorithm outperforms existing higher-order clustering algorithms.

The last two contributions, to our best knowledge, are new to the literature of dTBMs.

**Related work.** Our work is closely related to but also distinct from several lines of existing research. Table I summarizes the most relevant models.

- *Block model for clustering.* The block model such as stochastic block model (SBM) and degree-corrected SBM has been widely used for matrix clustering problems. The theoretical properties and algorithm performance for matrix block models have been well-studied (Gao et al., 2018); see the review paper (Abbe, 2017) and the references therein. However, The tensor counterparts are relatively less understood.

	Gao et al. (2018)	Han et al. (2020)	Ghoshdastidar et al. (2017)	Ke et al. (2019)	<b>Ours</b>
Allow tensors of arbitrary order	✗	✓	✓	✓	✓
Allow degree heterogeneity	✓	✗	✓	✓	✓
Singular-value gap-free clustering	✓	✓	✗	✗	✓
Misclustering rate (for order $K^*$ )	-	$\exp(-p^{K/2})$	$p^{-1}$	$p^{-2}$	$\exp(-p^{K/2})$

TABLE I: Comparison between previous methods with our method. \*We list the result for order-K tensors with  $K \geq 3$  and general number of communities  $r = \mathcal{O}(1)$ .

- *Tensor block model.* The (non-degree) tensor block model (TBM) is a higher-order extension of SBM, and its statistical-computational properties are investigated in recent literatures (Ghoshdastidar et al., 2017; Han et al., 2020; Wang and Zeng, 2019). Extending results from non-degree to degree-corrected model is highly challenging. Our dTBM parameter space is equipped with angle-based similarity and nuisance degree parameters. The extra complexity makes the Cartesian coordinates based analysis (Han et al., 2020) non-applicable to our setting. Towards this goal, we have developed a new polar coordinates based analysis to control the model complexity. We also develop a new angle-based iteration algorithm to achieve optimal clustering rates *without the need of estimating nuisance degree parameters*.

- *Degree-corrected block model.* The hypergraph degree-corrected block model (hDCBM) and its variant have been proposed in the literature (Ke et al., 2019; Yuan et al., 2018). For this popular model, however, the optimal statistical-computational rates remain an open problem. Our main contribution is to provide a sharp statistical and computational critical phase transition in dTBM literature. In addition, our algorithm results in a faster *exponential* error rate, in contrast to the *polynomial* rate in Ke et al. (2019). The original hDCBM (Ke et al., 2019) is designed for binary observations only, and we extend the model to both continuous and binary observations. We believe our results are novel and helpful to the community. See Fig 2 for overview of our results.

- *Global-to-local algorithm strategy.* Our methods generalize the recent global-to-local strategy for matrix learning (Chi et al., 2019; Gao et al., 2018; Yun and Proutiere, 2016) to tensors (Ahn et al., 2018; Han et al., 2020; Kim et al., 2018). Despite the conceptual similarity, we address several fundamental challenges associated with this non-convex, non-continuous problem. We show the insufficiency of the conventional tensor HOSVD (De Lathauwer et al., 2000), and we develop a weighted higher-order initialization that relaxes the singular-value gap separation condition. Furthermore, our local iteration leverages the angle-based clustering in order to avoid explicit estimation of degree heterogeneity. Our bounds reveal the interesting interplay between the computational and statistical errors. We show that our final estimate *provably* achieves the exact clustering within only polynomial-time complexity.

**Notation.** We use lower-case letters (e.g.,  $a, b$ ) for scalars, lower-case boldface letters (e.g.,  $\mathbf{a}, \boldsymbol{\theta}$ ) for vectors, upper-case boldface letters (e.g.,  $\mathbf{X}, \mathbf{Y}$ ) for matrices, and calligraphy letters (e.g.,  $\mathcal{X}, \mathcal{Y}$ ) for tensors of order three or greater. We use  $\mathbf{1}_p$  to denote a vector of length  $p$  with all entries to be 1. We use  $|\cdot|$  for the cardinality of a set and  $\mathbf{1}\{\cdot\}$  for the indicator function. For an integer  $p \in \mathbb{N}_+$ , we use the shorthand  $[p] = \{1, 2, \dots, p\}$ . For a length- $p$  vector  $\mathbf{a}$ , we use  $a(i) \in \mathbb{R}$  to denote the  $i$ -th entry of  $\mathbf{a}$ , and use  $\mathbf{a}_I$  to denote the sub-vector by restricting the indices in the set  $I \subset [p]$ . We use  $\|\mathbf{a}\| = \sqrt{\sum_i a^2(i)}$  to denote the  $\ell_2$ -norm,  $\|\mathbf{a}\|_1 = \sum_i |a_i|$  to denote the  $\ell_1$  norm of  $\mathbf{a}$ . For two vector  $\mathbf{a}, \mathbf{b}$  of the same dimension, we denote the angle between  $\mathbf{a}, \mathbf{b}$  by

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|},$$

where  $\langle \mathbf{a}, \mathbf{b} \rangle$  is the inner product of two vectors and  $\cos(\mathbf{a}, \mathbf{b}) \in [-1, 1]$ . We make the convention that  $\cos(\mathbf{a}, \mathbf{b}) = \cos(\mathbf{a}^T, \mathbf{b}^T)$ .

Let  $\mathcal{Y} \in \mathbb{R}^{p_1 \times \dots \times p_K}$  be an order- $K$  ( $p_1, \dots, p_K$ )-dimensional tensor. We use  $\mathcal{Y}(i_1, \dots, i_K)$  to denote the  $(i_1, \dots, i_K)$ -th entry of  $\mathcal{Y}$ . The multilinear multiplication of a tensor  $\mathcal{S} \in \mathbb{R}^{r_1 \times \dots \times r_K}$  by matrices  $\mathbf{M}_k \in \mathbb{R}^{p_k \times r_k}$  results in an order- $d$  ( $p_1, \dots, p_K$ )-dimensional tensor  $\mathcal{X}$ , denoted

$$\mathcal{X} = \mathcal{S} \times_1 \mathbf{M}_1 \times \dots \times_K \mathbf{M}_K,$$

where the entries of  $\mathcal{X}$  are defined by

$$\mathcal{X}(i_1, \dots, i_K) = \sum_{(j_1, \dots, j_K)} \mathcal{S}(j_1, \dots, j_K) \mathbf{M}_1(i_1, j_1) \cdots \mathbf{M}_K(i_K, j_K).$$

For a matrix  $\mathbf{Y}$ , we use  $\mathbf{Y}_{::i}$  (respectively,  $\mathbf{Y}_{:i}$ ) to denote the  $i$ -th row (respectively,  $i$ -th column) of the matrix. Similarly, for an order-3 tensor, we use  $\mathcal{Y}_{:::i}$  to denote the  $i$ -th matrix slide of the tensor. We use  $\text{Ave}(\cdot)$  to denote the operation of taking averages across elements and  $\text{Mat}_k(\cdot)$  to denote the unfolding operation that reshapes the tensor along mode  $k$  into a matrix. For a symmetric tensor  $\mathcal{X} \in \mathbb{R}^{p \times \cdots \times p}$ , we omit the subscript and use  $\text{Mat}(\mathcal{X}) \in \mathbb{R}^{p \times p^{K-1}}$  to denote the unfolding. For two sequences  $\{a_p\}, \{b_p\}$ , we denote  $a_p \lesssim b_p$  or  $a_p = \mathcal{O}(b_p)$  if  $\lim_{p \rightarrow \infty} a_p/b_p \leq c$  for some constant  $c \geq 0$ ,  $a_p = o(b_p)$  if  $\lim_{p \rightarrow \infty} a_p/b_p = 0$ , and  $a_p = \Omega(b_p)$  if both  $b_p \lesssim a_p$  and  $a_p \lesssim b_p$ . Throughout the paper, we use the terms ‘‘community’’ and ‘‘clusters’’ exchangeably.

**Organization.** The rest of this paper is organized as follows. Section II introduces the degree-corrected tensor block model (dTBM) with three motivating examples and presents the identifiability of dTBM under the angle gap condition. We show the phase transition and the existence of statistical-computational gaps for the higher-order dTBM in Section III. In Section IV, we provide a polynomial-time two-stage algorithm with misclustering rate guarantees. Extension to Bernoulli models is also presented. In Section V, we compare our work with non-degree tensor block models. Numerical studies including the simulation, comparison with other methods, and two real dataset analyses are in Sections VI-VII. The main technical ideas we develop for addressing main theorems are provided in Section VIII. Detailed proofs and extra theoretical results are provided in Appendix.

## II. MODEL FORMULATION AND MOTIVATIONS

### A. Degree-corrected tensor block model

Suppose we have an order- $K$  data tensor  $\mathcal{Y} \in \mathbb{R}^{p \times \cdots \times p}$ . Assume there exist  $r \geq 2$  disjoint communities among the  $p$  nodes. We represent the community assignment by a function  $z: [p] \mapsto [r]$ , where  $z(i) = a$  for  $i$ -th node that belongs to the  $a$ -th community. Then,  $z^{-1}(a) = \{i \in [p]: z(i) = a\}$  denotes the set of nodes that belong to the  $a$ -th community, and  $|z^{-1}(a)|$  denotes the number of nodes in the  $a$ -th community. Let  $\boldsymbol{\theta} = (\theta(1), \dots, \theta(p))^T$  denote the degree heterogeneity for  $p$  nodes. We consider the order- $K$  dTBM (Ghoshdastidar et al., 2017; Ke et al., 2019),

$$\mathcal{Y}(i_1, \dots, i_K) = \mathcal{S}(z(i_1), \dots, z(i_K)) \prod_{k=1}^K \theta_{i_k} + \mathcal{E}(i_1, \dots, i_K), \quad (1)$$

where  $\mathcal{S} \in \mathbb{R}^{r \times \cdots \times r}$  is an order- $K$  tensor collecting the block means among communities, and  $\mathcal{E} \in \mathbb{R}^{p \times \cdots \times p}$  is a noise tensor consisting of independent zero-mean sub-Gaussian entries with variance bounded by  $\sigma^2$ . The unknown parameters are  $z$ ,  $S$ , and  $\boldsymbol{\theta}$ . The dTBM can be equivalently written in a compact form of tensor-matrix product:

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \cdots \times_K \boldsymbol{\Theta} \mathbf{M}, \quad (2)$$

where  $\boldsymbol{\Theta} = \text{diag}(\theta(1), \dots, \theta(p)) \in \mathbb{R}^{p \times p}$  is a diagonal matrix,  $\mathbf{M} \in \{0, 1\}^{p \times r}$  is the membership matrix associated with community assignment  $z$  such that  $\mathbf{M}(i, j) = \mathbb{1}\{z(i) = j\}$ . By definition, each row of  $\mathbf{M}$  has one copy of 1’s and 0’s elsewhere. Note that the discrete nature of  $\mathbf{M}$  renders our model (2) more challenging than Tucker decomposition. We call a tensor  $\mathcal{Y}$  an  $r$ -block tensor with degree  $\boldsymbol{\theta}$  if  $\mathcal{Y}$  admits dTBM (2) and let  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  denote the mean tensor. The goal of clustering is to estimate  $z$  from a single noisy tensor  $\mathcal{Y}$ . We are particularly interested in the high-dimensional regime where  $p$  grows whereas  $r = \mathcal{O}(1)$ .

For ease of notation, we have focused on the case with symmetric mean tensor  $\mathbb{E}\mathcal{Y}$ . This assumption simplifies the notation because all modes have the same  $(\boldsymbol{\Theta}, \mathbf{M}, z)$ ; the noise tensor  $\mathcal{E}$  and the data tensor  $\mathcal{Y}$  are still possibly asymmetric. In general, we allow asymmetric mean tensors with  $\{(\boldsymbol{\Theta}_k, \mathbf{M}_k, z_k)\}_{k=1}^K$ , one for each mode. The extension can be found in Appendix II.

### B. Motivating examples

Here, we provide four applications to illustrate the practical necessity of dTBM.

a) *Tensor block model*: Consider the model (2). Let  $\theta(i) = 1$  for all  $i \in [p]$ . The model (2) reduces to the tensor block model, which is widely used in previous clustering algorithms (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019). The theoretical results in TBM serve as benchmarks for dTBM.

b) *Community detection in hypergraphs*: The hypergraph network is a powerful tool to represent the complex entity relations with higher-order interactions (Ke et al., 2019). A typical undirected hypergraph is denoted as  $H = (V, E)$ , where  $V = [p]$  is the set of nodes and  $E$  is the set of undirected hyperedges. Each hyperedge in  $E$  is a subset of  $V$ , and we call the hyperedge an order- $K$  edge if the corresponding subset involves  $K$  nodes. We call  $H$  a  $K$ -uniform hypergraph if  $E$  only contains order- $K$  edges.

It is natural to represent the  $K$ -uniform hypergraph using a binary order- $K$  adjacency tensor. Let  $\mathcal{Y} \in \{0, 1\}^{p \times \dots \times p}$  denote the adjacency tensor, where the entries encode the presence or absence of order- $K$  edges among  $p$  nodes. Specifically, for all  $(i_1, \dots, i_K) \in [p]^K$ , we have

$$\mathcal{Y}(i_1, \dots, i_K) = \begin{cases} 1 & \text{if } (i_1, \dots, i_K) \in E, \\ 0 & \text{if } (i_1, \dots, i_K) \notin E. \end{cases}$$

Assume there exist  $r$  disjoint communities among  $p$  nodes, and the connection probabilities depend on the community assignments and node-specific parameters. Then, the equation (2) models  $\mathbb{E}\mathcal{Y}$  with unknown degree heterogeneity  $\theta$  and sub-Gaussianity parameter  $\sigma^2 = 1/4$ .

c) *Multi-layer weighted network*: Multi-layer weighted network data consists of multiple networks over the same set of nodes. One representative example is the brain connectome data (Zhang et al., 2019). The multi-layer weighted network  $\mathcal{Y}$  has dimension of  $p \times p \times L$ , where  $p$  denotes the number of brain regions of interest, and  $L$  denotes the number of layers (networks). Each of the  $L$  networks describes one aspect of the brain connectivity, such as functional connectivity or structural connectivity. The resulting tensor  $\mathcal{Y}$  consists of a mixture of slices with various data types.

Assume there exist  $r$  disjoint communities among  $p$  nodes and  $r_l$  disjoint communities among the  $L$  layers. The multi-layer network community detection is modeled by the general asymmetric asymmetric dTBM model (2)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \Theta \mathbf{M} \times_2 \Theta \mathbf{M} \times_3 \Theta_l \mathbf{M}_l,$$

where  $(\Theta \in \mathbb{R}^p, \mathbf{M} \in \{0, 1\}^{p \times r})$  and  $(\Theta_l \in \mathbb{R}^L, \mathbf{M}_l \in \{0, 1\}^{L \times r_l})$  are the degree heterogeneity and membership matrices corresponding to the community structure for  $p$  nodes and  $L$  layers, respectively.

d) *Gaussian higher-order clustering*: Datasets in various fields such as medical image, genetics, and computer science are formulated as Gaussian tensors. One typical example is the multi-tissue gene expression dataset, which records the different gene expression in different individuals and different tissues. The dataset, denoted as  $\mathcal{Y} \in \mathbb{R}^{p \times n \times t}$ , consists of the expression data for  $p$  genes of  $n$  individuals in  $t$  tissues.

Assume there exist  $r_1, r_2, r_3$  disjoint clusters for  $p$  genes,  $n$  individuals, and  $t$  tissues, respectively. We apply the general asymmetric dTBM model (2)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \Theta_1 \mathbf{M}_1 \times_2 \Theta_2 \mathbf{M}_2 \times_3 \Theta_3 \mathbf{M}_3,$$

where  $\{(\Theta_k, \mathbf{M}_k)\}_{k=1}^3$  represents the heterogeneity and membership for genes, individuals, and tissues.

**Remark 1** (Comparison with non-degree models). Our dTBM uses fewer block parameters than TBM. In particular, every non-degree  $r_1$ -block tensor can be represented by a *degree-corrected*  $r_2$ -block tensor with  $r_2 \leq r_1$ . In particular, there exist tensors with  $r_1 = p$  but  $r_2 = 1$ , so the reduction in model complexity can be dramatic from  $p$  to 1. This fact highlights the benefits of introducing degree heterogeneity in higher-order clustering tasks.

### C. Identifiability under angle gap condition

The goal of clustering is to estimate the partition function  $z$  from model (2). For ease of notation, we focus on symmetric tensors; the extension to non-symmetric tensors are similar. We use  $\mathcal{P}$  to denote the following parameter

space for  $(z, \mathcal{S}, \boldsymbol{\theta})$ ,

$$\mathcal{P} = \left\{ (z, \mathcal{S}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}_+^p, \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, c_3 \leq \|\text{Mat}(\mathcal{S})_{a:}\| \leq c_4, \|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r] \right\}, \quad (3)$$

where  $c_i > 0$ 's are universal constants. We briefly describe the rationale of the constraints in (3). First, the entrywise positivity constraint on  $\boldsymbol{\theta} \in \mathbb{R}_+^p$  is imposed to avoid sign ambiguity between entries in  $\boldsymbol{\theta}_{z^{-1}(a)}$  and  $\mathcal{S}$ . This constraint allows the trigonometric cos to describe the angle similarity in the Assumption 1 below and Sub-algorithm 2 in Section IV. Note that the positivity constraint can be achieved without sacrificing model flexibility, by using a slightly larger dimension of  $\mathcal{S}$  in the factorization (2); see Example 1 below. Second, recall that the quantity  $|z^{-1}(a)|$  denotes the number of nodes in  $a$ -th community. The constants  $c_1, c_2$  in the  $|z^{-1}(a)|$  bound assume the roughly balanced size across  $r$  communities. Third, the constant  $c_3$  requires no purely zero slice in  $\mathcal{S}$ , and thus the model (2) is not trivially non-identifiable; see Example 2 below. The upper bound  $c_4$  is a technical constraint to avoid the unbounded entries in the core tensor; in practice, the constraint  $\max_a \|\text{Mat}(\mathcal{S})_{a:}\| \leq c_4$  would likely never be active with a large  $c_4 \geq \|\mathcal{Y}\|_F$ . Lastly, the  $\ell_1$  normalization  $\|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$  is imposed to avoid the scalar ambiguity between  $\boldsymbol{\theta}_{z^{-1}(a)}$  and  $\mathcal{S}$ . This constraint, again, incurs no restriction to model flexibility but makes our presentation cleaner. Our constraints in  $\mathcal{P}$  are mild compared with previous literature; see Table II for comparison.

**Example 1** (Positivity of degree parameters). Here we provide an example to show the positivity constraints on  $\boldsymbol{\theta}$  incurs no loss on the model flexibility. Consider an order-3 dTBM with core tensor  $\mathcal{S} = 1$  and degree  $\boldsymbol{\theta} = (1, 1, -1, -1)^T$ . We have the mean tensor

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \boldsymbol{\Theta} \mathbf{M} \times_3 \boldsymbol{\Theta} \mathbf{M},$$

where  $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta})$  and  $\mathbf{M} = (1, 1, 1, 1)^T$ . Note that  $\mathcal{X} \in \mathbb{R}^{4 \times 4 \times 4}$  is a 1-block tensor with *mixed-signed* degree  $\boldsymbol{\theta}$ , and the mode-3 slices of  $\mathcal{X}$  are

$$\mathcal{X}_{::1} = \mathcal{X}_{::2} = -\mathcal{X}_{::3} = -\mathcal{X}_{::4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

Now, instead of original decomposition, we encode  $\mathcal{X}$  as a 2-block tensor with *positive-signed* degree. Specifically, we write

$$\mathcal{X} = \mathcal{S}' \times_1 \boldsymbol{\Theta}' \mathbf{M}' \times_2 \boldsymbol{\Theta}' \mathbf{M}' \times_3 \boldsymbol{\Theta}' \mathbf{M}',$$

where  $\boldsymbol{\Theta}' = \text{diag}(\boldsymbol{\theta}') = \text{diag}(1, 1, 1, 1)$ , the core tensor  $\mathcal{S}' \in \mathbb{R}^{2 \times 2 \times 2}$  has mode-3 slices, and the membership matrix  $\mathbf{M}' \in \{0, 1\}^{4 \times 2}$  defines the clustering  $z' : [4] \rightarrow [2]$ ,

$$\mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{M}' = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

The triplet  $(z', \mathcal{S}', \boldsymbol{\theta}')$  lies in our parameter space (3). In general, we can always reparameterize a block- $r$  tensor with mixed-signed degree using a block- $2r$  tensor with positive-signed degree. Since we assume  $r = \mathcal{O}(1)$  throughout the paper, the splitting does not affect the error rates of our interest.

**Example 2** (Non-identifiability with purely zero core slice). Consider an order-2 dTBM with core tensor  $\mathcal{S} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$  and mean tensor

$$\mathcal{X} = \boldsymbol{\Theta}_1 \mathbf{M} \mathcal{S} \mathbf{M}^T \boldsymbol{\Theta}_2, \quad \text{with } \mathbf{M} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \text{diag}(1, 1, 1, 1).$$

Replacing  $\boldsymbol{\Theta}_1$  by  $\boldsymbol{\Theta}'_1 = (3/2, 1/2, 1, 1)$  leads to the same mean tensor  $\mathcal{X}$ .

Assumptions in parameter space	Gao et al. (2018)	Han et al. (2020)	Ke et al. (2019)	Ours
Balanced community sizes	✓	✓	✓	✓
Bounded core tensors	✓	✗	✓	✓
Balanced degrees	✓	-	✓	✓
Flexible in-group connections	✗	✓	✓	✓
Gaps among cluster centers	In-between cluster difference	Euclidean gap	Eigen gap	Angle gap

TABLE II: Parameter space comparison between previous work with our assumption.

We now provide the identifiability conditions for our model before estimation procedures. When  $r = 1$ , the decomposition (2) is always unique (up to cluster label permutation) in  $\mathcal{P}$ , because dTBM is equivalent to the rank-1 tensor family under this case. When  $r \geq 2$ , the Tucker rank of signal tensor  $\mathbb{E}\mathcal{Y}$  in (2) is bounded by, but not necessarily equal to, the number of blocks  $r$  (Wang and Zeng, 2019). Therefore, one can not apply the classical identifiability conditions for low-rank tensors to dTBM. Here, we introduce a key separation condition on the core tensor.

**Assumption 1** (Angle gap). Let  $\mathbf{S} = \text{Mat}(\mathcal{S})$ . Assume the minimal gap between normalized rows of  $\mathbf{S}$  is bounded away from zero; i.e.,

$$\Delta_{\min} := \min_{a \neq b \in [r]} \left\| \frac{\mathbf{S}_{a:}}{\|\mathbf{S}_{a:}\|} - \frac{\mathbf{S}_{b:}}{\|\mathbf{S}_{b:}\|} \right\| > 0, \quad \text{for } r \geq 2. \quad (4)$$

We make the convention  $\Delta_{\min} = 1$  for  $r = 1$ . Equivalently, (4) says that none of the two rows in  $\mathbf{S}$  are parallel; i.e.,  $\max_{a \neq b \in [r]} \cos(\mathbf{S}_{a:}, \mathbf{S}_{b:}) = 1 - \Delta_{\min}^2/2 < 1$ . The quantity  $\Delta_{\min}$  characterizes the non-redundancy among clusters measured by angle separation. The denominators involved in definition (4) are well posed because of the lower bound on  $\|\mathbf{S}_{a:}\|$  in (3).

Our first main result is the following theorem showing the sufficiency and necessity of the angle gap separation condition for the parameter identifiability under dTBM.

**Theorem 1** (Model identifiability). Consider the dTBM with  $r \geq 2$ . The parameterization (2) is unique in  $\mathcal{P}$  up to cluster label permutations, if and only if Assumption 1 holds.

The identifiability guarantee for the dTBM is stronger than classical Tucker model. In the Tucker model, the factor matrix  $\mathbf{M}$  is identifiable only up to orthogonal rotations. In contrast, our model does not suffer from rotational invariance. As we will show in Section IV, each column of the membership matrix  $\mathbf{M}$  can be precisely recovered under our algorithm. This property benefits the interpretation of dTBM in practice.

### III. STATISTICAL-COMPUTATIONAL CRITICAL VALUES FOR HIGHER-ORDER TENSORS

#### A. Assumptions

We propose the signal-to-noise ratio (SNR),

$$\text{SNR} := \Delta_{\min}^2 / \sigma^2 = p^\gamma, \quad (5)$$

with varying  $\gamma \in \mathbb{R}$  that quantifies different regimes of interest. We call  $\gamma$  the *signal exponent*. Intuitively, a larger SNR, or equivalently a larger  $\gamma$ , benefits the clustering in the presence of noise.

With quantification (5), we consider the following parameter space,

$$\mathcal{P}(\gamma) = \mathcal{P} \cap \{\mathcal{S} \text{ satisfies SNR condition (5) with } \gamma\}. \quad (6)$$

The 1-block dTBM does not belong to the space  $\mathcal{P}(\gamma)$  when  $\gamma < 0$  by Assumption 1. Our goal is to characterize the clustering accuracy with respect to  $\gamma$  under the space  $\mathcal{P}(\gamma)$ .

In our algorithmic development, we often refer to the regime of balanced degree heterogeneity. We call the degree  $\theta$  *balanced* if

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|. \quad (7)$$

The following lemma provides the rational of balanced degree assumption. We show the close relation between angle gaps in the mean tensor  $\mathcal{X}$  and the core tensor  $\mathcal{S}$  under balanced degree heterogeneity.

**Lemma 1** (Angle gaps in  $\mathcal{X}$  and  $\mathcal{S}$ ). Consider the dTBM model (2) under the parameter space  $\mathcal{P}$  in (3). Suppose Assumption 1 holds and  $\boldsymbol{\theta}$  is balanced satisfying (7). Then, for all  $i, j$  such that  $z(i) \neq z(j)$ , we have

$$\cos(\mathbf{X}_{i:}, \mathbf{X}_{j:}) \asymp \cos(\mathbf{S}_{z(i):}, \mathbf{S}_{z(j):}),$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$  and  $\mathbf{S} = \text{Mat}(\mathcal{S})$ .

In practice, an estimation algorithm has access to a noisy version of  $\mathcal{X}$  but not  $\mathcal{S}$ . Our goal is to establish the algorithm performance with respect to the signal  $\Delta_{\min}^2$  in the core tensor. By Lemma 1, the mapping from the core tensor  $\mathbf{S}_{z(i):}$  to the mean tensor  $\mathbf{X}_{z(i):}$  preserves the angle information  $\Delta_{\min}^2$  under balanced degree heterogeneity (7). Therefore, the balanced degree assumption helps to exclude the cases in which the degree heterogeneity distorts the algorithm guarantees.

Here, we provide an example to illustrate the insufficiency of  $\Delta_{\min}^2$  in the absence of balanced degrees.

**Example 3** (Insufficiency of  $\Delta_{\min}^2$  in the absence of balanced degrees). Consider an order-2  $(p, p)$ -dimensional dTBM with core matrix

$$\mathbf{S} = \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix}, \quad \text{and } \boldsymbol{\theta} \text{ such that } \|\boldsymbol{\theta}_{z^{-1}(1)}\|^2 = p^m \|\boldsymbol{\theta}_{z^{-1}(2)}\|^2, \quad (8)$$

where  $m \geq 0$  is a scalar parameter controlling the skewness of degrees. Let  $\Delta_{\mathbf{X}}^2$  denote the minimal angle gap of the mean tensor, defined by

$$\Delta_{\mathbf{X}}^2 := \min_{i,j \in [p], z(i) \neq z(j)} \left\| \frac{\mathbf{X}_{i:}}{\|\mathbf{X}_{i:}\|} - \frac{\mathbf{X}_{j:}}{\|\mathbf{X}_{j:}\|} \right\|, \quad \mathbf{X} = \text{Mat}_1(\mathcal{X}). \quad (9)$$

Take  $a = p^{-1/4}$  in the model setup (8). We have

$$\Delta_{\min}^2 = \frac{2a^2}{1+a^2} \asymp p^{-1/2}, \quad \Delta_{\mathbf{X}}^2 = \frac{2\|\boldsymbol{\theta}_{z^{-1}(2)}\|^2 a^2}{\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2 + \|\boldsymbol{\theta}_{z^{-1}(2)}\|^2 a^2} \asymp p^{-1/2-m}.$$

Based on our theory in later Sections, the stat-computational achievability of dTBM occurs when  $\Delta_{\mathbf{X}}^2 > p^{-1}$ ; that is, the dTBM estimation depends on the relative magnitude of  $m$  vs.  $1/2$ . In such a setting, the proposed signal notion  $\Delta_{\min}^2$  alone fails to fully characterize dTBM.

**Remark 2** (Flexibility in balanced degree assumption). One important note is that our balance assumption (7) does not preclude the mild degree heterogeneity. In fact, within each of the clusters, we allow the highest degree at the order  $\mathcal{O}(p)$ , whereas the lowest degree at the order  $\Omega(1)$ . This range is more relaxed than previous work (Gao et al., 2018) that restricts the highest degree in the sub-linear regime  $o(p)$  and the lowest degree at the order  $\Omega(1)$ .

**Remark 3** (Similar assumptions in literature). Similar degree regulations are not rare in literature. In higher-order tensor model (Ke et al., 2019), the degree assumption  $\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\| \leq C \min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|$  is made to ensure balancedness across communities. In Gao et al. (2018), the degree distribution is restricted to  $\frac{1}{|z^{-1}(a)|} \sum_{i \in z^{-1}(a)} \theta_i = 1 + o(1)$  for all communities.”

Last, let  $\hat{z}$  and  $z$  be the estimated and true clustering functions in the family (3). Define the misclustering error by

$$\ell(\hat{z}, z) = \frac{1}{p} \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq \pi \circ z(i)\},$$

where  $\pi : [r] \mapsto [r]$  is a permutation of cluster labels,  $\circ$  denotes the composition operation, and  $\Pi$  denotes the collection of all possible permutations. The infinitum over all permutations accounts for the ambiguity in cluster label permutation.

In Sections III-B and III-C, we provide the phase transition of  $\ell(\hat{z}, z)$  for general Gaussian dTBMs (1) without symmetric assumptions. For general (asymmetric) Gaussian dTBMs, we assume Gaussian noise  $\mathcal{E}(i_1, \dots, i_K) \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ , and we extend the parameter space (3) to allow  $K$  clustering functions  $(z_k)_{k \in [K]}$ , one for each mode. For notational simplicity, we still use  $z$  and  $\mathcal{P}(\gamma)$  for this general (asymmetric) model. All results should be interpreted as the worst-case results across  $K$  modes.

### B. Statistical critical value

The statistical critical value means the SNR required for solving dTBMs with *unlimited computational cost*. Our following result shows the minimax lower bound for exact recovery and the matching upper bound for maximum likelihood estimator (MLE). We consider the Gaussian MLE, denoted as  $(\hat{z}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}})$ , over the estimation space  $\mathcal{P}(\gamma)$ , where

$$(\hat{z}_{\text{MLE}}, \hat{\theta}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}) = \arg \min_{(z, \theta, \mathcal{S}) \in \mathcal{P}(\gamma)} \|\mathcal{Y} - \mathcal{X}(z, \theta, \mathcal{S})\|_F^2. \quad (10)$$

**Theorem 2** (Statistical critical value). Consider general Gaussian dTBMs with parameter space  $\mathcal{P}(\gamma)$  with  $K \geq 1$ . Then, we have the following statistical phase transition.

- **Impossibility.** Assume  $r \lesssim p^{1/3}$ . If the signal exponent satisfies  $\gamma < -(K-1)$ , then, no estimator  $\hat{z}_{\text{stat}}$  achieves exact recovery in expectation; that is,

$$\gamma < -(K-1) \Rightarrow \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \mathcal{S}, \theta) \in \mathcal{P}(\lambda)} \mathbb{E}[p\ell(\hat{z}_{\text{stat}}, z)] \geq 1.$$

- **MLE achievability.** Suppose the signal exponent satisfies  $\gamma > -(K-1) + c$  for an arbitrary constant  $c > 0$ . Furthermore, assume  $\theta$  is balanced and  $\min_{i \in [p]} \theta(i) \geq c$  from some constant  $c > 0$ . Then, the MLE in (10) achieves exact recovery in high probability; that is,

$$\gamma > -(K-1) + c \Rightarrow \ell(\hat{z}_{\text{MLE}}, z) \lesssim \text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right) \rightarrow 0,$$

with probability going to 1.

The proofs for the two parts in Theorem 2 are in the Appendix II, Section D and Section G, respectively. The first part of Theorem 2 demonstrates the impossibility of exact recovery when  $\gamma < -(K-1)$ . Inspection of the proof shows that the impossibility holds with or without the balanced degree assumption. The proof is information-theoretical, and therefore the results apply to all statistical estimators, including but not limited to MLE and trace maximization (Ghoshdastidar and Dukkipati, 2017).

The second part of Theorem 2 shows the exact recovery of MLE when the  $\gamma > -(K-1) + c$  for an arbitrary constant  $c > 0$ . Combining the impossibility and achievability results, we conclude that the boundary  $\gamma_{\text{stat}} := -(K-1)$  is the critical value for statistical performance of dTBM with respect to our SNR.

### C. Computational critical value

In this section, we derive the computational critical value of dTBMs. The computational critical value means the minimal SNR required for exactly recovery with *polynomial-time* computational cost. An important ingredient to establish the computational limits is the *hypergraphic planted clique (HPC) conjecture* (Brennan and Bresler, 2020; Zhang and Xia, 2018). The HPC conjecture indicates the impossibility of fully recovering the planted cliques with polynomial-time algorithm when the clique size is less than the number of vertices in the hypergraph. The formal statement of HPC detection conjecture is provided in Definition 1 and Conjecture 1 as follows.

**Definition 1** (Hypergraphic planted clique (HPC) detection). Consider an order- $K$  hypergraph  $H = (V, E)$  where  $V = [p]$  collects vertices and  $E$  collects all the order- $K$  edges. Let  $\mathcal{H}_k(p, 1/2)$  denote the Erdős-Rényi  $K$ -hypergraph where the edge  $(i_1, \dots, i_K)$  belongs to  $E$  with probability  $1/2$ . Further, we let  $\mathcal{H}_K(p, 1/2, \kappa)$  denote the hyhypergraph with planted cliques of size  $\kappa$ . Specifically, we generate a hypergraph from  $\mathcal{H}_k(p, 1/2)$ , pick  $\kappa$  vertices uniformly from  $[p]$ , denoted  $K$ , and then connect all the hyperedges with vertices in  $K$ . Note that the clique size  $\kappa$  can be a

function of  $p$ , denoted  $\kappa_p$ . The order- $K$  HPC detection aims to identify whether there exists a planted clique hidden in an Erdős-Rényi  $K$ -hypergraph. The HPC detection is formulated as the following hypothesis testing problem

$$H_0 : H \sim \mathcal{H}_K(p, 1/2) \quad \text{versus} \quad H_1 : H \sim \mathcal{H}_K(p, 1/2, \kappa_p).$$

**Conjecture 1** (HPC conjecture). Consider the HPC detection problem in Definition 1. Suppose the sequence  $\{\kappa_p\}$  such that  $\limsup_{p \rightarrow \infty} \log \kappa_p / \log \sqrt{p} \leq (1 - \tau)$ . Then, for every sequence of polynomial-time test  $\{\varphi_p\} : H \mapsto \{0, 1\}$  we have

$$\liminf_{p \rightarrow \infty} \mathbb{P}_{H_0}(\varphi_p(H) = 1) + \mathbb{P}_{H_1}(\varphi_p(H) = 0) \geq \frac{1}{2}.$$

Under the HPC conjecture, we establish the SNR lower bound that is necessary for any *polynomial-time* estimator to achieve exact clustering.

**Theorem 3** (Computational critical value). Consider general Gaussian dTBMs under the parameter space  $\mathcal{P}$  with  $K \geq 2$ . Then, we have the following computational phase transition.

- **Impossibility.** Assume HPC conjecture holds. If the signal exponent satisfies  $\gamma < -K/2$ , then, no *polynomial-time estimator*  $\hat{z}_{\text{comp}}$  achieves exact recovery in expectation as  $p \rightarrow \infty$ ; that is,

$$\gamma < -K/2 \Rightarrow \liminf_{p \rightarrow \infty} \sup_{(z, \mathcal{S}, \boldsymbol{\theta}) \in \mathcal{P}(\gamma)} \mathbb{E}[p\ell(\hat{z}_{\text{comp}}, z)] \geq 1.$$

- **Polynomial-time algorithm achievability.** Suppose the parameter space satisfies  $\gamma > -K/2 + c$  for an arbitrary constant  $c > 0$ . Furthermore, assume the degree is balanced, lower bounded in that  $\min_{i \in [p]} \theta_i \geq c$  for some constant  $c > 0$ , and satisfies the local linear stability in Definition 2 in the neighborhood  $\mathcal{N}(z, \varepsilon)$  for  $\varepsilon \geq \log^{-1}(p)$ . Then, there exists a polynomial-time algorithm  $\hat{z}_{\text{poly}}$  that achieves exact recovery in high probability; that is,

$$\gamma > -K/2 + c \Rightarrow \ell(\hat{z}_{\text{poly}}, z) \lesssim \text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right) \rightarrow 0,$$

with probability going to 1.

The proofs for the two parts in Theorem 3 are in the Appendix II, Section E and Section G, respectively. The first part of Theorem 3 indicates the impossibility of exact recovery by polynomial-time algorithms when  $\gamma < -K/2$ , and the second part shows the existence of such algorithm when  $\gamma > -K/2 + c$  under extra technical assumptions. In Section IV, we will present an efficient polynomial-time algorithm in this setting. Therefore, we conclude that  $\gamma_{\text{comp}} := -K/2$  is the critical value for computational performance of dTBM with respect to our SNR.

**Remark 4** (Statistical-computational gaps). Now, we have established the phase transition of exact clustering under order- $K$  dTBM by combining Theorems 2 and 3. Figure 2 summarizes our results of critical SNRs when  $K \geq 2$ . In the weak SNR region  $\gamma < -(K-1)$ , no statistical estimator succeeds in degree-corrected higher-order clustering. In the strong SNR region  $\gamma > -K/2$ , our proposed algorithm precisely recovers the clustering in polynomial time. In the moderate SNR regime,  $-(K-1) \leq \gamma \leq -K/2$ , the degree-corrected clustering problem is statistically easy but computationally hard. Particularly, dTBM reduces to matrix degree-corrected model when  $K = 2$ , and the statistical and computational bounds show the same critical value. When  $K = 1$ , dTBM reduces to the degree-corrected sub-Gaussian mixture model (GMM) with model

$$\mathbf{Y} = \boldsymbol{\Theta} \mathbf{M} \mathbf{S} + \mathbf{E},$$

where  $\mathbf{Y} \in \mathbb{R}^{p \times d}$  collects  $n$  data points in  $\mathbb{R}^d$ ,  $\mathbf{S} \in \mathbb{R}^{r \times d}$  collects the  $d$ -dimensional centroids for  $r$  clusters, and  $\boldsymbol{\Theta} \in \mathbb{R}^{p \times p}$ ,  $\mathbf{M} \in \{0, 1\}^{p \times r}$ ,  $\mathbf{E} \in \mathbb{R}^{p \times d}$  have the same meaning as in dTBM. Lu and Zhou (2016) implies that polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Therefore, we conclude that the statistical-to-computational gap emerges only for higher-order tensors with  $K \geq 3$ . The result reveals the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering.

#### IV. POLYNOMIAL-TIME ALGORITHM UNDER MILD SNR

In this section, we present an efficient polynomial-time clustering algorithm under mild SNR. The procedure takes a global-to-local approach. See Figure 3 for illustration. The global step finds the basin of attraction with polynomial misclustering error, whereas the local iterations improve the initial clustering to exact recovery. Both steps are critical to obtain a satisfactory algorithm output. In what follows, we first use the symmetric tensor as a working example to describe the algorithm procedures to gain insight. Our theoretical analysis focuses on the symmetric mean tensor with independent sub-Gaussian noises such as Gaussian and uniform observations. The extensions for Bernoulli observations, asymmetric mean tensors, and other practical issues are in Sections IV-C and IV-D.

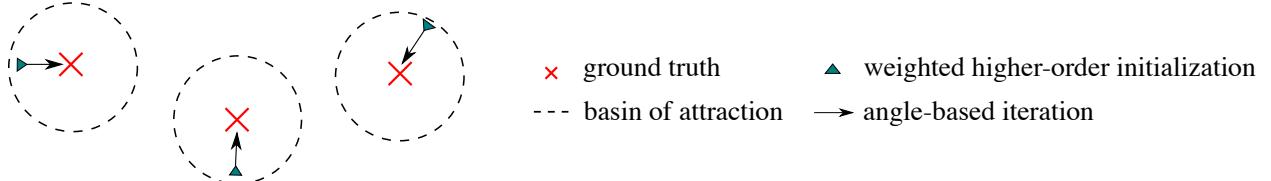


Fig. 3: Illustration of our global-to-local algorithm.

To construct algorithm guarantees, we introduce the misclustering loss between an estimator  $\hat{z}$  and the true  $z$ :

$$L(\hat{z}, z) = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1}\{\hat{z}(i) = b\} \|[\mathbf{S}_{z(i)}]_b^s - [\mathbf{S}_b]_b^s\|^2, \quad (11)$$

where the superscript  $\cdot^s$  denotes the normalized vector; i.e.,  $a^s := a / \|a\|$  if  $a \neq 0$  and  $a^s = 0$  if  $a = 0$  for any vector  $a$ . The following lemma indicates the close relationship between the loss  $L(\hat{z}, z)$  and error  $\ell(\hat{z}, z)$ . The loss  $L(\hat{z}, z)$  serves as an important intermediate quantity to control the misclustering error.

**Lemma 2** (Misclustering error and loss). Suppose  $\min_{i \in [p]} \theta(i) > c$  for some  $c > 0$ . We have  $\ell(\hat{z}, z) \Delta_{\min}^2 \leq L(\hat{z}, z)$ .

##### A. Weighted higher-order initialization

We start with weighted higher-order clustering algorithm as initialization. We take an order-3 tensor and the clustering on the first mode as illustration for insight. Consider noiseless case with  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  and  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . By model (2), for all  $i \in [p]$ , we have

$$\theta(i)^{-1} \mathbf{X}_{i:} = [\text{Mat}(\mathcal{S} \times_2 \Theta \mathbf{M} \times_3 \Theta \mathbf{M})]_{z(i):}.$$

This implies that, all node  $i$  belonging to  $a$ -th community (i.e.,  $z(i) = a$ ) share the same normalized mean vector  $\theta(i)^{-1} \mathbf{X}_{i:}$ , and vice versa. Intuitively, one can apply  $k$ -means clustering to the vectors  $\{\theta(i)^{-1} \mathbf{X}_{i:}\}_{i \in [p]}$ , which leads to main idea of our Sub-algorithm 1.

Specifically, our initialization consists of denoising step and clustering step. The denoising step (lines 1-2 in Sub-algorithm 1) estimates  $\mathcal{X}$  from  $\mathcal{Y}$  by a double projection spectral method. The first projection performs HOSVD (De Lathauwer et al., 2000) via  $\mathbf{U}_{\text{pre},k} = \text{SVD}_r(\text{Mat}_k(\mathcal{Y}))$ ,  $k \in [3]$ , where  $\text{SVD}_r(\cdot)$  returns the top- $r$  left singular vectors. The second projection performs HOSVD on the projected  $\mathcal{Y}$  onto the multilinear Kronecker space  $\mathbf{U}_{\text{pre},k} \otimes \mathbf{U}_{\text{pre},k}$ ; i.e.,

$$\hat{\mathbf{U}}_1 = \text{SVD}_r(\text{Mat}_1(\mathcal{Y} \times_2 \mathbf{U}_{\text{pre},2} \mathbf{U}_{\text{pre},2}^T \times_3 \mathbf{U}_{\text{pre},3} \mathbf{U}_{\text{pre},3}^T)).$$

and similar for  $\hat{\mathbf{U}}_2, \hat{\mathbf{U}}_3$ . The final denoised tensor  $\hat{\mathcal{X}}$  is defined by

$$\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\mathbf{U}}_1 \hat{\mathbf{U}}_1^T \times_2 \hat{\mathbf{U}}_2 \hat{\mathbf{U}}_2^T \times_3 \hat{\mathbf{U}}_3 \hat{\mathbf{U}}_3^T.$$

The double projection improves usual matrix spectral methods in order to alleviate the noise effects for  $K \geq 3$  (Han et al., 2020).

The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted  $k$ -means clustering. We write  $\hat{\mathbf{X}} = \text{Mat}_1(\hat{\mathcal{X}})$ , and normalize the rows into  $\hat{\mathbf{X}}_{i:}^s = \|\hat{\mathbf{X}}_{i:}\|^{-1} \hat{\mathbf{X}}_{i:}$  as a surrogate of  $\theta(i)^{-1} \mathbf{X}_{i:}$ . Then, a weighted  $k$ -means clustering is performed on the normalized rows with weights equal to  $\|\hat{\mathbf{X}}_{i:}\|^2$ . The choice of weights is to bound the  $k$ -means

objective function by the Frobenius-norm accuracy of  $\hat{\mathcal{X}}$ . Unlike existing clustering algorithm (Ke et al., 2019), we apply the clustering on the unfolded tensor  $\hat{\mathcal{X}}$  rather than on the factors  $\hat{\mathbf{U}}_1$ . This strategy relaxes the singular-value gap condition (Gao et al., 2018; Han et al., 2020). We assign degenerate rows with purely zero entries to an arbitrarily random cluster; these nodes are negligible in high-dimensions because of the lower bound on  $\|\text{Mat}(\mathcal{S})_{a:}\|$  in (3). The final result gives the initial cluster assignment  $z^{(0)}$ . Full procedures for clustering are provided in Sub-algorithm 1.

---

**Algorithm: Multiway spherical clustering for degree-corrected tensor block model**


---

**Sub-algorithm 1: Weighted higher-order initialization**


---

**Input:** Observation  $\mathcal{Y} \in \mathbb{R}^{p \times \dots \times p}$ , cluster number  $r$ , relaxation factor  $\eta > 1$  in  $k$ -means clustering.

1: Compute factor matrix  $\mathbf{U}_{\text{pre},k} = \text{SVD}_r(\text{Mat}_k(\mathcal{Y})), k \in [K]$  and the  $(K-1)$ -mode projection

$$\mathcal{X}_{\text{pre},k} = \mathcal{Y} \times_1 \mathbf{U}_{\text{pre},1} \mathbf{U}_{\text{pre},1}^T \times_2 \dots \times_{k-1} \mathbf{U}_{\text{pre},k-1} \mathbf{U}_{\text{pre},k-1}^T \times_{k+1} \mathbf{U}_{\text{pre},k+1} \mathbf{U}_{\text{pre},k+1}^T \times_{K+1} \dots \times_K \mathbf{U}_{\text{pre},K} \mathbf{U}_{\text{pre},K}^T.$$

2: Compute factor matrix  $\hat{\mathbf{U}}_k = \text{SVD}_r(\text{Mat}_k(\mathcal{X}_{\text{pre},k})), k \in [K]$  and denoised tensor

$$\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\mathbf{U}}_1 \hat{\mathbf{U}}_1^T \times_2 \dots \times_K \hat{\mathbf{U}}_K \hat{\mathbf{U}}_K^T.$$

3: **for**  $k \in [K]$  **do**

4: Let  $\hat{\mathbf{X}} = \text{Mat}_k(\hat{\mathcal{X}})$  and  $S_0 = \{i \in [p] : \|\hat{\mathbf{X}}_{i:}\| = 0\}$ . Set  $\hat{z}(i)$  randomly in  $[r]$  for  $i \in S_0$ .

5: For all  $i \in S_0^c$ , compute normalized rows  $\hat{\mathbf{X}}_{i:}^s := \|\hat{\mathbf{X}}_{i:}\|^{-1} \hat{\mathbf{X}}_{i:}$ .

6: Solve the clustering  $\hat{z}_k : [p] \rightarrow [r]$  and centroids  $(\hat{\mathbf{x}}_j)_{j \in [r_k]}$  using weighted  $k$ -means, such that

$$\sum_{i \in S_0^c} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{\hat{z}_k(i)}\|^2 \leq \eta \min_{\bar{\mathbf{x}}_j, j \in [r], \bar{z}_k(i), i \in S_0^c} \sum_{i \in S^c} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \bar{\mathbf{x}}_{\hat{z}_k(i)}\|^2.$$

7: **end for**

**Output:** Initial clustering  $z_k^{(0)} \leftarrow \hat{z}_k, k \in [K]$ .

---

**Sub-algorithm 2: Angle-based iteration**


---

**Input:** Observation  $\mathcal{Y} \in \mathbb{R}^{p \times \dots \times p}$ , initialization  $z_k^{(0)} : [p] \rightarrow [r], k \in [K]$  from Sub-algorithm 1, iteration number  $T$ .

8: **for**  $t = 0$  to  $T-1$  **do**

9: Update the block tensor  $\mathcal{S}^{(t)}$  via  $\mathcal{S}^{(t)}(a_1, \dots, a_K) = \text{Ave}\{\mathcal{Y}(i_1, \dots, i_K) : z_k^{(t)}(i_k) = a_k, k \in [K]\}$ .

10: **for**  $k \in [K]$  **do**

11: Calculate reduced tensor  $\mathcal{Y}_k^d \in \mathbb{R}^{r \times r \times p \times r \times \dots \times r}$  via

$$\mathcal{Y}_k^d(a_1, \dots, a_{k-1}, i, a_{k+1}, \dots, a_K) = \text{Ave}\{\mathcal{Y}(i_1, \dots, i_{k-1}, i, i_{k+1}, \dots, i_K) : z^{(t)}(i_j) = a_j, j \neq k\}$$

12: Let  $\mathbf{Y}_k^d = \text{Mat}_k(\mathcal{Y}_k^d)$  and  $J_0 = \{i \in [p] : \|\mathbf{Y}_k^d_{i:}\| = 0\}$ . Set  $z_k^{(t+1)}(i)$  randomly in  $[r]$  for  $i \in J_0$ .

13: Let  $\mathcal{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$ . For all  $i \in J_0^c$  update the cluster assignment by

$$z(i)_k^{(t+1)} = \arg \max_{a \in [r]} \cos \left( \mathbf{Y}_{k,i:}^d, \mathcal{S}_{a:}^{(t)} \right).$$

14: **end for**

15: **end for**

**Output:** Estimated clustering  $z_k^{(T)} \in [r]^p, k \in [K]$ .

---

We now establish the misclustering error rate of initialization.

**Theorem 4** (Error for weighted higher-order initialization). Consider the general sub-Gaussian dTBM with i.i.d. noise under the parameter space  $\mathcal{P}$  and Assumption 1. Assume  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$  and let  $\Delta_{\mathbf{X}}$  denote the minimal gap in mean tensor defined in (9). Let  $z_k^{(0)}$  denote the output of Sub-algorithm 1. With

probability going to 1, we have

$$\ell(z_k^{(0)}, z) \lesssim \frac{\sigma^2 r^K p^{-K/2}}{\Delta_{\mathbf{X}}^2}.$$

Further, assume  $\boldsymbol{\theta}$  is balanced as (7). We have

$$\ell(z_k^{(0)}, z) \lesssim \frac{r^K p^{-K/2}}{\text{SNR}} \quad \text{and} \quad L(z_k^{(0)}, z) \lesssim \sigma^2 r^K p^{-K/2}. \quad (12)$$

**Remark 5** (Comparison to previous results). For fixed SNR, our initialization error rate with  $K = 2$  agrees with the initialization error rate  $\mathcal{O}(p^{-1})$  in matrix models (Gao et al., 2018). Furthermore, in the special case of non-degree TBMs with  $\theta_1 = \dots = \theta_p = 1$ , we achieve the same initial misclustering error  $\mathcal{O}(p^{-K/2})$  as in non-degree models (Han et al., 2020). Theorem 4 implies the advantage of our algorithm in achieving both accuracy and model flexibility.

**Remark 6** (Failure of conventional tensor HOSVD). If we use conventional HOSVD for tensor denoising; that is, we use  $\mathbf{U}_{\text{pre},k}$  in place of  $\hat{\mathbf{U}}$ ,  $k$  in line 2, then the misclustering rate becomes  $\mathcal{O}(p^{-1})$  for all  $K \geq 2$ . This rate is substantially worse than our current rate (12).

**Remark 7** (Singular-value gap-free clustering). Note that our clustering directly applies to the estimated mean tensor  $\hat{\mathcal{X}}$  rather than the leading tensor factors  $\hat{\mathbf{U}}_k$ . Applying clustering to the tensor factors suffers from the non-identifiability issue due to the infinitely many orthogonal rotations when the number of blocks  $r \geq 3$  in the absence of singular-value gaps. Such ambiguity causes the trouble for effective clustering (Abbe et al., 2020). In contrast, our initialization algorithm applies the clustering to the overall mean tensor  $\hat{\mathcal{X}}$ . This strategy avoids the non-identifiability issue regardless of the number of blocks and singular-value gaps.

### B. Angle-based iteration

Our Theorem 4 has shown the polynomially decaying error rate from our initialization. Now we improve the error rate to exponential decay using local iterations. We propose an angle-based local iteration to improve the outputs from Sub-algorithm 1. To gain the intuition, consider an one-dimensional degree-corrected clustering problem with data vectors  $\mathbf{x}_i = \theta(i)\mathbf{s}_{z(i)} + \epsilon_i$ ,  $i \in [p]$ , where  $\mathbf{s}_i$ 's are known cluster centroids,  $\theta(i)$ 's are unknown positive degrees, and  $z: [p] \mapsto [r]$  is the cluster assignment of interest. The angle-based  $k$ -means algorithm estimates the assignment  $z$  by minimizing the angle between data vectors and centroids; i.e.,

$$z(i) = \arg \max_{a \in [r]} \cos(\mathbf{x}_i, \mathbf{s}_a), \quad \text{for all } i \in [p]. \quad (13)$$

The classical Euclidean-distance based clustering (Han et al., 2020) fails to recover  $z$  in the presence of degree heterogeneity, even under noiseless case. In contrast, the proposed angle-based  $k$ -means achieves accurate recovery without explicit estimation of  $\boldsymbol{\theta}$ .

Our Sub-algorithm 2 shares the same spirit as in angle-based  $k$ -means. We still take the order-3 tensor for illustration. Specifically, Sub-algorithm 2 updates estimated core tensor and cluster assignment in each iteration. We use superscript  $.(t)$  to denote the estimate from  $t$ -th iteration, where  $t = 1, \dots$ . For core tensor, we consider the following update strategy

$$\mathcal{S}^{(t)}(a_1, a_2, a_3) = \text{Ave}\{\mathcal{Y}(i_1, i_2, i_3): z_k^{(t)}(i_k) = a_k, k \in [3]\}.$$

Intuitively,  $\mathcal{S}^{(t)}$  becomes closer to the true core  $\mathcal{S}$  as  $z^{(t)}$  is more precise. For cluster assignment, we first aggregate the slices of  $\mathcal{Y}$  and obtain a reduced tensor  $\mathcal{Y}^d \in \mathbb{R}^{p \times r \times r}$  with given  $z^{(t)}$ , where

$$\mathcal{Y}_1^d(i, a_2, a_3) = \text{Ave}\{\mathcal{Y}(i, i_2, i_3): z_k^{(t)}(i_k) = a_k, k \neq 1\}.$$

We use  $\mathbf{Y}_k^d$  and  $\mathbf{S}_k^{(t)}$  to denote the  $\text{Mat}_k(\mathcal{Y}^d)$  and  $\text{Mat}_k(\mathcal{S}^{(t)})$ . The rows  $\mathbf{Y}_{k,i,:}^d$  and  $\mathbf{S}_{k,a,:}^{(t)}$  correspond to the  $\mathbf{x}_i$  and  $\mathbf{s}_a$  in the one-dimensional clustering (13). Then, we obtain the updated assignment by

$$z_k(i)^{(t+1)} = \arg \max_{a \in [r]} \cos(\mathbf{Y}_{k,i,:}^d, \mathbf{S}_{k,a,:}^{(t)}), \quad \text{for all } i \in [p],$$

provided that  $S_{k,a}^{(t)}$  is a non-zero vector. Otherwise, if  $S_{k,a}^{(t)}$  is a zero vector, then we make the convention to assign  $z_k^{(t+1)}(i)$  randomly in  $[p]$ . Full procedures for our angle-based iteration are described in Sub-algorithm 2.

We now establish the misclustering error rate of iterations under the stability assumption.

**Definition 2** (Locally linear stability). Define the  $\varepsilon$ -neighborhood of  $z$  by  $\mathcal{N}(z, \varepsilon) = \{\bar{z}: \ell(\bar{z}, z) \leq \varepsilon\}$ . Let  $\bar{z}: [p] \rightarrow [r]$  be a clustering function. We define two vectors associated with  $\bar{z}$ ,

$$\mathbf{p}(\bar{z}) = (|\bar{z}^{-1}(1)|, \dots, |\bar{z}^{-1}(r)|)^T, \quad \mathbf{p}_\theta(\bar{z}) = (\|\theta_{\bar{z}^{-1}(1)}\|_1, \dots, \|\theta_{\bar{z}^{-1}(r)}\|_1)^T.$$

We call the degree is  $\varepsilon$ -locally linearly stable if and only if

$$\sin(\mathbf{p}(\bar{z}), \mathbf{p}_\theta(\bar{z})) \lesssim \varepsilon \Delta_{\min}, \quad \text{for all } \bar{z} \in \mathcal{N}(z, \varepsilon). \quad (14)$$

Roughly speaking, the vector  $\mathbf{p}(\bar{z})$  represents the raw cluster sizes, and  $\mathbf{p}_\theta(\bar{z})$  represents the relative cluster sizes weighted by degrees. The local stability holds trivially for  $\varepsilon = 0$  based on the construction of parameter space (3). The condition (14) controls the impact of node degree to the  $p_\theta(\cdot)$  with respect to the misclassification rate  $\varepsilon$  and angle gap. Intuitively, the condition (14) controls the skewness of degree so that the angle between raw cluster size and degree-weighted cluster size is well controlled. The stability assumption is proposed for technical convenience, and we relax this condition in numerical studies; see Section VI.

**Theorem 5** (Error for angle-based iteration). Consider the general sub-Gaussian dTBM with independent noise under the parameter space  $\mathcal{P}$  and Assumption 1. Assume the local linear stability of degree holds in the neighborhood  $\mathcal{N}(z, \varepsilon)$  for  $\varepsilon \geq \log^{-1}(p)$ . Let  $\{z_k^{(0)}\}_{k=1}^K$  be the initialization for Sub-Algorithm 2 and  $z_k^{(t)}$  be the  $t$ -th iteration output on  $k$ -th mode. Suppose  $\min_{i \in [p]} \theta(i) \geq c$  for some  $c > 0$ , the SNR  $\geq \tilde{C} p^{-(K-1)} \log p$  for some sufficiently large constant  $\tilde{C}$ , and the initialization satisfies

$$L(z_k^{(0)}, z) \lesssim \frac{\Delta_{\min}^2}{r \log p}, \quad k \in [K].$$

With probability going to 1, there exists a contraction parameter  $\rho \in (0, 1)$  such that

$$\ell(z, \hat{z}_k^{(t+1)}) \lesssim \underbrace{\text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right)}_{\text{statistical error}} + \underbrace{\rho^t \ell(z, z_k^{(0)})}_{\text{computational error}}. \quad (15)$$

From the conclusion (15), we find that the iteration error is decomposed into two parts: statistical error and computational error. The statistical error is unavoidable with noisy data regardless  $t$ , whereas the computational error decays in an exponential rate as the number of iterations  $t \rightarrow \infty$ .

**Corollary 1** (Exact recovery of dtBM with weighted higher-order initialization). Let the initialization  $\{z_k^{(0)}\}_{k=1}^K$  be the output from Sub-algorithm 1. Assume  $\text{SNR} \gtrsim p^{-K/2} \log p$ . Combining Theorems 4 and 5, with probability going to 1, our estimate  $z_k^{(T)}$  achieves exact recovery within polynomial iterations; more precisely,

$$z_k^{(T)} = \pi_k \circ z, \quad \text{for all } T \gtrsim \log_{1/\rho} p \text{ and } k \in [K].$$

for some permutation  $\pi_k \in \Pi$ .

Therefore, our combined algorithm is *computationally efficient* as long as  $\text{SNR} \gtrsim p^{-K/2} \log p$ . Note that, ignoring the logarithmic term, the minimal SNR requirement,  $p^{-K/2}$ , coincides with the computational lower bound in Theorem 3. Therefore, our algorithm is optimal regarding the signal requirement and lies in the sharpest *computationally efficient* regime in Figure 2.

### C. Extension to Bernoulli observations

Bernoulli or network observations are common in multiple fields. Our iteration Theorem 5 holds for Bernoulli models, but our initialization Theorem 4 does not. Moreover, our current dtBM is insufficient to address sparsity

with decaying mean tensor. Here, we provide extra discussions for Bernoulli initialization and strategies under sparse settings.

- *Extension to dense binary dTBMs.* The main difficulty to establish initialization guarantees for Bernoulli observations lies in the denoising step (lines 1-2 in Sub-algorithm 1). We now provide a high-level explanation for the technical difficulty when applying Theorem 4 to Bernoulli observations.

The derivation of Theorem 4 relies on the upper bound of the estimation error for the mean tensor in Lemma 7; i.e., with high probability

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \quad (16)$$

where  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  and  $\hat{\mathcal{X}}$  is defined in Step 2 of Sub-algorithm 1. Unfortunately, the inequality (16) holds only for i.i.d. sub-Gaussian observations, while Bernoulli observations are generally not identically distributed.

One possible remedy is to apply singular value decomposition to the *square unfolding* (Mu et al., 2014) of Bernoulli tensor  $\mathcal{Y}$ . We summarize the procedure in Sub-algorithm 3.

---

### Sub-algorithm 3: Weighted higher-order initialization for Bernoulli observation

---

**Input:** Bernoulli tensor  $\mathcal{Y} \in \{0, 1\}^{p \times \dots \times p}$ , cluster number  $r$ , relaxation factor  $\eta > 1$  in  $k$ -means clustering.

1: Let the matrix  $\text{Mat}_{sq}(\mathcal{Y}) \in \{0, 1\}^{\lfloor p^{K/2} \rfloor \times \lceil p^{K/2} \rceil}$  denote the nearly square unfolded tensor. Compute the estimate  $\hat{\mathcal{X}}'$ , where

$$\hat{\mathcal{X}}' = \arg \min_{\text{rank}(\text{Mat}_{sq}(\mathcal{X})) \leq r^{\lceil K/2 \rceil}} \|\text{Mat}_{sq}(\mathcal{X}) - \text{Mat}_{sq}(\mathcal{Y})\|_F^2. \quad (17)$$

2: Implement lines 3-5 of Sub-algorithm 1 with  $\hat{\mathcal{X}}$  replaced by  $\hat{\mathcal{X}}'$  in (17).

**Output:** Initial clustering  $z^{(0)} \leftarrow \hat{z}$ .

---

**Proposition 1** (Error for Bernoulli initialization). Consider the Bernoulli dTBM in the parameter space  $\mathcal{P}$  and Assumption 1 holds. Assume that  $\theta$  is balanced and  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$ . Let  $z^{(0)}$  denote the output of Sub-algorithm 3. With probability going to 1, we have

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-\lfloor K/2 \rfloor}}{\text{SNR}}, \quad \text{and} \quad L(z^{(0)}, z) \lesssim \sigma^2 r^K p^{-\lfloor K/2 \rfloor}.$$

*Proof of Proposition 1.* Following Lemma 7 in Gao et al. (2018), with high probability, the new estimate satisfies

$$\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2 \lesssim p^{\lceil K/2 \rceil}.$$

We obtain the upper bound by using  $\hat{\mathcal{X}}'$  in place of  $\hat{\mathcal{X}}$  in Theorem 4.  $\square$

**Remark 8** (Comparison with previous methods). The Bernoulli bound  $\mathcal{O}(p^{-\lfloor K/2 \rfloor})$  in Proposition 1 is relatively looser than the Gaussian bound  $\mathcal{O}(p^{-K/2})$  in Theorem 4. The gap between Bernoulli and Gaussian error decreases as the order  $K$  increases. Nevertheless, combining with angle iteration Sub-algorithm 2, Bernoulli clustering still achieves exponential error rate  $\exp(-p^{(K-1)})$  at a price of a larger SNR. The exponential rate is tighter than the rate error  $\mathcal{O}(p^{-2})$  in previous work (Ke et al., 2019). The investigation of the gap between upper bound  $p^{-\lfloor K/2 \rfloor}$  and the lower bound  $p^{-K/2}$  for Bernoulli tensors will be left as future work. In numerical experiments, we will use our original initialization, Sub-algorithm 1, to verify the robustness to Bernoulli observations.

- *Extension to sparse binary dTBMs.* The sparsity is often a popular feature in hypergraphs (Florescu and Perkins, 2016; Ke et al., 2019). Specifically, the sparse binary dTBM assumes that, the entries of  $\mathcal{Y}$  follow independent Bernoulli distributions with the mean

$$\mathbb{E}\mathcal{Y} = \alpha_p \mathcal{S} \times_1 \Theta M \times_2 \dots \times_K \Theta M, \quad (18)$$

where the extra scalar parameter  $\alpha_p \in (0, 1]$  is function of  $p$  that controls the sparsity. A smaller  $\alpha_p$  indicates a higher level of sparsity. Our current work focuses on dense dTBM with  $\alpha_p = 1$ . While sparse dTBM is an interesting application, the algorithm and its analysis require different techniques. Below, we discuss possible modifications of the algorithm.

The sparsity affects our initialization guarantee in our Theorem 4. In our initialization, the spectral denoising step (lines 1-2 in Sub-algorithm 1) implements matrix SVD to unfolded tensors. However, SVD-based methods are believed to fail in extremely sparse SBM due to the localization phenomenon in the singular vectors (Florescu and Perkins, 2016). Inspired by Florescu and Perkins (2016), we adapt the diagonal-deleted HOSVD (D-HOSVD) (Ke et al., 2019) as the initialization in our higher-order clustering.

The sparsity also affects the iteration guarantee in our Theorem 5. The decaying mean tensor leads to a worse statistical error of order  $\mathcal{O}(-\alpha_p p^{K-1})$  on  $\hat{\mathcal{X}}$ . The theoretical analyses for sparse binary dTBM and algorithms are left as future direction. Instead, we add numerical experiments to evaluate the robustness of our algorithm and the improvement of D-HOSVD initialization in the sparse dTBM; see Appendix I.

#### D. Practical issues

**Computational complexity.** Our two-stage algorithm has a computational cost polynomial in tensor dimension  $p$ . Specifically, the complexity of Sub-algorithm 1 is  $\mathcal{O}(Kp^{K+1} + Krp^K)$ , where the first term is contributed by the double projection and the calculation of  $\hat{\mathcal{X}}$ , and the second term comes from normalization and the  $k$ -means. The cost of each update in Sub-algorithm 2 is  $\mathcal{O}(p^K + pr^K)$ , where  $p^K$  comes from the calculation of  $\mathcal{S}^{(t)}$  and  $\mathcal{Y}^d$ , and  $pr^K$  comes from the normalization of  $\mathcal{Y}^d$ , the calculation of  $\mathcal{S}^{(t)}$ , and the cluster assignment update in Step 10.

**Hyper-parameter selection.** In our theoretical analysis, we have assumed the true cluster number  $r$  is given to our algorithm. In practice, the cluster number  $r$  is often unknown, and we now propose a method to choose  $r$  from data. We impose the Bayesian information criterion (BIC) and choose the cluster number that minimizes BIC; i.e., under the symmetric Gaussian dTBM (2),

$$\hat{r} = \arg \min_{r \in \mathbb{Z}_+} \left( p^K \log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_F^2) + p_e(r)K \log p \right), \quad (19)$$

$$\text{where } \hat{\mathcal{X}} = \hat{\mathcal{S}}(r) \times_1 \hat{\Theta}(r) \hat{\mathcal{M}}(r) \times_2 \cdots \times_K \hat{\Theta}(r) \hat{\mathcal{M}}(r),$$

where the triplet  $(\hat{z}(r), \hat{\mathcal{S}}(r), \hat{\theta}(r))$  are estimated parameters with cluster number  $r$ , and  $p_e(r) = r^K + p(\log r + 1) - r$  is the effective number of parameters. Note that we have added the argument  $(r)$  to related quantities as functions of  $r$ . In particular, the estimate  $\hat{\theta}(r)$  in (19) is obtained by first calculating the reduced tensor  $\hat{\mathcal{Y}}^d$  with  $\hat{z}(r)$ , and then normalizing the row norms  $\|\hat{\mathcal{Y}}_{i:}^d\|$  to 1 in each cluster; i.e.,

$$\hat{\theta}(r) = (\hat{\theta}(1, r), \dots, \hat{\theta}(p, r))^T, \quad \text{where } \hat{\theta}(i, r) = \frac{\|\hat{\mathcal{Y}}^d(r)_{i:}\|}{\sum_{j:\hat{z}(j,r)=\hat{z}(i,r)} \|\hat{\mathcal{Y}}^d(r)_{j:}\|},$$

where  $\hat{\mathcal{Y}}^d(r) = \text{Mat}(\hat{\mathcal{Y}}^d(r))$ ,  $\hat{\mathcal{Y}}^d(r)(i, a_2, \dots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \dots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}$ , and  $\hat{z}(i, r)$  denotes the community label for the  $i$ -th node with given cluster number  $r$ . We evaluate the performance of the BIC criterion in Section VI-A.

## V. COMPARISON WITH NON-DEGREE TENSOR BLOCK MODEL

We discuss the connections and difference between our dTBM and TBM (Han et al., 2020) from three aspects: signal notions, theoretical results, and algorithms. Without loss of generality, let  $\sigma^2 = 1$ .

- *Signal notion.* The signal levels in both TBM (Han et al., 2020) and our dTBM are functions of the core tensor  $\mathcal{S}$ . We emphasize that the signal notions are different between the two models. In particular, the Euclidean-based signal notion in TBM Han et al. (2020) fails to accurately describe the phase transition in our dTBM due to the possible heterogeneity in degree  $\theta$ . To compare, we denote our angle-based signal notion in (5) and the Euclidean-based SNR in Han et al. (2020) as  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$ , respectively:

$$\Delta_{\text{ang}}^2 = 2(1 - \max_{a \neq b \in [r]} \cos(\mathcal{S}_{a:}, \mathcal{S}_{b:})), \quad \Delta_{\text{Euc}}^2 = \min_{a \neq b \in [r]} \|\mathcal{S}_{a:} - \mathcal{S}_{b:}\|^2.$$

By Lemma 4 in the Appendix II, we have

$$\Delta_{\text{ang}}^2 \max_{a \in [r]} \|S_{a:}\|^2 \leq \Delta_{\text{Euc}}^2.$$

The above inequality indicates the condition  $\Delta_{\text{Euc}}^2 \leq p^\gamma$  is sufficient but not necessary for  $\Delta_{\text{ang}}^2 \leq p^\gamma$ . In fact, if we were to use  $\Delta_{\text{Euc}}^2$  for both models, then the phase transition of dTBM can be arbitrarily worse than that for TBM.

Here, we provide an example to illustrate the dramatical difference between TBM and dTBM with the same core tensor.

**Example 4** (Comparison with Euclidean-based signal notion). Consider a biclustering model with  $\theta = 1$  and an order-2 core matrix

$$S = \begin{pmatrix} p^{(\gamma+1)/2} + 2 & 2p^{(\gamma+1)/2} + 4 \\ 2 & 4 \end{pmatrix}, \quad \text{with } \gamma \leq -1.$$

The core matrix  $S$  lies in the parameter spaces of TBM and our dTBM. Here, the constraint  $\gamma \leq -1$  is added to ensure the bounded condition of  $S$  in our parameter space in (3). The angle-based and Euclidean-based signal levels of  $S$  are

$$\Delta_{\text{ang}}^2(S) = 0 \ (\leq p^\gamma), \quad \Delta_{\text{Euc}}^2(S) = 5p^{\gamma+1} \ (\geq p^\gamma).$$

We conclude that TBM with  $S$  achieves exact recovery with a polynomial-time algorithm; see Han et al. (2020, Theorem 4). By contrast, the dTBM with the same  $S$  and input  $r = 2$  violates the identifiability condition, and thus fails to be solved by all estimators; see our Theorem 1.

- *Theoretical results.* In both works, we study the phase transition of TBM and dTBM with respect to the Euclidean and angle-based SNRs. We briefly summarize the results in Han et al. (2020) and compare with ours.

*Statistical critical value:*

$$\begin{aligned} \text{Ours: } \Delta_{\text{ang}}^2 \lesssim p^{-(K-1)} &\Rightarrow \text{statistically impossible}; \quad \Delta_{\text{ang}}^2 \gtrsim p^{-(K-1)} \Rightarrow \text{MLE achieves exact recovery}; \\ \text{Han's: } \Delta_{\text{Euc}}^2 \lesssim p^{-(K-1)} &\Rightarrow \text{statistically impossible}; \quad \Delta_{\text{Euc}}^2 \gtrsim p^{-(K-1)} \Rightarrow \text{MLE achieves exact recovery}. \end{aligned}$$

*Computational critical value:*

$$\begin{aligned} \text{Ours: } \Delta_{\text{ang}}^2 \lesssim p^{-K/2} &\Rightarrow \text{computationally impossible}; \quad \Delta_{\text{ang}}^2 \gtrsim p^{-K/2} \Rightarrow \text{computationally efficient}; \\ \text{Han's: } \Delta_{\text{Euc}}^2 \lesssim p^{-K/2} &\Rightarrow \text{computationally impossible}; \quad \Delta_{\text{Euc}}^2 \gtrsim p^{-K/2} \Rightarrow \text{computationally efficient}. \end{aligned}$$

The above comparison reveals three major differences.

First, none of our results in Section III are corollaries of Han et al. (2020). Both models show the similar conclusion but under different conditions. While the TBM impossibility (Han et al., 2020) provides a necessary condition for our dTBM impossibility, we find that such a condition is often loose. There exists a regime of  $S$  in which TBM problems are computationally efficient but dTBM problems are statistically impossible; see Example 4. This observation has motivated us to develop the new signal notion  $\Delta_{\text{ang}}^2$  for sharp dTBM phase transition conditions.

Second, to find the phase transition, we need to show both the impossibility and achievability when SNR is below and above the critical value, respectively. While the TBM impossibility can serve as a loose condition of our dTBM impossibility, more efforts are required to show the achievability. In particular, since TBM is a more restrictive model than dTBM, the achievability in Han et al. (2020) does not imply the achievability of dTBM in a larger parameter space. The latter requires us to develop new MLE and polynomial algorithms for dTBM achievability.

Third, from the perspective of proofs, we develop new tools to handle the extra degree heterogeneity. In our Theorem 2, we consider the profile MLE by maximizing out the nuisance degree parameter, while TBM (Han et al., 2020) considers the usual MLE without degree parameter. In our Theorem 3, we construct a rank-2 tensor to relate HPC conjecture to  $\Delta_{\text{ang}}^2$ , while TBM (Han et al., 2020) constructs a rank-1 tensor to relate HPC conjecture to  $\Delta_{\text{Euc}}^2$ . The asymptotic non-equivalence between  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$  renders our proof technically more involved.

- *Algorithms.* Both Han et al. (2020) and our work propose the two-step algorithm, which combines warm initialization and iterative refinement to achieve exact recovery. This local-to-global strategy is not new in clustering literature (Chien et al., 2019; Gao and Zhang, 2022). The highlight of our algorithm is the angle-based update in lines 10-14, Sub-algorithm 2, which is specifically designed for dTBM to avoid the estimation of  $\theta$ . This angle-based update brings new proof challenges. We develop polar-coordinate based techniques to establish the error rate for the proposed algorithm. ”

## VI. NUMERICAL STUDIES

We evaluate the performance of the weighted higher-order initialization and angle-based iteration in this section. We report average errors and standard deviations across 30 replications in each experiment. Clustering accuracy is assessed by clustering error rate (CER, i.e., one minus rand index). Note that CER between  $(\hat{z}, z)$  is equivalent to misclustering error  $\ell(\hat{z}, z)$  up to constant multiplications (Meilă, 2012), and a lower CER indicates a better performance.

We generate order-3 tensors with *assortative* (Gao et al., 2018) core tensors to control SNR; i.e., we set  $S_{aaa} = s_1$  for  $a \in [r]$  and others be  $s_2$ , where  $s_1 > s_2 > 0$ . Let  $\alpha = s_1/s_2$ . We set  $\alpha$  close to 1 such that  $1 - \alpha = o(p)$ . In particular, we have  $\alpha = 1 + \Omega(p^{\gamma/2})$  with  $\gamma < 0$  by Assumption 1 and definition (5). Hence, we easily adjust SNR via varying  $\alpha$ . Note that the assortative setting is proposed for simulations, and our algorithm is applicable for general tensors in practice. The cluster assignment  $z$  is randomly generated with equal probability across  $r$  clusters for each mode. Without further explanation, we generate degree heterogeneity  $\theta$  from absolute normal distribution by  $\theta(i) = |X_i| + 1 - 1/\sqrt{2\pi}$  with  $|X_i| \stackrel{i.i.d.}{\sim} N(0, 1)$ ,  $i \in [p]$  and normalize  $\theta$  to satisfy (3). Also, we set  $\sigma^2 = 1$  for Gaussian data without further specification.

### A. Verification of theoretical results

The first experiment verifies statistical-computational gap described in Section III. Consider the Gaussian model with  $p = \{80, 100\}$ ,  $r = 5$ . We vary  $\gamma$  in  $[-1.2, -0.4]$  and  $[-2.1, -1.4]$  for matrix ( $K = 2$ ) and tensor ( $K = 3$ ) clustering, respectively. Note that finding MLE under dTBM is computationally intractable. We approximate MLE using an oracle estimator, i.e., the output of Sub-algorithm 2 initialized from true assignment. Figure 4a shows that both our algorithm and oracle estimator start to decrease around the critical value  $\gamma_{\text{stat}} = \gamma_{\text{comp}} = -1$  in matrix case. In contrast, Figure 4b shows a significant gap in the phase transitions between the algorithm estimator and oracle estimator in tensor case. The oracle error rapidly decreases to 0 when  $\gamma_{\text{stat}} = -2$ , whereas the algorithm estimator tends to achieve exact clustering when  $\gamma_{\text{comp}} = -1.5$ . Figure 4 confirms the existence of the statistical-computational gap in our Theorems 2 and 3.

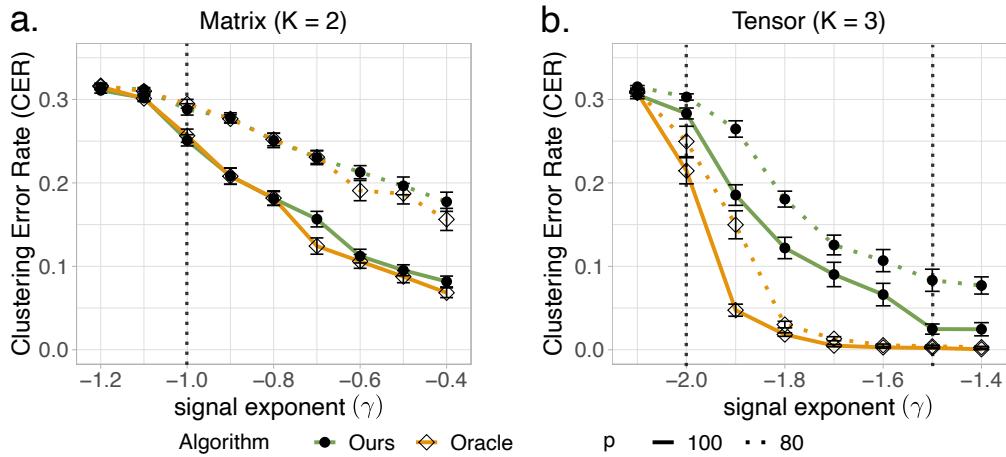


Fig. 4: SNR phase transitions for clustering in dTBM with  $p = \{80, 100\}$ ,  $r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

The second experiment verifies the performance guarantees of two algorithms: (i) weighted higher-order initialization; (ii) combined algorithm of weighted higher-order initialization and angle-based iteration. We consider both the Gaussian and Bernoulli models with  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$ . Figure 5 shows the substantial improvement of combined algorithm over initialization, especially under weak and intermediate signals. This phenomenon agrees with the error rates in Theorems 4 and 5 and confirms the necessity of the local iterations.

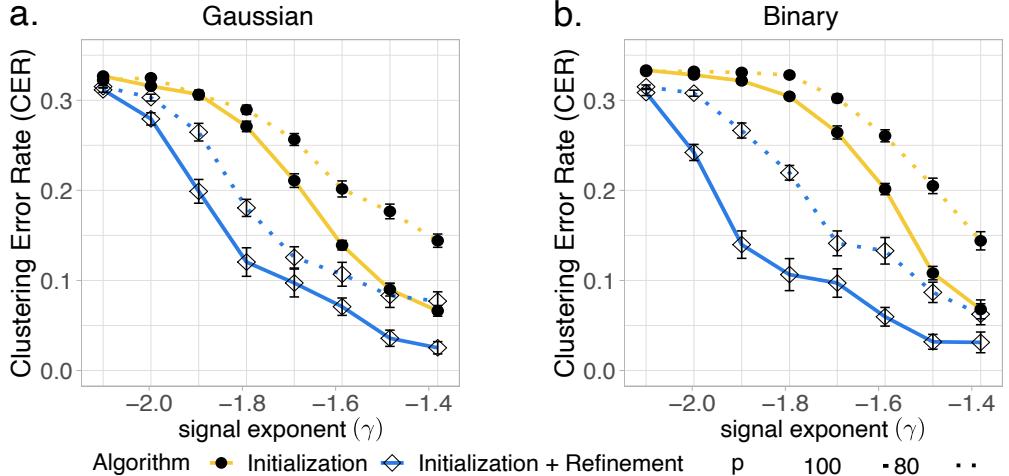


Fig. 5: CER versus signal exponent ( $\gamma$ ) for initialization only and for combined algorithm. We set  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$  under (a) Gaussian models and (b) Bernoulli models.

Settings	$p = 50, \sigma^2 = 0.25$		$p = 50, \sigma^2 = 1$		$p = 80, \sigma^2 = 0.25$		$p = 80, \sigma^2 = 1$	
True cluster number $r$	2	4	2	4	2	4	2	4
Estimated cluster number $\hat{r}$	2(0)	3.9(0.25)	2(0)	3.1(0.52)	2(0)	4(0)	2(0)	3.9(0.31)

TABLE III: Estimated cluster number given by BIC criterion under the low noise level ( $\sigma^2 = 0.25$ ) and high noise level ( $\sigma^2 = 0.5$ ) settings. Numbers in parentheses are standard deviations of  $\hat{r}$  over 30 replications.

The third experiment evaluates the empirical performance of the BIC criterion to select unknown cluster number. We generate the data from an order-3 Gaussian model with  $p = \{50, 80\}$ ,  $r = \{2, 4\}$ , and noise level  $\sigma^2 \in \{0.25, 1\}$ . Table III shows that our BIC criterion well chooses the true  $r$  under most settings. Note that the BIC slightly underestimates the true cluster number ( $r = 4$ ) with smaller dimension and higher noise ( $p = 50, \sigma = 1$ ), and the accuracy immediately increases with larger dimension  $p = 80$ . The improvement follows from the fact that a larger dimension  $p$  indicates a larger sample size in the tensor block model. Therefore, we conclude that BIC criterion is a reasonable way to tune the cluster number.

### B. Comparison with other methods

We compare our algorithm with following higher-order clustering methods:

- **HOSVD**: HOSVD on data tensor and  $k$ -means on the rows of the factor matrix;
- **HOSVD+**: HOSVD on data tensor and  $k$ -means on the  $\ell_2$ -normalized rows of the factor matrix;
- **HLloyd** (Han et al., 2020): High-order clustering algorithm developed for non-degree tensor block models;
- **SCORE** (Ke et al., 2019): Tensor-SCORE for clustering developed for sparse binary tensors.

Among the four alternative algorithms, the **SCORE** is the closest method to ours. We set the tuning parameters of **SCORE** as in previous literature (Ke et al., 2019). The methods **SCORE** and **HOSVD+** are designed for degree models, whereas **HOSVD** and **HLloyd** are designed for non-degree models. We conduct two experiments to assess

the impacts of (i) signal strength and (ii) degree heterogeneity, based on Gaussian and Bernoulli models with  $p = 100, r = 5$ . We refer to our algorithm as **dTBM** in the comparison.

We investigate the effects of signal to clustering performance by varying  $\gamma \in [-1.5, -1.1]$ . Figure 6 shows that our method **dTBM** outperforms all other algorithms. The sub-optimality of **SCORE** and **HOSVD+** indicates the necessity of local iterations on the clustering. Furthermore, Figure 6 shows the inadequacy of non-degree algorithms in the presence of mild degree heterogeneity. The experiment demonstrates the benefits of addressing heterogeneity in higher-order clustering tasks.

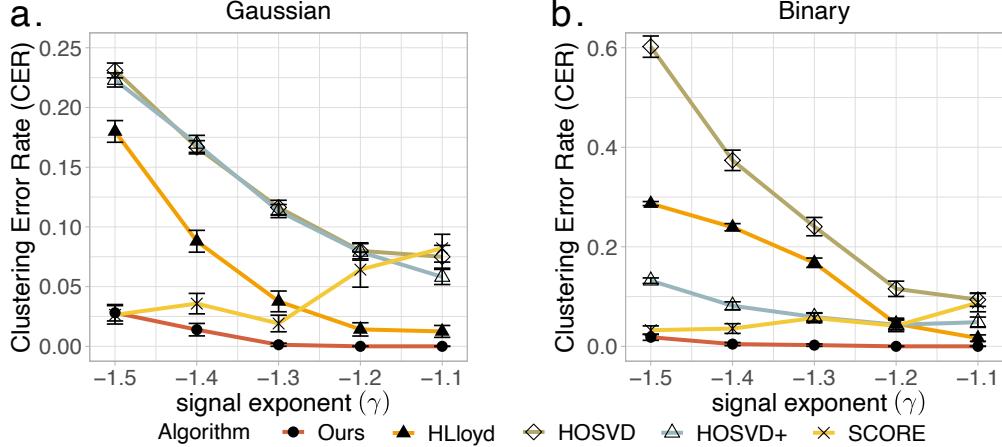


Fig. 6: CER versus signal exponent (denoted  $\gamma$ ) for different methods. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under (a) Gaussian and (b) Bernoulli models.

The only exception in Figure 6 is the slightly better performance of **HLlloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLlloyd** disappears with higher degree heterogeneity. We perform extra simulations to verify the impact of degree effects. We use the same setting as in the first experiment in the Section VI-B, except that we now generate the degree heterogeneity  $\theta$  from Pareto distribution prior to normalization. The density function of Pareto distribution is  $f(x|a, b) = ab^a x^{-(a+1)} \mathbb{1}\{x \geq b\}$ , where  $a$  is called *shape* parameter. We vary  $a \in \{2, 6\}$  and choose  $b$  such that  $\mathbb{E}X = a(a-1)^{-1}b = 1$  for  $X$  following Pareto( $a, b$ ). Note that a smaller  $a$  leads to a larger variance in  $\theta$  and hence a larger degree heterogeneity. We consider the Gaussian model under low ( $a = 6$ ) and high ( $a = 2$ ) degree heterogeneity. Figure 8 shows that the errors for non-degree algorithms (**HLlloyd**, **HOSVD**) increase with degree heterogeneity. In addition, the advantage of **HLlloyd** over **HOSVD+** disappears with higher degree heterogeneity.

The last experiment investigates the effects of degree heterogeneity to clustering performance. We fix the signal exponent  $\gamma = -1.2$  and vary the extent of degree heterogeneity. In this experiment, we generate  $\theta$  from Pareto distribution prior to normalization. We vary the shape parameter  $a \in [3, 6]$  in the Pareto distribution to investigate a range of degree heterogeneities. Figure 7 demonstrates the stability of degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over the entire range of degree heterogeneity under consideration. In contrast, non-degree algorithms (**HLlloyd**, **HOSVD**) show poor performance with large heterogeneity, especially in Bernoulli cases. This experiment, again, highlights the benefit of addressing degree heterogeneity in higher-order clustering.

## VII. REAL DATA APPLICATIONS

### A. Human brain connectome data analysis

The Human Connectome Project (HCP) aims to construct the structural and functional neural connections in human brains (Van Essen et al., 2013). We preprocess the original dataset following Desikan et al. (2006) and partition the brain into 68 regions. The cleaned dataset includes brain networks for 136 individuals. Each brain network is represented by a 68-by-68 binary symmetric matrix, where the entry with value 1 indicates the presence of

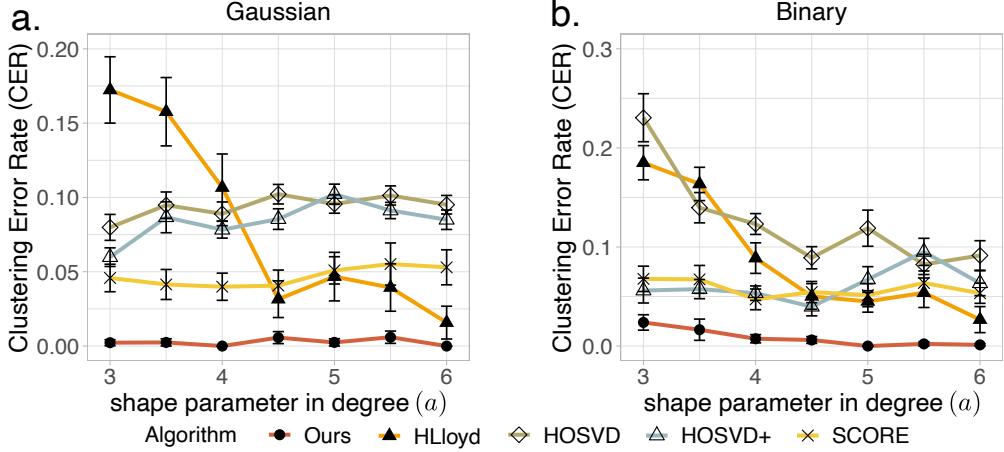


Fig. 7: CER versus shape parameter in degree (denoted  $a \in [3, 6]$ ) for different methods. We set  $p = 100, r = 5, \gamma = -1.2$  under (a) Gaussian and (b) Bernoulli models.

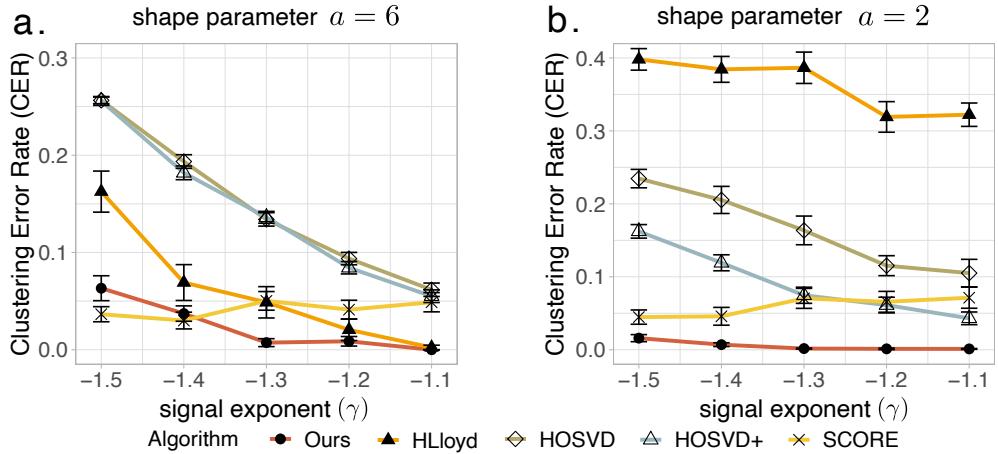


Fig. 8: CER comparison versus signal exponent (denoted  $\gamma$ ) under (a) low (shape parameter  $a = 6$ ) (b) high (shape parameter  $a = 2$ ) degree heterogeneity. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under Gaussian model.

connection between node pairs, while the value 0 indicates the absence. We use  $\mathcal{Y} \in \{0, 1\}^{68 \times 68 \times 136}$  to denote the binary tensor. Individual attributes such as gender and sex are recorded.

We apply our general asymmetric algorithm to the HCP data with the numbers of clusters on three modes  $r_1 = r_2 = 4$  and  $r_3 = 3$ . The selection of  $r_1$  and  $r_2$  follows the human brain anatomy and the symmetry in the brain network, and the  $r_3$  is specified following previous analysis (Hu et al., 2021). Because of the symmetry in the data, the estimated brain node clustering results are the same on the first and second modes. Figure 9 shows that brain connection exhibits a strong spatial separation structure. Specifically, the first cluster, named *L.Hemis*, involves all the nodes in the left hemisphere. The nodes in the right hemisphere are further separated into three clusters led by the middle-part tissues in Temporal and Parietal lobes (*R.Temporal*), the back-part tissues in Occipital lobe (*R.Occipital*), and the front-part tissues in Frontal and Parietal lobes (*R.Supra*). This clustering result is reasonable since the left and right hemispheres often play different roles in human brains.

Figure 10 illustrates the estimated core tensor  $\hat{\mathcal{S}}$  with estimated clustering, and Figure 11 visualizes the average brain connections and the connection enrichment in contrast to average networks in each group. In general, we find that the inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Figure 10a).

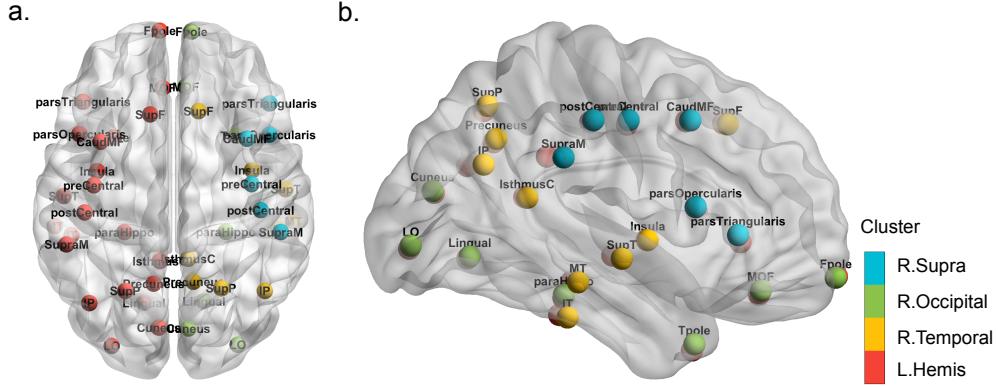


Fig. 9: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

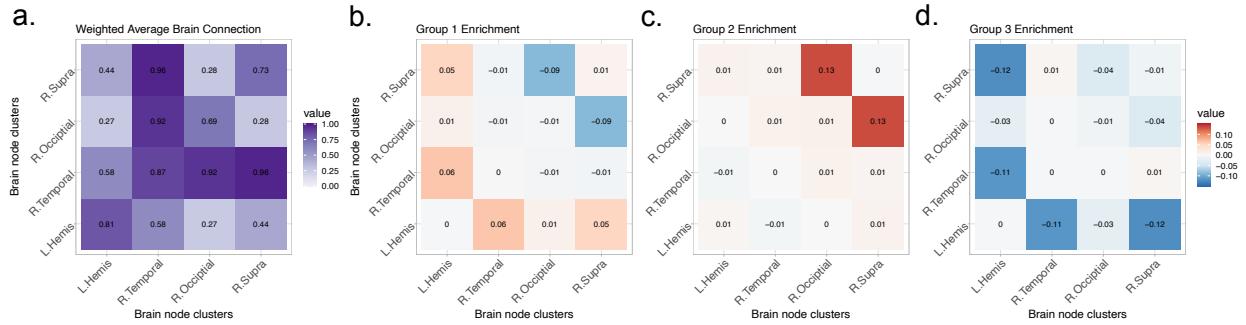


Fig. 10: Mode 3 slices of estimated core tensor  $\hat{S}$ . (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the difference between each slice of  $\hat{S}$  and the averaged slice.

Also, the back and front parts (*R.Occipital*, *R.Supra*) are shown to have more interactions with temporal tissues than inner-cluster connections. In addition, the group 1 with 54% females shows an enrichment on the inter-hemisphere connections (Figure 10b), while group 4 with only 36% females exhibits a reduction (Figure 10d). This result agrees with previous findings in Hu et al. (2021). The enrichment on the back-front connection is also recognized in group 3 (Figure 10c). The interpretive patterns in our results demonstrate the usefulness of our clustering methods

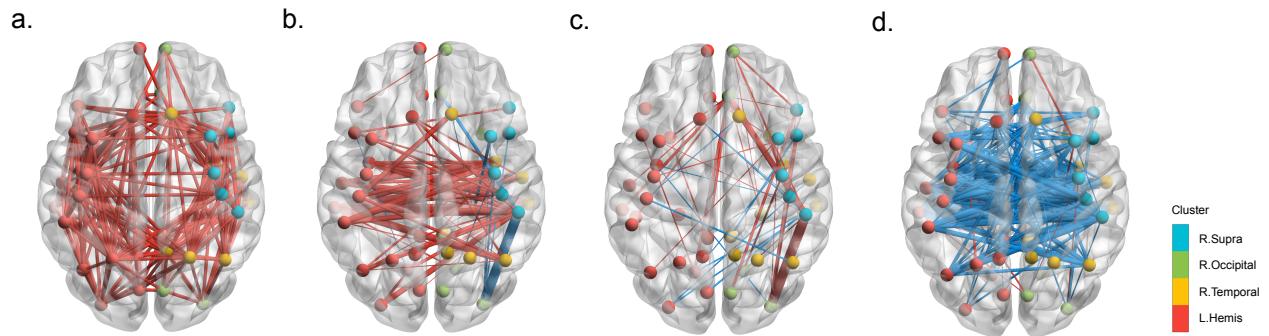


Fig. 11: Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain network enrichments in Groups 1-3. Red edges represent the positive enrichment and blue edges represent the negative enrichment.

in the human brain connectome data application.

### B. Peru Legislation data analysis

We also apply our method to the legislation networks in the Congress of the Republic of Peru (Lee et al., 2017). Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the data for the first half of 2006-2007 year. The dataset records the co-sponsorship of 116 legislators from top 5 parties and 802 bill proposals. We reconstruct legislation network as an order-3 binary tensor  $\mathcal{Y} \in \{0, 1\}^{116 \times 116 \times 116}$ , where  $\mathcal{Y}_{ijk} = 1$  if the legislators  $(i, j, k)$  have sponsored the same bill, and  $\mathcal{Y}_{ijk} = 0$  otherwise. The true party affiliations of legislators are provided and serve as the ground truth. We apply various higher-order clustering methods to  $\mathcal{Y}$  with  $r = 5$ . Table IV shows that our **dTBM** achieves the best performance compared to others. The second best method is the two-stage algorithm **HLloyd**, followed by the spectral methods **SCORE** and **HOSVD+**. This result is consistent with our simulations under strong signal and moderate degree heterogeneity. The comparison suggests that our method **dTBM** is more appealing in real-world applications.

Method	<b>dTBM</b>	<b>HOSVD</b>	<b>HOSVD+</b>	<b>HLloyd</b>	<b>SCORE</b>
CER	<b>0.116</b>	0.22	0.213	0.149	0.199

TABLE IV: Clustering errors (measured by CER) for various methods in the analysis of Peru Legislation dataset.

## VIII. PROOF SKETCHES

In this section, we provide the proof sketches for the main Theorems 4-5. Detail proofs and extra theoretical results are provided in Appendix II.

### A. Proof sketch of Theorem 4

The proof of Theorem 4 is inspired by the proof idea of Gao et al. (2018, Lemma 1). The extra difficulties are the angle gap characterization and multilinear algebra property in tensors; we address both challenges in our proof. Specifically, we control the misclustering error by the estimation error of  $\hat{\mathcal{X}}$  calculated in Step 2 of Sub-algorithm 1. We prove the following inequality

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K} \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \quad (20)$$

where  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  is the true mean. The first inequality in (20) holds with the assumption  $\min_{i \in [p]} \theta(i) \geq c > 0$  in Theorem 4. The second inequality relies on an important conclusion that the angle gap of mean tensor  $\mathcal{X}$  is lower bounded by that of core tensor  $\mathcal{S}$ , i.e., the minimal angle gap  $\Delta_{\min}$  defined in Assumption 1. Let  $\mathbf{a}^s := \mathbf{a} / \|\mathbf{a}\|$  denote the normalized vector with the convention that  $\mathbf{a}^s = 0$  if  $\mathbf{a} = 0$ . We want to show that

$$\min_{z(i) \neq z(j)} \|[\mathbf{X}_{i:}]^s - [\mathbf{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \quad (21)$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (21), in which the proof of Gao et al. (2018) is no longer applicable due to different angle gap assumption in our dTBM. We develop the extra padding technique in Lemma 5 and balance assumption (7) to derive (21). Last, we finish the proof of Theorem 4 by showing the third inequality of (20) using Han et al. (2020, Proposition 1).

### B. Proof sketch of Theorem 5

The proof of Theorem 5 is inspired by the proof idea of Han et al. (2020, Theorem 2). We develop extra polar-coordinate based techniques with angle gap characterization to address the nuisance degree heterogeneity. Recall the intermediate quantity, misclustering loss, defined in (11)

$$L^{(t)} := L(z, z^{(t)}) = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ z^{(t)}(i) = b \right\} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2,$$

where the superscript  $\cdot^s$  denotes the normalized vector; i.e.,  $\mathbf{a}^s := \mathbf{a}/\|\mathbf{a}\|$  if  $\mathbf{a} \neq 0$  and  $\mathbf{a}^s = 0$  if  $\mathbf{a} = 0$  for any vector  $\mathbf{a}$ . We show that  $L^{(t)}$  provides an upper bound for the misclassification error of interest via the inequality  $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2}$ . Therefore, it suffices to control  $L^{(t)}$ . Further, we introduce the oracle estimators for core tensor under the true cluster assignment via

$$\tilde{\mathcal{S}} = \mathcal{Y} \times_1 \mathbf{W}^T \times_2 \cdots \times_K \mathbf{W}^T,$$

where  $\mathbf{W} = \mathbf{M} (\text{diag}(\mathbf{1}_p^T \mathbf{M}))^{-1}$  is the weighted true membership matrix. Let  $\mathbf{V} = \mathbf{W}^{\otimes(K-1)}$  denote the Kronecker product of  $(K-1)$  copies of  $\mathbf{W}$  matrices, and we define the  $t$ -th iteration quantities  $\mathbf{W}^{(t)}, \mathbf{V}^{(t)}$  corresponding to  $\mathbf{M}^{(t)}$  (or equivalently  $z^{(t)}$ ). To evaluate  $L^{(t+1)}$ , we prove the bound

$$\mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} = \mathbb{1} \left\{ \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right\} \leq A_{ib} + B_{ib}, \quad (22)$$

where  $\mathbf{Y} = \text{Mat}(\mathcal{Y})$ ,  $\mathbf{S} = \text{Mat}(\mathcal{S})$ ,  $\mathbf{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$  and

$$\begin{aligned} A_{ib} &= \mathbb{1} \left\{ \left\langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \lesssim - \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \right\}, \\ B_{ib} &= \mathbb{1} \left\{ \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \lesssim F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\}. \end{aligned}$$

The terms  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$  are controlled by  $z^{(t)}, \mathcal{S}^{(t)}$ ; see the detailed definitions in (63), (64), (65). Note that the event  $A_{ib}$  only involves the oracle estimator independent of  $t$ , while all the terms related to the  $t$ -th iteration are in  $B_{ib}$ . Thus, the inequality (22) decomposes the misclustering loss in the  $(t+1)$ -th iteration into the oracle loss and the loss in  $t$ -th iteration. This decomposition leads to the separation of statistical error and computational error in the final upper bound of Theorem 5.

Specifically, we prove the contraction inequality

$$L^{(t+1)} \leq M\xi + \rho L^{(t)}, \quad \xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \quad (23)$$

where  $M$  is a positive constant,  $\rho \in (0, 1)$  is the contraction parameter, and we call  $\xi$  the oracle loss. Controlling the probability of event  $B_{ib}$  and obtaining the  $\rho L^{(t)}$  term in the right hand side of (23) are the most challenging parts in the proof of Theorem 5. Note that the true and estimated core tensors are involved via their normalized rows such as  $\mathbf{S}_{a:}^s, \tilde{\mathbf{S}}_{a:}^s, [\mathbf{S}_{a:}^{(t)}]^s$ . The Cartesian coordinate based analysis in Han et al. (2020) is no longer applicable in our case. Instead, we use the polar-coordinate based analysis and the geometry property of trigonometric functions to derive the high probability upper bounds for  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ .

Further, by sub-Gaussian concentration, we prove the high probability upper bound for oracle loss

$$\xi \lesssim \text{SNR}^{-1} \exp \left( - \frac{p^{K-1} \text{SNR}}{r^{K-1}} \right). \quad (24)$$

Combining the decomposition (23) and the oracle bound (24), we finish the proof of Theorem 5.

**The proof of MLE error shares the similar idea as Theorems 4-5.** we first show a weaker polynomial rate for MLE, and then improve the rate from polynomial to exponential throughout the iterations. The only difference is that the MLE remains the same over iterations due to its global optimality. See Appendix II, Section G for the detailed proof.

#### ACKNOWLEDGMENTS

This research is supported in part by NSF CAREER DMS-2141865, DMS-1915978, DMS-2023239, EF-2133740, and funding from the Wisconsin Alumni Research foundation. We thank Zheng Tracy Ke, Rungang Han, Yuetian Luo for helpful discussions and for sharing software packages.

#### REFERENCES

- Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

- Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.
- Ahn, K., Lee, K., and Suh, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974.
- Ahn, K., Lee, K., and Suh, C. (2019). Community recovery in hypergraphs. *IEEE Transactions on Information Theory*, 65(10):6561–6579.
- Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.
- Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.
- Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR.
- Chi, E. C., Gaines, B. J., Sun, W. W., Zhou, H., and Yang, J. (2020). Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58.
- Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.
- Chien, I. E., Lin, C.-Y., and Wang, I.-H. (2019). On the minimax misclassification ratio of hypergraph community detection. *IEEE Transactions on Information Theory*, 65(12):8095–8118.
- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.
- Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959. PMLR.
- Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.
- Gao, C. and Zhang, A. Y. (2022). Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2):1066–1094.
- Ghoshdastidar, D. and Dukkipati, A. (2017). Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *The Journal of Machine Learning Research*, 18(1):1638–1678.
- Ghoshdastidar, D. et al. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315.
- Han, R., Luo, Y., Wang, M., and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.
- Han, R., Willett, R., and Zhang, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29.
- Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094.
- Hu, J., Lee, C., and Wang, M. (2021). Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, pages 1–15.
- Ke, Z. T., Shi, F., and Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.
- Kim, C., Bandeira, A. S., and Goemans, M. X. (2018). Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*.
- Koniusz, P. and Cherian, A. (2016). Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5395–5403.
- Lee, S. H., Magallanes, J. M., and Porter, M. A. (2017). Time-dependent community structure in legislation cosponsorship networks in the congress of the republic of peru. *Journal of Complex Networks*, 5(1):127–144.
- Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd’s algorithm and its variants. *arXiv*

*preprint arXiv:1612.02099.*

- Meilă, M. (2012). Local equivalences of distances between clusterings—a geometric perspective. *Machine Learning*, 86(3):369–389.
- Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., and WU-Minn HCP Consortium (2013). The WU-Minn human connectome project: An overview. *Neuroimage*, 80:62–79.
- Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network-response regression. *Bioinformatics*, 33(12):1859–1866.
- Wang, M., Fischer, J., and Song, Y. S. (2019). Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics*, 13(2):1103–1127.
- Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723.
- Yuan, M., Liu, R., Feng, Y., and Shang, Z. (2018). Testing community structures for hypergraphs. *arXiv preprint arXiv:1810.04617*.
- Yun, S.-Y. and Proutiere, A. (2016). Optimal cluster recovery in the labeled stochastic block model. *Advances in Neural Information Processing Systems*, 29:965–973.
- Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338.
- Zhang, Z., Allen, G. I., Zhu, H., and Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *Neuroimage*, 197:330–343.

## APPENDIX

### APPENDIX I. ADDITIONAL NUMERICAL EXPERIMENTS

**Bernoulli phase transition.** The first additional experiment verifies the statistical-computational gap in Section III under the Bernoulli model. Consider the Bernoulli model with  $p = \{80, 100\}$ ,  $r = 5$ . We vary  $\gamma$  in  $[-1.2, -0.4]$  and  $[-2.1, -1.4]$  for matrix ( $K = 2$ ) and tensor ( $K = 3$ ) clustering, respectively. We approximate MLE using an oracle estimator, i.e., the output of Sub-algorithm 2 initialized from the true assignment. Figure 12 shows a similar pattern as Figure 4. The algorithm and oracle estimators have no gap in the matrix case, while an error gap emerges between the critical values  $\gamma_{\text{stat}} = -2$  and  $\gamma_{\text{comp}} = -1.5$  in the tensor case. Figure 4 suggests the statistical-computational gap in Bernoulli models.

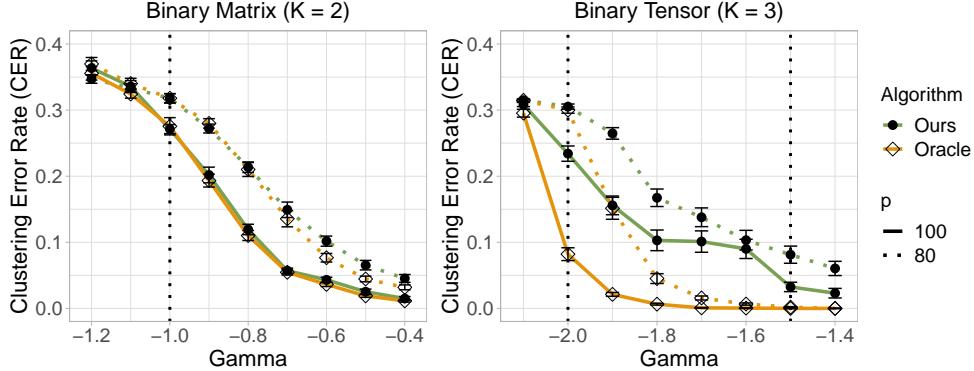


Fig. 12: SNR phase transitions for Bernoulli dTBM with  $p = \{80, 100\}, r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

**Sparsity.** The second additional experiment evaluates the algorithm performances under the sparse binary dTBM (18). We fix the signal exponent  $\gamma = -1.2$  and vary the sparsity parameter  $\alpha_p \in [0.05, 0.9]$ . A smaller  $\alpha_p$  leads to a higher probability of zero entries in the observation. In addition to the three algorithms mentioned in Section VI-B (denoted **Initialization**, **dTBM**, and **SCORE**), we consider other three algorithms based on the discussion in Section IV-C:

- **D-HOSVD**, the diagonal-deleted HOSVD in Ke et al. (2019);
- **D-HOSVD + Angle**, the combined algorithm of our angle-based iteration with initialization from **D-HOSVD**;
- **SCORE + Angle**, the combined algorithms of our angle-based iteration with initialization from **SCORE**.

Figure 13 shows a slightly larger error in **dTBM** than that in **SCORE**, **D-HOSVD + Angle**, and **SCORE + Angle** under the sparse setting with  $\alpha_p < 0.3$ . The small gap between **dTBM** and other sparse-specific methods implies the robustness of our algorithm. In addition, comparing **SCORE** versus **SCORE + Angle** (or **D-HOSVD** versus **D-HOSVD + Angle**) indicates the benefit of our angle iterations under the sparse dTBM. In the intermediate and dense cases with  $\alpha_p \geq 0.3$ , our proposed **dTBM** has a clear improvement over others, which again verifies the success of our algorithm in dense settings.

### APPENDIX II. PROOFS

We provide the proofs for all the theorems in our main paper. In each sub-section, we first show the proof of main theorem and then collect the useful lemmas in the end. We combine the proofs of MLE achievement in Theorem 2 and polynomial-time achievement in Theorem 5 in the last section due to the similar idea.

#### A. Notation

Before the proofs, we first introduce the notation used throughout the appendix and the general dTBM without symmetric assumptions. The parameter space and minimal gap assumption are also extended for the general asymmetric dTBM.

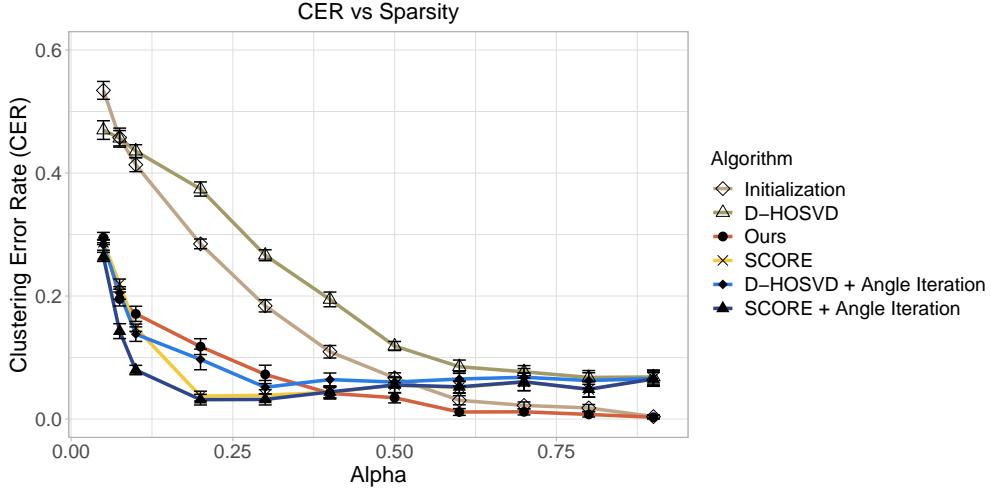


Fig. 13: CER comparison versus sparsity parameter  $\alpha_p$  in  $[0.05, 0.9]$ . We set  $p = 100, r = 5$  and  $\gamma = -1.2$  under sparse binary dTBM.

## Preliminaries.

- 1) For mode  $k \in [K]$ , denote the mode- $k$  tensor matricizations by

$$\mathbf{Y}_k = \text{Mat}_k(\mathcal{Y}), \quad \mathbf{S}_k = \text{Mat}_k(\mathcal{S}), \quad \mathbf{E}_k = \text{Mat}_k(\mathcal{E}), \quad \mathbf{X}_k = \text{Mat}_k(\mathcal{X}).$$

- 2) For a vector  $a$ , let  $a^s := a / \|a\|$  denote the normalized vector. We make the convention that  $a^s = \mathbf{0}$  if  $a = \mathbf{0}$ .
- 3) For a matrix  $A \in \mathbb{R}^{n \times m}$ , let  $A^{\otimes K} := A \otimes \cdots \otimes A \in \mathbb{R}^{n^K \times m^K}$  denote the Kronecker product of  $K$  copies of matrices  $A$ .
- 4) For a matrix  $A$ , let  $\|A\|_\sigma$  denote the spectral norm of matrix  $A$ , which is equal to the maximal singular value of  $A$ ; let  $\lambda_k(A)$  denote the  $k$ -th largest singular value of  $A$ ; let  $\|A\|_F$  denote the Frobenius norm of matrix  $A$ .
- 5) For two sequence  $a$  and  $b$ , let  $a \asymp b$  if there exist two positive constants  $c, C$  such that  $cb \leq a \leq Cb$ .

## Extension to general asymmetric dTBM.

The general order- $K$  ( $p_1, \dots, p_K$ )-dimensional dTBM with  $r_k$  communities and degree heterogeneity  $\boldsymbol{\theta}_k = [\theta_k(i)] \in \mathbb{R}_+^{p_k}$  is represented by

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where } \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 M_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K M_K, \quad (25)$$

where  $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$  is the data tensor,  $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$  is the mean tensor,  $\mathcal{S} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$  is the core tensor,  $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$  is the noise tensor consisting of independent zero-mean sub-Gaussian entries with variance bounded by  $\sigma^2$ ,  $\boldsymbol{\Theta}_k = \text{diag}(\boldsymbol{\theta}_k)$ , and  $M_k \in \{0, 1\}^{p_k \times r_k}$  is the membership matrix corresponding to the assignment  $z_k : [p_k] \mapsto [r_k]$ , for all  $k \in [K]$ .

For ease of notation, we use  $\{z_k\}$  to denote the collection  $\{z_k\}_{k=1}^K$ , and  $\{\boldsymbol{\theta}_k\}$  to denote the collection  $\{\boldsymbol{\theta}_k\}_{k=1}^K$ . Correspondingly, we consider the parameter space for the triplet  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ ,

$$\mathcal{P}(\{r_k\}) = \left\{ (\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) : \right.$$

$$\left. \begin{aligned} \boldsymbol{\theta}_k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} |z_k^{-1}(a)| \leq \frac{c_2 p_k}{r_k}, c_3 \leq \|S_{k,a:}\| \leq c_4, \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|_1 = |z_k^{-1}(a)|, a \in [r_k], k \in [K] \end{aligned} \right\} \quad (26)$$

We call the degree heterogeneity  $\{\boldsymbol{\theta}_k\}$  is balanced if for all  $k \in [K]$ ,

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|.$$

We also consider the generalized Assumption 1 on angle gap.

**Assumption 2** (Generalized angle gap). Recall  $S_k = \text{Mat}_k(\mathcal{S})$ . We assume the minimal gap between normalized rows of  $S_k$  is bounded away from zero for all  $k \in [K]$ ; i.e.,

$$\Delta_{\min} := \min_{k \in [K]} \min_{a \neq b \in [r_k]} \|S_{k,a:}^s - S_{k,b:}^s\| > 0.$$

Similarly, let  $\text{SNR} = \Delta_{\min}^2 / \sigma^2$  with the generalized minimal gap  $\Delta_{\min}^2$  defined in Assumption 2. We define the regime

$$\mathcal{P}(\gamma) = \mathcal{P}(\{r_k\}) \cap \{\mathcal{S} \text{ satisfies } \text{SNR} = p^\gamma \text{ and } p_k \asymp p, \text{ for all } k \in [K]\}.$$

### B. Proof of Theorem 1

*Proof of Theorem 1.* To study the identifiability, we consider the noiseless model with  $\mathcal{E} = 0$ . Assume there exist two parameterizations satisfying

$$\mathcal{X} = \mathcal{S} \times_1 \Theta_1 \mathbf{M}_1 \times_2 \cdots \times_K \Theta_K \mathbf{M}'_K = \mathcal{S}' \times_1 \Theta'_1 \mathbf{M}'_1 \times_2 \cdots \times_K \Theta'_K \mathbf{M}'_K, \quad (27)$$

where  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$  and  $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\}) \in \mathcal{P}(\{r'_k\})$  are two sets of parameters. We prove the sufficient and necessary conditions separately.

( $\Leftarrow$ ) For the necessity, it suffices to construct two distinct parameters up to cluster label permutation, if the model (25) violates Assumption 2. Without loss of generality, we assume  $\|S_{1,1:}^s - S_{1,2:}^s\| = 0$ .

By constraints in parameter space (26), neither  $\mathbf{S}_{1,1:}$  nor  $\mathbf{S}_{1,2:}$  is a zero vector. There exists a positive constant  $c$  such that  $\mathbf{S}_{1,1:} = c \mathbf{S}_{1,2:}$ . Thus, there exists a core tensor  $\mathcal{S}_0 \in \mathbb{R}^{r_1-1 \times \cdots \times r_K}$  such that

$$\mathcal{S} = \mathcal{S}_0 \times_1 \mathbf{C} \mathbf{R}, \quad \text{where } \mathbf{C} = \text{diag}(1, c, 1, \dots, 1) \in \mathbb{R}^{r_1 \times r_1}, \quad \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{1}_{r_1-2} \end{pmatrix} \in \mathbb{R}^{r_1 \times (r_1-1)}.$$

Let  $\mathbf{D} = \text{diag}(1 + c, 1, \dots, 1) \in \mathbb{R}^{r_1-1 \times r_1-1}$ . Consider the parameterization

$$\mathbf{M}'_1 = \mathbf{M}_1 \mathbf{R}, \quad \mathcal{S}' = \mathcal{S}_0 \times_1 \mathbf{D}, \quad \theta'_1(i) = \begin{cases} \frac{1}{1+c} \theta_1(i) & i \in z_1^{-1}(1), \\ \frac{c}{1+c} \theta_1(i) & i \in z_1^{-1}(2), \\ \theta_1(i) & \text{otherwise,} \end{cases}$$

and  $\mathbf{M}'_k = \mathbf{M}_k, \boldsymbol{\theta}'_k = \boldsymbol{\theta}_k$  for all  $k = 2, \dots, K$ . Then we have constructed a triplet  $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$  that is distinct from  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$  up to label permutation.

( $\Rightarrow$ ) For the sufficiency, it suffices to show that all possible triplets  $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$  are identical to  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$  up to label permutation if the model (25) satisfies Assumption (2). We show the uniqueness of the three parameters,  $\{\mathbf{M}_k\}, \{\mathcal{S}\}, \{\boldsymbol{\theta}_k\}$  separately.

First, we show the uniqueness of  $\mathbf{M}_k$  for all  $k \in [K]$ . Without loss of generality, we consider  $k = 1$  and show the first mode membership matrix; i.e.,  $\mathbf{M}'_1 = \mathbf{M}_1 \mathbf{P}_1$  where  $\mathbf{P}_1$  is a permutation matrix. The conclusion for  $k \geq 2$  can be showed similarly and thus omitted.

Consider an arbitrary node pair  $(i, j)$ . If  $z_1(i) = z_1(j)$ , then we have  $\|\mathbf{X}_{1,z_1(i):}^s - \mathbf{X}_{1,z_1(j):}^s\| = 0$  and thus  $\|(\mathcal{S}')_{1,z'_1(i):}^s - (\mathcal{S}')_{1,z'_1(j):}^s\| = 0$  by Lemma 3. Then, by Assumption (2), we have  $z'_1(i) = z'_1(j)$ . Conversely, if

$z_1(i) \neq z_1(j)$ , then we have  $\|\mathbf{X}_{1,i:}^s - \mathbf{X}_{1,j:}^s\| \neq 0$  and thus  $\|(S')_{1,z'_1(i):}^s - (S')_{1,z'_1(j):}^s\| \neq 0$  by Lemma 3. Hence, we have  $z'_1(i) \neq z'_1(j)$ . Therefore, we have proven that  $z'_1$  is identical  $z_1$  up to label permutation.

Next, we show the uniqueness of  $\boldsymbol{\theta}_k$  for all  $k \in [K]$  provided that  $z_k = z'_k$ . Similarly, consider  $k = 1$  only, and omit the procedure for  $k \geq 2$ .

Consider an arbitrary  $j \in [p_1]$  such that  $z_1(j) = a$ . Then for all the nodes  $i \in z_1^{-1}(a)$  in the same cluster of  $j$ , we have

$$\frac{\mathbf{X}_{1,z_1(i):}}{\mathbf{X}_{1,z_1(j):}} = \frac{\mathbf{X}'_{1,z_1(i):}}{\mathbf{X}'_{1,z_1(j):}}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta'_1(j)}{\theta'_1(i)}. \quad (28)$$

Let  $\theta'_1(j) = c\theta_1(j)$  for some positive constant  $c$ . By equation (28), we have  $\theta'_1(i) = c\theta_1(i)$  for all  $i \in z_1^{-1}(a)$ . By the constraint  $(\{z_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\}) \in \mathcal{P}(\{r_k\})$ , we have

$$\sum_{j \in z_1^{-1}(a)} \theta'_1(j) = c \sum_{j \in z_1^{-1}(a)} \theta_1(j) = 1,$$

which implies  $c = 1$ . Hence, we have proven  $\boldsymbol{\theta}_1 = \boldsymbol{\theta}'_1$  provided that  $z_1 = z'_1$ .

Last, we show the uniqueness of  $\mathcal{S}$ ; i.e.,  $\mathcal{S}' = \mathcal{S} \times_1 \mathbf{P}_1^{-1} \times_2 \cdots \times_K \mathbf{P}_K^{-1}$ , where  $\mathbf{P}_k$ 's are permutation matrices for all  $k \in [K]$ . Provided  $z'_k = z_k$ ,  $\boldsymbol{\theta}'_k = \boldsymbol{\theta}_k$ , we have  $\mathbf{M}'_k = \mathbf{M}_k \mathbf{P}_k$  and  $\boldsymbol{\Theta}'_k = \boldsymbol{\Theta}_k$  for all  $k \in [K]$ .

Let  $\mathbf{D}_k = [(\boldsymbol{\Theta}'_k \mathbf{M}'_k)^T (\boldsymbol{\Theta}'_k \mathbf{M}'_k)]^{-1} (\boldsymbol{\Theta}'_k \mathbf{M}'_k)^T$ ,  $k \in [K]$ . By the parameterization (27), we have

$$\begin{aligned} \mathcal{S}' &= \mathcal{X} \times_1 \mathbf{D}_1 \times_2 \cdots \times_K \mathbf{D}_K \\ &= \mathcal{S} \times_1 \mathbf{D}_1 \boldsymbol{\Theta}_1 \mathbf{M}_1 \times_2 \cdots \times_K \mathbf{D}_K \boldsymbol{\Theta}_K \mathbf{M}_K \\ &= \mathcal{S} \times_1 \mathbf{P}_1^{-1} \times_2 \cdots \times_K \mathbf{P}_K^{-1}. \end{aligned}$$

Therefore, we finish the proof of Theorem 1.  $\square$

### Useful Lemma for the Proof of Theorem 1

**Lemma 3** (Motivation of angle-based clustering). Consider the signal tensor  $\mathcal{X}$  in the general asymmetric dTBM (25) with  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$  and  $r_k \geq 2$ . Then, for any  $k \in [K]$  and index pair  $(i, j) \in [p_k]^2$ , we have

$$\left\| \mathbf{S}_{k,z_k(i):}^s - \mathbf{S}_{k,z_k(j):}^s \right\| = 0 \quad \text{if and only if} \quad \left\| \mathbf{X}_{k,z_k(i):}^s - \mathbf{X}_{k,z_k(j):}^s \right\| = 0.$$

*Proof of Lemma 3.* Without loss of generality, we prove  $k = 1$  only and drop the subscript  $k$  in  $\mathbf{X}_k, \mathbf{S}_k$  for notational convenience. By tensor matricization, we have

$$\mathbf{X}_{j:} = \theta_1(j) \mathbf{S}_{z_1(j):} [\boldsymbol{\Theta}_2 \mathbf{M}_2 \otimes \cdots \otimes \boldsymbol{\Theta}_K \mathbf{M}_K]^T.$$

Let  $\tilde{\mathbf{M}} = \boldsymbol{\Theta}_2 \mathbf{M}_2 \otimes \cdots \otimes \boldsymbol{\Theta}_K \mathbf{M}_K$ . Notice that for two vectors  $\mathbf{a}, \mathbf{b}$  and two positive constants  $c_1, c_2 > 0$ , we have

$$\|\mathbf{a}^s - \mathbf{b}^s\| = \|(c_1 \mathbf{a})^s - (c_2 \mathbf{b})^s\|.$$

Thus it suffices to show the following statement holds for any index pair  $(i, j) \in [p_1]^2$ ,

$$\left\| \mathbf{S}_{z_1(i):}^s - \mathbf{S}_{z_1(j):}^s \right\| = 0 \quad \text{if and only if} \quad \left\| [\mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T]^s - [\mathbf{S}_{z_1(j):} \tilde{\mathbf{M}}^T]^s \right\| = 0.$$

$(\Leftarrow)$  Suppose  $\left\| [\mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T]^s - [\mathbf{S}_{z_1(j):} \tilde{\mathbf{M}}^T]^s \right\| = 0$ . There exists a positive constant  $c$  such that  $\mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T = c \mathbf{S}_{z_1(j):} \tilde{\mathbf{M}}^T$ . Note that

$$\mathbf{S}_{z_1(i):} = \mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T \left[ \tilde{\mathbf{M}} (\tilde{\mathbf{M}}^T \tilde{\mathbf{M}})^{-1} \right],$$

where  $\tilde{\mathbf{M}}^T \tilde{\mathbf{M}}$  is an invertible diagonal matrix with positive diagonal elements. Thus, we have  $\mathbf{S}_{z_1(i)\cdot} = c\mathbf{S}_{z_1(j)\cdot}$ , which implies  $\|\mathbf{S}_{z_1(i)\cdot}^s - \mathbf{S}_{z_1(j)\cdot}^s\| = 0$ .

$(\Rightarrow)$  Suppose  $\|\mathbf{S}_{z_1(i)\cdot}^s - \mathbf{S}_{z_1(j)\cdot}^s\| = 0$ . There exists a positive constant  $c$  such that  $\mathbf{S}_{z_1(i)\cdot} = c\mathbf{S}_{z_1(j)\cdot}$ , and thus  $\mathbf{S}_{z_1(i)\cdot} \tilde{\mathbf{M}}^T = c\mathbf{S}_{z_1(j)\cdot} \tilde{\mathbf{M}}^T$ , which implies  $\left\| \left[ \mathbf{S}_{z_1(i)\cdot} \tilde{\mathbf{M}}^T \right]^s - \left[ \mathbf{S}_{z_1(j)\cdot} \tilde{\mathbf{M}}^T \right]^s \right\| = 0$ .

Therefore, we finish the proof of Lemma 3.  $\square$

### C. Proof of Lemma 1 and Lemma 2

*Proof of Lemma 1.* Note that the vector  $\mathbf{S}_{z(i)\cdot}$  can be folded to a tensor  $\mathcal{S}' = [\mathcal{S}'_{a_2, \dots, a_K}] \in \mathbb{R}^{r^{K-1}}$ ; i.e.,  $\text{vec}(\mathcal{S}') = \mathbf{S}_{z(i)\cdot}$ . Define weight vectors  $\mathbf{w}_{a_2, \dots, a_K}$  correspond to the elements in  $\mathcal{S}'_{a_2, \dots, a_K}$  by

$$\mathbf{w}_{a_2 \dots a_K} = [\boldsymbol{\theta}_{z^{-1}(a_2)}^T \otimes \dots \otimes \boldsymbol{\theta}_{z^{-1}(a_K)}^T] \in \mathbb{R}^{|z^{-1}(a_2)| \times \dots \times |z^{-1}(a_K)|},$$

for all  $a_k \in [r]$ ,  $k = 2, \dots, K$ , where  $\otimes$  denotes the Kronecker product. Therefore, we have  $\mathbf{X}_{i\cdot} = \theta(i) \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(i)\cdot})$  where  $\mathbf{w} = \{\mathbf{w}_{a_2, \dots, a_K}\}_{a_k \in [r], k \in [K]/\{1\}}$ . Specifically, we have  $\|\mathbf{w}_{a_2, \dots, a_K}\|^2 = \prod_{k=2}^K \|\boldsymbol{\theta}_{z^{-1}(a_k)}\|^2$ , and by the balanced assumption (7) we have

$$\max_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2 = (1 + o(1)) \min_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2. \quad (29)$$

Consider the inner product of  $\mathbf{X}_{i\cdot}$  and  $\mathbf{X}_{j\cdot}$  for  $z(i) \neq z(j)$ . By the definition of weighted padding operator (53) and the balanced assumption (29), we have

$$\begin{aligned} \langle \mathbf{X}_{i\cdot}, \mathbf{X}_{j\cdot} \rangle &= \theta(i)\theta(j) \langle \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(i)\cdot}), \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(j)\cdot}) \rangle \\ &= \theta(i)\theta(j) \min_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2 \langle \mathbf{S}_{z(i)\cdot}, \mathbf{S}_{z(j)\cdot} \rangle (1 + o(1)). \end{aligned}$$

Therefore, when  $p$  large enough, the inner product  $\langle \mathbf{X}_{i\cdot}, \mathbf{X}_{j\cdot} \rangle$  has the same sign as  $\langle \mathbf{S}_{z(i)\cdot}, \mathbf{S}_{z(j)\cdot} \rangle$ .

Notice that  $\|\mathbf{v}_1^s - \mathbf{v}_2^s\|^2 = 2(1 - \cos(\mathbf{v}_1, \mathbf{v}_2))$  for two vectors  $\mathbf{v}_1, \mathbf{v}_2$  of the same dimension. We have

$$\begin{aligned} \|\mathbf{S}_{z_1(i)\cdot}^s - \mathbf{S}_{z_1(j)\cdot}^s\| &= 2 - 2 \frac{\langle \mathbf{S}_{z_1(i)\cdot}, \mathbf{S}_{z_1(j)\cdot} \rangle}{\|\mathbf{S}_{z_1(i)\cdot}\| \|\mathbf{S}_{z_1(j)\cdot}\|} \\ &= 2 - 2(1 + o(1)) \frac{\langle \mathbf{X}_{i\cdot}, \mathbf{X}_{j\cdot} \rangle}{\|\mathbf{X}_{i\cdot}\| \|\mathbf{X}_{j\cdot}\|} \\ &= (1 + o(1)) \|\mathbf{X}_{i\cdot}^s - \mathbf{X}_{j\cdot}^s\|, \end{aligned}$$

where the second inequality follows by the balance assumption on  $\theta$ , and the last inequality holds when  $p$  is large enough.

Hence, we conclude that for all  $i, j$  such that  $z(i) \neq z(j)$ ,

$$\|\mathbf{X}_{i\cdot}^s - \mathbf{X}_{j\cdot}^s\| \asymp \|\mathbf{S}_{z_1(i)\cdot}^s - \mathbf{S}_{z_1(j)\cdot}^s\| \gtrsim \Delta_{\min}.$$

$\square$

*Proof of Lemma 2.* By the definition of minimal gap in Assumption 1, we have

$$L^{(t)} = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ z^{(t)}(i) = b \right\} \|[\mathbf{S}_{z(i)\cdot}]^s - [\mathbf{S}_{b\cdot}]^s\|^2 \geq \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ z^{(t)}(i) = b \right\} \Delta_{\min}^2 \geq c\ell^{(t)} \Delta_{\min}^2,$$

where the last inequality follows from the assumption  $\min_{i \in [p]} \theta(i) \geq c > 0$ .  $\square$

#### D. Proof of Theorem 2 (Impossibility)

*Proof of Theorem 2 (Impossibility).* We will prove a more general conclusion than the main paper by allowing growing  $r_k$ 's. Consider the general asymmetric dTBM (25) in the special case that  $p_k = p$  and  $r_k = r$  for all  $k \in [K]$ . Specifically, we will show that, under the assumptions  $K \geq 1$ ,  $r \lesssim p^{1/3}$  and SNR condition

$$\frac{\Delta_{\min}^2}{\sigma^2} \lesssim \frac{r^{K-1}}{p^{K-1}}, \quad \text{or equivalently, } \gamma \leq -(K-1)(1 + \log_p r),$$

the desired conclusion in Theorem 2 holds; i.e., for all  $k \in [K]$ , every estimator  $\hat{z}_{k,\text{stat}}$  obeys

$$\sup_{(\{z_k\}, \mathcal{S}, \{\theta_k\}) \in \mathcal{P}(\gamma)} \mathbb{E} [\rho \ell(\hat{z}_{k,\text{stat}}, z_k)] \geq 1. \quad (30)$$

Since the inequality (30) is a minimax lower bound, it suffices to show the inequality holds for a particular  $(\{z_k\}, \mathcal{S}, \{\theta_k\}) \in \mathcal{P}(\gamma)$ .

Specifically, we consider the estimation problem based on a particular parameter point  $(\{z_k\}, \mathcal{S}, \{\theta_k\})$  with the following two properties:

$$(i) \min_{a \in [r]} \|\theta_{z^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\theta_{z^{-1}(a)}\|; \quad (ii) \Delta_{\min} \lesssim \left(\frac{p}{r}\right)^{-\frac{K-1}{2}} \sigma; \quad (iii) |z_k^{-1}(a)| = \frac{p}{r} \in \mathbb{Z}_+ \text{ for all } a \in [r], \quad (31)$$

for all  $k \in [K]$ .

Furthermore, we define a subset of indices  $T_k \subset [p_k]$ ,  $k \in [K]$  in order to avoid the complication of label permutation. Based on Han et al. (2020, Proof of Theorem 6), we consider the minimax rate over the restricted family of  $\hat{z}_k$ 's for which the following three conditions are satisfied:

$$(iv) \hat{z}_k(i) = z_k(i) \text{ for all } i \in T_k; \quad (v) |T_k^c| \asymp \frac{p}{r}; \quad (vi) \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{\hat{z}_k(i) \neq \pi \circ z_k(i)\} = \sum_{i \in [p]} \mathbb{1}\{\hat{z}_k(i) \neq z_k(i)\},$$

for all  $k \in [K]$ .

The construction of  $T$  is precisely the same as Gao et al. (2018, Proof of Theorem 2). By the constraint  $\|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r]$  in parameter space (3), the  $\max_{i \in T^c} \theta(i) \leq C$  for some positive constant  $C$ . Then, following the proof of Gao et al. (2018, Theorem 2), for all  $k \in [K]$ , we have

$$\inf_{\hat{z}_k} \sup_{z_k} \mathbb{E} \ell(\hat{z}_k, z_k) \gtrsim \frac{1}{r^3 |T_k^c|} \sum_{i \in T_k^c} \inf_{\hat{z}_k} \{\mathbb{P}[\hat{z}_k(i) = 2 | z_k(i) = 1] + \mathbb{P}[\hat{z}_k(i) = 1 | z_k(i) = 2]\}, \quad (32)$$

where  $\hat{z}_k$  and  $z_k$  on the left hand side denote the generic clustering functions in  $\mathcal{P}(\gamma)$ ,  $z_k$  on the right hand side denotes a particular parameter satisfying properties (i)-(vi), and the infimum on the right hand side is taken over the restricted family of  $\hat{z}$  satisfying (iv)-(vi). Here, the factor  $r^3 = r \cdot r^2$  in (32) comes from two sources:  $r^2 \asymp \binom{r}{2}$  comes from the multiple testing burden for all pairwise comparisons among  $r$  clusters; and another  $r$  comes from the number of elements  $|T_k^c| \asymp p/r$  to be clustered.

Next, we need to find the lower bound of the rightmost side in (32). For simplicity, we show the bound for the mode-1 case  $k = 1$  only. We drop the subscripts 1 in  $z_1, T_1, S_1, \theta_1$  and omit the repeated procedures for the cases of  $k = 2, \dots, K$ .

We consider the hypothesis test based on model (25). First, we reparameterize the model under the construction (31)

$$\mathbf{x}_a = [\text{Mat}_1 (\mathcal{S} \times_2 \mathbf{M}_2 \times_3 \cdots \times_K \mathbf{M}_K)]_{a:}, \quad \text{for all } a \in [r],$$

where  $\mathbf{x}_a$ 's are centroids in  $\mathbb{R}^{p^{K-1}}$ . Without loss of generality, we consider the lower bound for the summand in (32) for  $i = 1$ . The analysis for other  $i \in T^c$  are similar. For notational simplicity, we suppress the subscript  $i$  and write  $\mathbf{y}, \theta, z$  in place of  $\mathbf{y}_1, \theta_1$  and  $z(1)$ , respectively. The equivalent vector problem for assessing the summand in (32) is

$$\mathbf{y} = \theta \mathbf{x}_z + \mathbf{e}, \quad (33)$$

where  $\theta \in \mathbb{R}_+$  and  $z \in \{1, 2\}$  are unknown parameters,  $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{p^{K-1}}$  are given centroids, and  $\mathbf{e} \in \mathbb{R}^{p^{K-1}}$  consists of i.i.d.  $N(0, \sigma^2)$  entries. Then, we consider the hypothesis testing under the model (33):

$$H_0: z = 1, \quad \text{v.s.} \quad H_\alpha: z = 2.$$

Note that the profile log-likelihood with respect to  $z$  is

$$\mathcal{L}(z, \theta(z); \mathbf{y}) \propto -\inf_{\theta > 0} \|\mathbf{y} - \theta \mathbf{x}_z\|^2 \propto \cos^2(\mathbf{y}, \mathbf{x}_z) \mathbb{1}\{\langle \mathbf{y}, \mathbf{x}_z \rangle > 0\},$$

and the MLE's of  $\theta$  and  $z$  are

$$\hat{\theta}_{\text{MLE}} = \hat{\theta}(\hat{z}_{\text{MLE}}) = \frac{\langle \mathbf{y}, \mathbf{x}_{\hat{z}_{\text{MLE}}} \rangle}{\|\mathbf{x}_{\hat{z}_{\text{MLE}}}\|^2} \vee 0, \quad \hat{z}_{\text{MLE}} = \arg \max_{a \in \{1, 2\}} \{\cos(\mathbf{y}, \mathbf{x}_a) \vee 0\}.$$

Then, the decision rule  $\hat{z}_{\text{MLE}} \in \{1, 2\}$  based on profile log-likelihood ratio is defined as

$$\hat{z}_{\text{MLE}} = \begin{cases} 1 & \text{if } \cos(\mathbf{y}, \mathbf{x}_1) \geq \cos(\mathbf{y}, \mathbf{x}_2) \text{ and } \langle \mathbf{y}, \mathbf{x}_1 \rangle > 0, \\ 2 & \text{if } \cos(\mathbf{y}, \mathbf{x}_1) < \cos(\mathbf{y}, \mathbf{x}_2) \text{ and } \langle \mathbf{y}, \mathbf{x}_2 \rangle > 0, \\ 1 \text{ or } 2 \text{ with equal probability} & \text{otherwise.} \end{cases} \quad (34)$$

The Neyman-Pearson Lemma implies

$$\inf_{\hat{z}} \{\mathbb{P}[\hat{z} = 2|z = 1] + \mathbb{P}[\hat{z} = 1|z = 2]\} = \mathbb{P}[\hat{z}_{\text{MLE}} = 1|z = 2] + \mathbb{P}[\hat{z}_{\text{MLE}} = 2|z = 1]. \quad (35)$$

By symmetric, it suffices to bound  $\mathbb{P}[\hat{z}_{\text{MLE}} = 1|z = 2]$ . Using (34), we obtain

$$\begin{aligned} \mathbb{P}[\hat{z}_{\text{MLE}} = 1|z = 2] &= \mathbb{P}[\cos(\theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_1) \geq \cos(\theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_2) \text{ and } \langle \theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_1 \rangle > 0] \\ &\stackrel{(*)}{\geq} \mathbb{P}\left[\left\langle \mathbf{e}, \frac{\mathbf{x}_1^s - \mathbf{x}_2^s}{\|\mathbf{x}_1^s - \mathbf{x}_2^s\|} \right\rangle \geq \frac{\theta}{2} \|\mathbf{x}_2\| \|\mathbf{x}_1^s - \mathbf{x}_2^s\| \right] - \\ &\quad \mathbb{P}\left[\left\langle \mathbf{e}, \mathbf{x}_1^s \right\rangle \leq -\frac{\theta}{2} \|\mathbf{x}_2\| (2 - \|\mathbf{x}_1^s - \mathbf{x}_2^s\|^2)\right] \\ &\stackrel{(**)}{=} \Phi\left(\frac{\theta}{2} \|\mathbf{x}_2\| (2 - \|\mathbf{x}_1^s - \mathbf{x}_2^s\|^2)\right) - \Phi\left(\frac{\theta}{2} \|\mathbf{x}_2\| \|\mathbf{x}_1^s - \mathbf{x}_2^s\|\right), \end{aligned} \quad (36)$$

where  $\Phi(\cdot)$  denotes the CDF for standard normal distribution. Here step (\*) is based on the inequality  $\mathbb{P}(A \cup B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$  and the identity  $1 - \langle \mathbf{x}_1^s, \mathbf{x}_2^s \rangle = \frac{1}{2} \|\mathbf{x}_1^s - \mathbf{x}_2^s\|^2$ ; and step (\*\*) is based on isotropic property of i.i.d. Gaussian distribution

$$\left\langle \mathbf{e}, \frac{\mathbf{x}_1^s - \mathbf{x}_2^s}{\|\mathbf{x}_1^s - \mathbf{x}_2^s\|} \right\rangle \sim N(0, \sigma^2), \quad \langle \mathbf{e}, \mathbf{x}_1^s \rangle \sim N(0, \sigma^2).$$

By construction (31) of  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$  with three properties and lower bound  $\min_{a \in [r]} \|\mathbf{S}_{a:}\| \geq c_3$  in the definition of  $\mathcal{P}(\gamma)$ , we have  $\theta^* = 1$ ,  $\|\mathbf{x}_2\| \geq \|\mathbf{S}_{2:}\| \min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^{K-1} \gtrsim \left(\frac{p}{r}\right)^{(K-1)/2}$ . Also, by the balance assumption (i), we have  $\|\mathbf{x}_1^s - \mathbf{x}_2^s\| = (1 + o(1))\|\mathbf{S}_{1:}^s - \mathbf{S}_{2:}^s\| = \Delta_{\min} \leq 1$ .

Therefore, the equation (36) is lower bounded by

$$\mathbb{P}[\hat{z}_{\text{MLE}} = 1|z = 2] \geq \mathbb{P}\left[\left(\frac{p}{r}\right)^{(K-1)/2} \Delta_{\min} \lesssim N(0, 1) \lesssim \left(\frac{p}{r}\right)^{(K-1)/2}\right] \geq C > 0, \quad (37)$$

where the existence of strictly positive constant  $C$  is based on the SNR assumption (31). Combining (32), (35) and (37) yields

$$\inf_{\hat{z}_1} \sup_{(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\gamma)} \mathbb{E}[\ell(\hat{z}_1, z_1)] \gtrsim C > 0,$$

and henceforth for all  $k \in [K]$

$$\inf_{\hat{z}_k} \sup_{(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\gamma)} \mathbb{E}[p\ell(\hat{z}_k, z_k)] \geq 1.$$

□

### E. Proof of Theorem 3 (Impossibility)

*Proof of Theorem 3 (Impossibility).* The idea of proving computational hardness is to show the computational lower bound for a special class of degree-corrected tensor clustering model with  $K \geq 2$ . We construct the following special class of higher-order degree-corrected tensor clustering model. For a given signal level  $\gamma \in \mathbb{R}$  and noise variance  $\sigma$ , define a rank-2 symmetric tensor  $\mathcal{S} \in \mathbb{R}^{3 \times \dots \times 3}$  subject to

$$\mathcal{S} = \mathcal{S}(\gamma) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}^{\otimes K}. \quad (38)$$

Then, we consider the signal tensor family

$$\mathcal{P}_{\text{shifted}}(\gamma) = \{\mathcal{X}: \mathcal{X} = \mathcal{S} \times_1 \mathbf{M}_1 \times_2 \dots \times_K \mathbf{M}_K, \mathbf{M}_k \in \{0, 1\}^{p \times 3} \text{ is a membership matrix that satisfies } |\mathbf{M}_k(:, i)| \asymp p \text{ for all } i \in [3] \text{ and } k \in [K]\}.$$

We claim that the constructed family satisfies the following two properties:

- (i) For every  $\gamma \in \mathbb{R}$ ,  $\mathcal{P}_{\text{shifted}}(\gamma) \subset \mathcal{P}(\gamma)$ , where  $\mathcal{P}(\gamma)$  is the degree-corrected cluster tensor family (6).
- (ii) For every  $\gamma \in \mathbb{R}$ ,  $\{\mathcal{X} - 1: \mathcal{X} \in \mathcal{P}_{\text{shifted}}(\gamma)\} \subset \mathcal{P}_{\text{non-degree}}(\gamma)$ , where  $\mathcal{P}_{\text{non-degree}}(\gamma)$  denotes the sub-family of rank-one tensor block model constructed in the proof of Han et al. (2020, Theorem 7).

The verification of the above two properties is provided in the end of this proof.

Now, following the proof of Han et al. (2020, Theorem 7), when  $\gamma < -K/2$ , every polynomial-time algorithm estimator  $(\hat{\mathbf{M}}_k)_{k \in [K]}$  obeys

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{non-degree}}(\gamma)} \mathbb{P}(\exists k \in [K], \hat{\mathbf{M}}_k \neq \mathbf{M}_k) \geq 1/2, \quad (39)$$

under the HPC Conjecture 1. The inequality (39) implies

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{non-degree}}(\gamma)} \max_{k \in [K]} \mathbb{E}[p\ell(z_k, \hat{z}_k)] \geq 1.$$

Based on properties (i)-(ii), we conclude that

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}(\gamma)} \max_{k \in [K]} \mathbb{E}[p\ell(z_k, \hat{z}_k)] \geq 1.$$

We complete the proof by verifying the properties (i)-(ii). For (i), we verify that the angle gap for the core tensor  $\mathcal{S}$  in (38) is on the order of  $\sigma p^{-\gamma/2}$ . Specifically, write  $\mathbf{1} = (1, 1, 1)$  and  $\mathbf{e} = (1, -1, 0)$ . We have

$$\text{Mat}(\mathcal{S}) = \begin{bmatrix} \text{Vec}(\mathbf{1}^{\otimes K-1}) + \sigma p^{-\gamma/2} \text{Vec}(\mathbf{e}^{\otimes(K-1)}) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) - \sigma p^{-\gamma/2} \text{Vec}(\mathbf{e}^{\otimes(K-1)}) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) \end{bmatrix}.$$

Based on the orthogonality  $\langle \mathbf{1}, \mathbf{e} \rangle = 0$ , the minimal angle gap among rows of  $\text{Mat}(\mathcal{S})$  is

$$\Delta_{\min}^2(\mathcal{S}) \asymp \tan^2(\text{Mat}(\mathcal{S})_{1:}, \text{Mat}(\mathcal{S})_{3:}) = \left( \frac{\|\mathbf{e}\|_2}{\|\mathbf{1}\|_2} \right)^{2(K-1)} \sigma^2 d^{-\gamma} \asymp \sigma^2 d^{-\gamma}.$$

Therefore, we have shown that  $\mathcal{P}_{\text{shifted}}(\gamma) = \mathcal{P}(\gamma)$ . Finally, the property (ii) follows directly by comparing the definition of  $\mathcal{S}$  in (38) with that in the proof of Han et al. (2020, Theorem 7). □

### F. Proof of Theorem 4

*Proof of Theorem 4.* We prove Theorem 4 under the dTBM (2) with symmetric mean tensor and parameters  $(z, \mathcal{S}, \boldsymbol{\theta})$  and i.i.d. noise. We focus on the proof of the first mode clustering  $z_1^{(0)}$ ; the proofs for the other modes can be

extended similarly. We drop the subscript  $k$  in the matricizations  $\mathbf{M}_k, \mathbf{X}_k, \mathbf{S}_k$  and in the estimate  $z_1^{(0)}$ . We firstly show the proof under the balance  $\theta$ .

**We firstly show the upper bound for misclustering error  $\ell(z^{(0)}, z)$ .**

First, by Lemma 1, there exists a positive constant such that  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\min}$ . By the balance assumption on  $\theta$  and Lemma 8, we have

$$\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_I} \theta(i)^2 + 4 \sum_{i \in S} \theta(i)^2, \quad (40)$$

where

$$S_0 = \{i : \|\hat{\mathbf{X}}_{i:}\| = 0\}, \quad S = \{i \in S_0^c : \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\| \geq c_0 \Delta_{\min}/2\}.$$

On one hand, note that for any set  $P \in [p]$ ,

$$\begin{aligned} \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 &= \sum_{i \in P} \|\theta(i) \mathbf{S}_{z(i):} (\Theta \mathbf{M})^{T, \otimes(K-1)}\|^2 \\ &\geq \sum_{i \in P} \theta(i)^2 \min_{a \in [r]} \|\mathbf{S}_{a:}\|^2 \lambda_r^{2(K-1)}(\Theta \mathbf{M}) \\ &\gtrsim \sum_{i \in P} \theta(i)^2 p^{K-1} r^{-(K-1)}, \end{aligned}$$

where the last inequality follows Lemma 6, the assumption that  $\min_{i \in [p]} \theta(i) \geq c$ , and the constraint  $\min_{a \in [r]} \|\mathbf{S}_{a:}\| \geq c_3$  in the parameter space (3). Thus, we have

$$\sum_{i \in P} \theta(i)^2 \lesssim \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 p^{-(K-1)} r^{K-1}. \quad (41)$$

On the other hand, note that

$$\sum_{i \in S} \|\mathbf{X}_{i:}\|^2 \leq 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 + 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \quad (42)$$

$$\leq \frac{8}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (43)$$

$$\leq \frac{16}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \left[ \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \hat{\mathbf{X}}_{i:}^s\|^2 + \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 \right] + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (44)$$

$$\leq \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (45)$$

$$\leq \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (46)$$

$$\lesssim \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (47)$$

where inequalities (42) and (44) follow from the triangle inequality, (43) follows from the definition of  $S$ , (45) follows from the update rule of  $k$ -means in Step 5 of Sub-algorithm 1, (46) follows from Lemma 4, and the last inequality (47) follows from Lemma 7. Also, note that

$$\sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 = \sum_{i \in S_0} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \leq \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (48)$$

where the equation follows from the definition of  $S_0$ . Therefore, combining the inequalities (40), (41), (47), and

(48), we have

$$\begin{aligned} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 &\lesssim \left( \sum_{i \in S} \|\mathbf{X}_{i:}\|^2 + \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \right) p^{-(K-1)} r^{K-1} \\ &\lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}} \left( p^{K/2} r + pr^2 + r^K \right). \end{aligned}$$

With the assumption that  $\min_{i \in [p]} \theta(i) \geq c$ , we finally obtain the result

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{r^K p^{-K/2}}{\text{SNR}},$$

where the last inequality follows from the definition  $\text{SNR} = \Delta_{\min}^2 / \sigma^2$ .

Without the balanced  $\theta$ , we have  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\mathbf{X}}$ . Replacing the definition of  $S$  with  $\Delta_{\mathbf{X}}$ , we obtain the desired result.

**Next, we show the bound for  $L(z^{(0)}, z)$ .**

Note that  $\mathbf{X}_{i:}^s$  have only  $r$  different values. We let  $\mathbf{X}_a^s = \mathbf{X}_{i:}^s$  for all  $i$  such that  $z(i) = a$ ,  $a \in [r]$ . Notice that

$$\|\mathbf{X}_{i:}\|^2 \gtrsim p^{K-1} r^{-(K-1)} \quad \text{and} \quad \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 \leq \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2} r + pr^2 + r^K.$$

Therefore, when  $p$  is large enough, we have

$$\begin{aligned} \sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 &\lesssim \sum_{i \in [p]} \left( \|\mathbf{X}_{i:}\|^2 - \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 \right) \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ &\lesssim \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ &\lesssim \eta \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 \\ &\lesssim \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ &\lesssim p^{K/2} r + pr^2 + r^K. \end{aligned} \tag{49}$$

Hence, we have

$$\begin{aligned} \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ &\lesssim \frac{r^{K-1}}{p^{K-1}} \sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ &\lesssim \frac{r^{K-1}}{p^{K-1}} \left( p^{K/2} r + pr^2 + r^K \right), \end{aligned} \tag{50}$$

where the first inequality follows from the assumption  $\min_{i \in [p]} \theta(i) \geq c$ , the second inequality follows from the inequality (41), and the last inequality comes from the inequality (49).

Next, we consider the following quantity,

$$\begin{aligned} \sum_{i \in [p]} \theta(i) \|\mathbf{X}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\mathbf{X}_{i:}^s - \hat{\mathbf{X}}_{i:}^s\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ &\lesssim \sum_{i \in [p]} \frac{\theta(i)^2}{\|\mathbf{X}_{i:}\|^2} \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \end{aligned}$$

$$\lesssim \frac{r^{K-1}}{p^{K-1}} \left( p^{K/2} r + pr^2 + r^K \right), \quad (51)$$

where the first inequality follows from the assumption of  $\theta(i)$  and triangle inequality, the second inequality follows from Lemma 4, and the last inequality follows from (50). In addition, with Theorem 4 and the condition SNR  $\gtrsim p^{-K/2} \log p$ , for all  $a \in [r]$ , we have

$$|z^{-1}(a) \cap (z^{(0)})^{-1}(a)| \geq |z^{-1}(a)| - p\ell(z^{(0)}, z) \gtrsim \frac{p}{r} - \frac{p}{\log p} \gtrsim \frac{p}{r},$$

when  $p$  is large enough. Therefore, for all  $a \in [r]$ , we have

$$\begin{aligned} \|\hat{\mathbf{x}}_a - \mathbf{X}_a^s\|^2 &= \frac{\sum_{i \in z^{-1}(a) \cap (z^{(0)})^{-1}(a)} \|\mathbf{X}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2}{|z^{-1}(a) \cap (z^{(0)})^{-1}(a)|} \\ &\lesssim \frac{r}{p} \left( \sum_{i \in [p]} \|\mathbf{X}_{i:}^s - \hat{\mathbf{X}}_{i:}^s\|^2 + \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \right) \\ &\lesssim \frac{r^K}{p^K} \left( p^{K/2} r + pr^2 + r^K \right), \end{aligned} \quad (52)$$

where the last inequality follows from the inequality (50).

Finally, we obtain

$$\begin{aligned} L^{(0)} &= \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ z^{(0)}(i) = b \right\} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \\ &\lesssim \frac{1}{p} \sum_{i \in [p], z^{(0)}(i) \neq z(i)} \theta(i) \|\mathbf{X}_{i:}^s - \mathbf{X}_{z^{(0)}(i)}^s\|^2 \\ &\lesssim \frac{1}{p} \sum_{i \in [p], z^{(0)}(i) \neq z(i)} \theta(i) \left( \|\mathbf{X}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 + \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{z^{(0)}(i)}^s\|^2 \right) \\ &\leq \bar{C} \frac{r^K}{p^K} \left( p^{K/2} r + pr^2 + r^K \right), \\ &\leq \frac{\bar{C} \Delta_{\min}^2}{\tilde{C} r \log p} \end{aligned}$$

where the first inequality follows from Lemma 1, the third inequality follows from inequalities (51) and (52), and the last inequality follows from the assumption that SNR  $\geq \tilde{C} p^{-K/2} \log p$ .

□

### Useful Definitions and Lemmas for the Proof of Theorem 4

**Lemma 4** (Basic inequality). For any two nonzero vectors  $\mathbf{v}_1, \mathbf{v}_2$  of same dimension, we have

$$\sin(\mathbf{v}_1, \mathbf{v}_2) \leq \|\mathbf{v}_1^s - \mathbf{v}_2^s\| \leq \frac{2 \|\mathbf{v}_1 - \mathbf{v}_2\|}{\max(\|\mathbf{v}_1\|, \|\mathbf{v}_2\|)}.$$

*Proof of Lemma 4.* For the first inequality, let  $\alpha \in [0, \pi]$  denote the angle between  $\mathbf{v}_1$  and  $\mathbf{v}_2$ . We have

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = \sqrt{2(1 - \cos \alpha)} = 2 \sin \frac{\alpha}{2} \geq \sin \alpha,$$

where the equations follows from the properties of trigonometric function and the inequality follows from the fact the  $\cos \frac{\alpha}{2} \leq 1$  and  $\sin \alpha = 2 \sin \frac{\alpha}{2} \cos \frac{\alpha}{2} > 0$  for  $\alpha \in [0, \pi]$ .

For the second inequality, without loss of generality, we assume  $\|\mathbf{v}_1\| \geq \|\mathbf{v}_2\|$ . Then

$$\begin{aligned}\|\mathbf{v}_1^s - \mathbf{v}_2^s\| &= \left\| \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} - \frac{\mathbf{v}_2}{\|\mathbf{v}_1\|} + \frac{\mathbf{v}_2}{\|\mathbf{v}_1\|} - \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} \right\| \\ &\leq \frac{\|\mathbf{v}_1 - \mathbf{v}_2\|}{\|\mathbf{v}_1\|} + \frac{\|\mathbf{v}_2\| \|\mathbf{v}_1\| - \|\mathbf{v}_2\|}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \\ &\leq \frac{2 \|\mathbf{v}_1 - \mathbf{v}_2\|}{\|\mathbf{v}_2\|}.\end{aligned}$$

Therefore, Lemma 4 is proved.  $\square$

**Definition 3** (Weighted padding vectors). For a vector  $\mathbf{a} = [\![a_i]\!] \in \mathbb{R}^d$ , we define the padding vector of  $\mathbf{a}$  with the weight collection  $\mathbf{w} = \{\mathbf{w}_i : \mathbf{w}_i = [\![w_{ik}]\!] \in \mathbb{R}^{p_i}\}_{i=1}^d$  as

$$\text{Pad}_{\mathbf{w}}(\mathbf{a}) = [a_1 \circ \mathbf{w}_1, \dots, a_d \circ \mathbf{w}_d]^T, \quad \text{where } a_i \circ \mathbf{w}_i = [a_i w_{i1}, \dots, a_i w_{ip_i}]^T, \text{ for all } i \in [d]. \quad (53)$$

Here we also view  $\text{Pad}_{\mathbf{w}}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^{\sum_{i \in [d]} p_i}$  as an operator. We have the bounds of the weighted padding vector

$$\min_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2 \leq \|\text{Pad}_{\mathbf{w}}(\mathbf{a})\|^2 \leq \max_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2. \quad (54)$$

Further, we define the inverse weighted padding operator  $\text{Pad}^{-1} : \mathbb{R}^{\sum_{i \in [d]} p_i} \mapsto \mathbb{R}^d$  which satisfies

$$\text{Pad}_{\mathbf{w}}^{-1}(\text{Pad}_{\mathbf{w}}(\mathbf{a})) = \mathbf{a}.$$

**Lemma 5** (Angle for weighted padding vectors). Suppose we have two non-zero vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ . Given the weight collection  $\mathbf{w}$ , we have

$$\frac{\min_{i \in [d]} \|\mathbf{w}_i\|}{\max_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}) \stackrel{*}{\leq} \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \stackrel{**}{\leq} \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}). \quad (55)$$

*Proof of Lemma 5.* We prove the two inequalities separately with similar ideas.

First, we prove the inequality \*\* in (55). Decomposing  $\mathbf{b}$  yields

$$\mathbf{b} = \cos(\mathbf{a}, \mathbf{b}) \frac{\|\mathbf{b}\|}{\|\mathbf{a}\|} \mathbf{a} + \sin(\mathbf{a}, \mathbf{b}) \frac{\|\mathbf{b}\|}{\|\mathbf{a}^\perp\|} \mathbf{a}^\perp,$$

where  $\mathbf{a}^\perp \in \mathbb{R}^d$  is in the orthogonal complement space of  $\mathbf{a}$ . By the Definition 3, we have

$$\text{Pad}_{\mathbf{w}}(\mathbf{b}) = \cos(\mathbf{a}, \mathbf{b}) \frac{\|\mathbf{b}\|}{\|\mathbf{a}\|} \text{Pad}_{\mathbf{w}}(\mathbf{a}) + \sin(\mathbf{a}, \mathbf{b}) \frac{\|\mathbf{b}\|}{\|\mathbf{a}^\perp\|} \text{Pad}_{\mathbf{w}}(\mathbf{a}^\perp).$$

Note that  $\text{Pad}_{\mathbf{w}}(\mathbf{a}^\perp)$  is not necessary equal to the orthogonal vector of  $\text{Pad}_{\mathbf{w}}(\mathbf{a})$ ; i.e.,  $\text{Pad}_{\mathbf{w}}(\mathbf{a}^\perp) \neq (\text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp$ . By the geometry property of trigonometric functions, we obtain

$$\begin{aligned}\sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) &\leq \frac{\|\mathbf{b}\| \|\text{Pad}_{\mathbf{w}}(\mathbf{a}^\perp)\|}{\|\mathbf{a}^\perp\| \|\text{Pad}_{\mathbf{w}}(\mathbf{b})\|} \sin(\mathbf{a}, \mathbf{b}) \\ &\leq \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}),\end{aligned}$$

where the second inequality follows by applying the property (54) to vectors  $\mathbf{b}$  and  $\mathbf{a}^\perp$ .

Next, we prove inequality \* in (55). With the decomposition of  $\text{Pad}_{\mathbf{w}}(\mathbf{b})$  and the inverse weighted padding operator, we have

$$\mathbf{b} = \cos(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \frac{\|\text{Pad}_{\mathbf{w}}(\mathbf{b})\|}{\|\text{Pad}_{\mathbf{w}}(\mathbf{a})\|} \mathbf{a} + \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \frac{\|\text{Pad}_{\mathbf{w}}(\mathbf{b})\|}{\|( \text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp\|} \text{Pad}_{\mathbf{w}}^{-1}((\text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp).$$

Therefore, we obtain

$$\begin{aligned}\sin(\mathbf{a}, \mathbf{b}) &\leq \frac{\|\text{Pad}_{\mathbf{w}}(\mathbf{b})\| \|\text{Pad}_{\mathbf{w}}^{-1}((\text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp)\|}{\|(\text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp\| \|\mathbf{b}\|} \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \\ &\leq \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})),\end{aligned}$$

where the second inequality follows by applying the property (54) to vectors  $\mathbf{b}$  and  $\text{Pad}_{\mathbf{w}}^{-1}((\text{Pad}_{\mathbf{w}}(\mathbf{a}))^\perp)$ .  $\square$

**Lemma 6** (Singular value of weighted membership matrix). Under the parameter space (3) and assumption that  $\min_{i \in [p]} \theta(i) \geq c$ , the singular values of  $\Theta M$  are bounded as

$$\sqrt{p/r} \lesssim \sqrt{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \leq \lambda_r(\Theta M) \leq \|\Theta M\|_\sigma \leq \sqrt{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \lesssim p/r.$$

*Proof of Lemma 6.* Note that

$$(\Theta M)^T \Theta M = D, \quad \text{with } D = \text{diag}(D_1, \dots, D_r), \quad D_a = \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2, a \in [r].$$

By the definition of singular values, we have

$$\sqrt{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \leq \lambda_r(\Theta M) \leq \|\Theta M\|_\sigma \leq \sqrt{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2}.$$

Since that  $\min_{i \in [p]} \theta(i) \geq c$  by the constraints in parameter space, we have

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \geq c^2 \min_{a \in [r]} |z^{-1}(a)| \gtrsim \frac{p}{r},$$

where the last inequality follows from the constraint in parameter space (3). Finally, notice that

$$\sqrt{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \leq \max_{a \in [r]} \sqrt{\|\boldsymbol{\theta}_{z^{-1}(a)}\|_1^2} \lesssim \frac{p}{r}.$$

Therefore, we complete the proof of Lemma 6.  $\square$

**Lemma 7** (Singular-value gap-free tensor estimation error bound). Consider an order- $K$  tensor  $\mathcal{A} = \mathcal{X} + \mathcal{Z} \in \mathbb{R}^{p \times \dots \times p}$ , where  $\mathcal{X}$  has Tucker rank  $(r, \dots, r)$  and  $\mathcal{Z}$  has independent sub-Gaussian entries with parameter  $\sigma^2$ . Let  $\hat{\mathcal{X}}$  denote the double projection estimated tensor in Step 2 of Sub-algorithm 1 in the main paper. Then with probability at least  $1 - C \exp(-cp)$ , we have

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \leq C\sigma^2 \left( p^{K/2}r + pr^2 + r^K \right),$$

where  $C, c$  are some positive constants.

*Proof of Lemma 7.* See Han et al. (2020, Proposition 1).  $\square$

**Lemma 8** (Upper bound of misclustering error). Let  $z : [p] \mapsto [r]$  be a cluster assignment such that  $|z^{-1}(a)| \asymp p/r$  for all  $a \in [r]$ . Let node  $i$  correspond to a vector  $\mathbf{x}_i = \theta(i)\mathbf{v}_{z(i)} \in \mathbb{R}^d$ , where  $\{\mathbf{v}_a\}_{a=1}^r$  are the cluster centers and  $\boldsymbol{\theta} = [\theta(i)] \in \mathbb{R}_+^p$  is the positive degree heterogeneity. Assume that  $\boldsymbol{\theta}$  satisfies the balanced assumption (7) such that  $\frac{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} = 1 + o(1)$ . Consider an arbitrary estimate  $\hat{z}$  with  $\hat{\mathbf{x}}_i = \hat{\mathbf{v}}_{\hat{z}(i)}$  for all  $i \in S$ . Then, if

$$\min_{a \neq b \in [r]} \|\mathbf{v}_a - \mathbf{v}_b\| \geq 2c, \tag{56}$$

for some constant  $c > 0$ , we have

$$\min_{\pi \in \Pi} \sum_{i: \hat{z}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_0} \theta(i)^2 + 4 \sum_{i \in S} \theta(i)^2,$$

where  $S_0$  is defined in Step 3 of Sub-algorithm 1 and

$$S = \{i \in S_0^c : \|\hat{\mathbf{x}}_i - \mathbf{v}_{z(i)}\| \geq c\}.$$

*Proof of Lemma 8.* For each cluster  $u \in [r]$ , we use  $C_u$  to collect the subset of points for which the estimated and true positions  $\hat{\mathbf{x}}_i, \mathbf{x}_i$  are within distance  $c$ . Specifically, define

$$C_u = \{i \in z^{-1}(u) \cap S_0^c : \|\hat{\mathbf{x}}_i - \mathbf{v}_{z(i)}\| < c\},$$

and divide  $[r]$  into three groups based on  $C_u$  as

$$\begin{aligned} R_1 &= \{u \in [r] : C_u = \emptyset\}, \\ R_2 &= \{u \in [r] : C_u \neq \emptyset, \text{ for all } i, j \in C_u, \hat{z}(i) = \hat{z}(j)\}, \\ R_3 &= \{u \in [r] : C_u \neq \emptyset, \text{ there exist } i, j \in C_u, \hat{z}(i) \neq \hat{z}(j)\}. \end{aligned}$$

Note that  $\cup_{u \in [r]} C_u = S_0^c / S^c$  and  $C_u \cap C_v = \emptyset$  for any  $u \neq v$ . Suppose there exist  $i \in C_u$  and  $j \in C_v$  with  $u \neq v \in [r]$  and  $\hat{z}(i) = \hat{z}(j)$ . Then we have

$$\|\mathbf{v}_{z(i)} - \mathbf{v}_{z(j)}\| \leq \|\mathbf{v}_{z(i)} - \hat{\mathbf{x}}_i\| + \|\mathbf{v}_{z(j)} - \hat{\mathbf{x}}_j\| < 2c,$$

which contradicts to the assumption (56). Hence, the estimates  $\hat{z}(i) \neq \hat{z}(j)$  for the nodes  $i \in C_u$  and  $j \in C_v$  with  $u \neq v$ . By the definition of  $R_2$ , the nodes in  $\cup_{u \in R_2} C_u$  have the same assignment with  $z$  and  $\hat{z}$ . Then, we have

$$\min_{\pi \in \Pi} \sum_{i: \hat{z}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_0} \theta(i)^2 + \sum_{i \in S} \theta(i)^2 + \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2.$$

We only need to bound  $\sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2$  to finish the proof. Note that every  $C_u$  with  $u \in R_3$  contains at least two nodes assigned to different clusters by  $\hat{z}$ . Then, we have  $|R_2| + 2|R_3| \leq r$ . Since  $|R_1| + |R_2| + |R_3| = r$ , we have  $|R_3| \leq |R_1|$ . Hence, we obtain

$$\begin{aligned} \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2 &\leq |R_3| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ &\leq |R_1| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ &\leq \frac{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \sum_{i \in \cup_{u \in R_1} z^{-1}(u)} \theta(i)^2 \\ &\leq 2 \sum_{i \in S} \theta(i)^2, \end{aligned}$$

where the last inequality holds by the balanced assumption on  $\boldsymbol{\theta}$  when  $p$  is large enough, and the fact that  $\cup_{u \in R_1} z^{-1}(u) \subset S$ .  $\square$

### G. Proofs of Theorem 2 (Achievability) and Theorem 5

*Proof of Theorem 2 (Achievability) and Theorem 5.* The proofs of Theorem 2 (Achievability) and Theorem 5 share the same idea. We prove the contraction step by step. In each step, we show the specific procedures for the algorithm loss and address the MLE loss by stating the difference.

We consider dTBM (2) with symmetric mean tensor and parameters  $(z, \mathcal{S}, \boldsymbol{\theta})$ . Let  $(\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}})$  denote the MLE in (10), and  $(z_k^{(0)}, \boldsymbol{\theta}^{(0)}, \mathcal{S}^{(0)})$  denote parameters related to the initialization. We focus on the proof of  $z_1^{(t+1)}$ ; the extension for other modes can be obtained similarly. We drop the subscript  $k$  in the matricizations  $\boldsymbol{\Theta}, \mathbf{M}_k, \mathbf{S}_k, \mathbf{X}_k$  and in estimates  $z^{(0)}, z_1^{(t+1)}, z_k^{(t)}$  for ease of the notation. Without loss of generality, we assume that the variance  $\sigma = 1$ , and that the

identity permutation minimizes the initial misclustering error; i.e.,  $\pi^{(0)} = \arg \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{z^{(0)}(i) \neq \pi \circ z(i)\}$  and  $\pi^{(0)}(a) = a$  for all  $a \in [r]$ , and so for  $\hat{z}$ .

**Step 1 (Notation and conditions).** We first introduce additional notations and the necessary conditions used in the proof. We will verify that the conditions hold in our context under high probability in the last step of the proof.

### Notation.

- 1) Projection. We use  $\mathbf{I}_d$  to denote the identity matrix of dimension  $d$ . For a vector  $\mathbf{v} \in \mathbb{R}^d$ , let  $\text{Proj}(\mathbf{v}) \in \mathbb{R}^{d \times d}$  denote the projection matrix to  $\mathbf{v}$ . Then,  $\mathbf{I}_d - \text{Proj}(\mathbf{v})$  is the projection matrix to the orthogonal complement  $\mathbf{v}^\perp$ .
- 2) We define normalized and weighted normalized membership matrices

$$\mathbf{W} = \mathbf{M} (\text{diag}(\mathbf{1}_p^T \mathbf{M}))^{-1}, \quad \mathbf{W}^{(t)} = \mathbf{M}^{(t)} (\text{diag}(\mathbf{1}_p^T \mathbf{M}^{(t)}))^{-1},$$

$$\mathbf{P} = \Theta \mathbf{M} (\text{diag}(\|\theta_{z^{-1}(1)}\|^2, \dots, \|\theta_{z^{-1}(r)}\|^2))^{-1}, \quad \hat{\mathbf{P}} = \hat{\Theta} \hat{\mathbf{M}} (\text{diag}(\|\hat{\theta}_{z^{-1}(1)}\|^2, \dots, \|\hat{\theta}_{z^{-1}(r)}\|^2))^{-1},$$

and the dual normalized and dual weighted normalized membership matrices

$$\mathbf{V} = \mathbf{W}^{\otimes(K-1)}, \quad \mathbf{V}^{(t)} = (\mathbf{W}^{(t)})^{\otimes(K-1)}, \quad \mathbf{Q} = \mathbf{P}^{\otimes K-1}, \quad \hat{\mathbf{Q}} = \hat{\mathbf{P}}^{\otimes K-1}.$$

Also, let  $\mathbf{B} = (\Theta \mathbf{M})^{\otimes(K-1)}$ ,  $\hat{\mathbf{B}} = (\hat{\Theta} \hat{\mathbf{M}})^{\otimes(K-1)}$ . By the definition, we have  $\mathbf{B}^T \mathbf{Q} = \hat{\mathbf{B}}^T \hat{\mathbf{Q}} = \mathbf{I}_{r^{K-1}}$ .

- 3) We use  $\mathcal{S}^{(t)}$  to denote the estimator of  $\mathcal{S}$  in the  $t$ -th iteration,  $\hat{\mathcal{S}}$  for MLE,  $\tilde{\mathcal{S}}$  to denote the oracle estimator of  $\mathcal{S}$  given true assignment  $z$ , and  $\bar{\mathcal{S}}$  for weighted oracle estimator; i.e.,

$$\mathcal{S}^{(t)} = \mathcal{Y} \times_1 (\mathbf{W}^{(t)})^T \times_2 \cdots \times_K (\mathbf{W}^{(t)})^T, \quad \tilde{\mathcal{S}} = \mathcal{Y} \times_1 \mathbf{W}^T \times_2 \cdots \times_K \mathbf{W}^T$$

$$\hat{\mathcal{S}} = \mathcal{Y} \times_1 \hat{\mathbf{P}}^T \times_2 \cdots \times_K \hat{\mathbf{P}}^T, \quad \bar{\mathcal{S}} = \mathcal{Y} \times_1 \mathbf{P}^T \times_2 \cdots \times_K \mathbf{P}^T.$$

- 4) We define the matricizations of tensors

$$\mathbf{S} = \text{Mat}(\mathcal{S}), \quad \mathbf{Y} = \text{Mat}(\mathcal{Y}), \quad \mathbf{X} = \text{Mat}(\mathcal{X}), \quad \mathbf{E} = \text{Mat}(\mathcal{E})$$

$$\mathbf{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)}), \quad \hat{\mathbf{S}} = \text{Mat}(\hat{\mathcal{S}}), \quad \tilde{\mathbf{S}} = \text{Mat}(\tilde{\mathcal{S}}), \quad \bar{\mathbf{S}} = \text{Mat}(\bar{\mathcal{S}}).$$

- 5) We define the extended core tensor on  $K-1$  modes

$$\mathbf{A} = \mathbf{S} \mathbf{B}^T, \quad \bar{\mathbf{A}} = \bar{\mathbf{S}} \mathbf{B}^T, \quad \hat{\mathbf{A}} = \hat{\mathbf{S}} \hat{\mathbf{B}}^T.$$

By the assumption in parameter space (3), we have  $\mathbf{A} = \mathbf{P} \mathbf{X} = \mathbf{W} \mathbf{X}$ ,  $\hat{\mathbf{A}} = \hat{\mathbf{P}} \hat{\mathbf{X}} = \hat{\mathbf{W}} \hat{\mathbf{X}}$ .

- 6) We define the angle-based misclustering loss in the  $t$ -th iteration and loss for MLE

$$L^{(t)} = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1}\{z^{(t)}(i) = b\} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2,$$

$$L(\hat{z}) = \frac{1}{p} \sum_{i \in [p]} \theta(i)^2 \sum_{b \in [r]} \mathbb{1}\{\hat{z}(i) = b\} \|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2.$$

We also define the loss for oracle and weighted oracle estimators

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1}\left\{ \langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \right\} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2,$$

$$\xi' = \frac{1}{p} \sum_{i \in [p]} \theta(i)^2 \sum_{b \in [r]} \mathbb{1}\left\{ \langle \mathbf{E}_{i:}, [\bar{\mathbf{A}}_{z(i)}]_s - [\bar{\mathbf{A}}_b]_s \rangle \leq -\frac{m'}{4} \sqrt{\frac{p^{K-1}}{r^{K-1}}} \|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|_F^2 \right\} \|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2.$$

where  $m$  and  $m'$  are some positive universal constants.

**Condition 1.** (Intermediate results) Let  $\mathbb{O}_{p,r}$  denote the collection of all the  $p$ -by- $r$  matrices with orthonormal columns. We have

$$\|\mathbf{EV}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}} \left( p^{1/2} + r^{(K-1)/2} \right), \quad \|\mathbf{EV}\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}}, \quad \|\mathbf{W}_{a:}^T \mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}} \text{ for all } a \in [r], \quad (57)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_\sigma \lesssim \left( \sqrt{r^{K-1}} + K\sqrt{pr} \right), \quad (58)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_F \lesssim \left( \sqrt{pr^{K-1}} + K\sqrt{pr} \right), \quad (59)$$

$$\xi \leq \exp \left( -M \frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad \xi' \lesssim \exp \left( -\frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad (60)$$

$$L^{(t)} \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \quad \text{for } t = 0, 1, \dots, T, \quad L(\hat{z}) \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p} \quad (61)$$

where  $M$  is a positive universal constant in inequality (80),  $\bar{C}, \tilde{C}$  are universal constants in Corollary ?? and assumption  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$ , respectively. Further, inequality (57) holds by replacing  $\mathbf{V}$  to  $\mathbf{V}^{(t)}, \mathbf{Q}, \hat{\mathbf{Q}}$  and  $\mathbf{W}_{:a}$  to  $\mathbf{W}_{:a}^{(t),T}, \mathbf{P}_{:a}^T, \hat{\mathbf{P}}_{:a}^T$  when initialization condition (61) holds.

**Step 2 (Misclustering loss decomposition).** Next, we derive the upper bound of  $L^{(t+1)}$  for  $t = 0, 1, \dots, T-1$ . By Sub-algorithm 2, we update the assignment in  $t$ -th iteration via

$$z^{(t+1)}(i) = \arg \min_{a \in [r]} \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{a:}^{(t)}]^s\|^2,$$

following the facts that  $\|\mathbf{a}^s - \mathbf{b}^s\|^2 = 1 - \cos(\mathbf{a}, \mathbf{b})$  for vectors  $\mathbf{a}, \mathbf{b}$  of same dimension and  $\text{Mat}(\mathcal{Y}^d) = \mathbf{Y} \mathbf{V}^{(t)}$  where  $\mathcal{Y}^d$  is the reduced tensor defined in Step 8 of Sub-algorithm 2. Then the event  $z^{(t+1)}(i) = b$  implies

$$\|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2. \quad (62)$$

Note that the event (62) also holds for the degenerate entity  $i$  with  $\|\mathbf{Y}_{i:} \mathbf{V}^{(t)}\| = 0$  due to the convention that  $\mathbf{a}^s = \mathbf{0}$  if  $\mathbf{a} = \mathbf{0}$ . Arranging the terms in (62) yields the decomposition

$$2 \left\langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\| \left( -\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 + G_{ib}^{(t)} + H_{ib}^{(t)} \right) + F_{ib}^{(t)},$$

where

$$F_{ib}^{(t)} = 2 \left\langle \mathbf{E}_{i:} \mathbf{V}^{(t)}, \left( [\tilde{\mathbf{S}}_{z(i):}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right) - \left( [\tilde{\mathbf{S}}_{b:}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right) \right\rangle + 2 \left\langle \mathbf{E}_{i:} \left( \mathbf{V} - \mathbf{V}^{(t)} \right), [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \quad (63)$$

$$\begin{aligned} G_{ib}^{(t)} &= \left( \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{S}_{z(i):}^{(t)}\|^2 - \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}\|^2 \right) \\ &\quad - \left( \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{S}_{b:}^{(t)}\|^2 - \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}\|^2 \right), \end{aligned} \quad (64)$$

$$H_{ib}^{(t)} = \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}\|^2 - \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^2 - \|\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}\|^2 + \|\mathbf{S}_{z(i):}^s - [\mathbf{S}_{b:}]^s\|^2. \quad (65)$$

Therefore, the event  $\mathbb{1} \{ z^{(t+1)}(i) = b \}$  can be upper bounded as

$$\begin{aligned} \mathbb{1} \{ z^{(t+1)}(i) = b \} &\leq \mathbb{1} \left\{ z^{(t+1)}(i) = b, \left\langle \mathbf{E}_{j:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq -\frac{1}{4} \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\| \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \right\} \\ &\quad + \mathbb{1} \left\{ z^{(t+1)}(i) = b, \frac{1}{2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \leq \|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|^{-1} F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\}. \end{aligned} \quad (66)$$

Note that

$$\|\mathbf{X}_{i:}\mathbf{V}^{(t)}\| = \theta(i)\|\mathbf{S}_{i:}(\Theta\mathbf{M})^{\otimes(K-1),T}\mathbf{W}^{(t),\otimes K-1}\| \geq \theta(i)\|\mathbf{S}_{z(i):}\|\lambda_r^{K-1}(\Theta\mathbf{M})\lambda_r^{K-1}(\mathbf{W}^{(t)}) \geq \theta(i)m, \quad (67)$$

where the first inequality follows from the property of eigenvalues; the last inequality follows from Lemma 6, Lemma 9, and assumption that  $\min_{a \in [r]}\|\mathbf{S}_{z(a):}\| \geq c_3 > 0$ ; and  $m > 0$  is a positive constant related to  $c_3$ . Plugging the lower bound of  $\|\mathbf{X}_{i:}\mathbf{V}^{(t)}\|$  (67) into the inequality (66) gives

$$\mathbb{1}\left\{z^{(t+1)}(i) = b\right\} \leq A_{ib} + B_{ib}, \quad (68)$$

where

$$A_{ib} = \mathbb{1}\left\{z^{(t+1)}(i) = b, \langle \mathbf{E}_{i:}\mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \rangle \leq -\frac{\theta(i)m}{4}\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2\right\},$$

$$B_{ib} = \mathbb{1}\left\{z^{(t+1)}(i) = b, \frac{1}{2}\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \leq (\theta(i)m)^{-1}F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}\right\}.$$

Taking the weighted summation of (68) over  $i \in [p]$  yields

$$L^{(t+1)} \leq \xi + \frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta_{ib}^{(t)},$$

where  $\xi$  is the oracle loss such that

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2. \quad (69)$$

Similarly to  $\xi$  in (69), we define

$$\zeta_{ib}^{(t)} = \theta(i)B_{ib}\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2.$$

### Now, we show the decomposition for MLE loss.

By the definition of Gaussian MLE, the estimator  $\hat{\theta}(i) = \langle \mathbf{Y}_{i:}, \hat{\mathbf{A}}_{\hat{z}(i):} \rangle / \|\hat{\mathbf{A}}_{\hat{z}(i)}\|_F^2$  for all  $i \in [p]$ . Hence, we have

$$\hat{z}(i) = \arg \min_{a \in [r_1]} \|[\mathbf{Y}_{i:}]^s - [\hat{\mathbf{A}}_{a:}]^s\|_F^2,$$

and the decomposition

$$L(\hat{z}) \leq \xi' + \frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta'_{ib},$$

where  $\zeta'_{ib} = \theta(i)^2 B'_{ib} \|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|^2$  and

$$A'_{ib} = \mathbb{1}\left\{\hat{z}(i) = b, \langle \mathbf{E}_{i:}, [\bar{\mathbf{A}}_{z(i):}]^s - [\bar{\mathbf{A}}_{b:}]^s \rangle \leq -\frac{m'}{4}\sqrt{\frac{p^{K-1}}{r^{K-1}}}\|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2\right\},$$

$$B'_{ib} = \mathbb{1}\left\{\hat{z}(i) = b, -\frac{1}{2}\|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2 \leq \sqrt{\frac{r^{K-1}}{(m')^2 p^{K-1}}} \hat{F}_{ib} + \hat{G}_{ib} + \hat{H}_{ib}\right\}$$

with terms

$$\hat{F}_{ib} = 2\left\langle \mathbf{E}_{i:}, ([\bar{\mathbf{A}}_{z(i):}]^s - [\hat{\mathbf{A}}_{a:}]^s) - ([\bar{\mathbf{A}}_{b:}]^s - [\hat{\mathbf{A}}_{b:}]^s) \right\rangle,$$

$$\hat{G}_{ib} = \left(\|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{z(i):}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2\right) - \left(\|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{b:}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2\right),$$

$$\hat{H}_{ib} = \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 + \|\mathbf{A}_{z(i):}^s - \mathbf{A}_{b:}^s\|_F^2.$$

**Step 3 (Derivation of contraction inequality).** In this step we derive the upper bound of  $\zeta_{ib}$  and obtain the

contraction inequality (23). Choose the constant  $\tilde{C}$  in the condition  $\text{SNR} \geq \tilde{C}p^{-K/2} \log p$  that satisfies the condition of Lemma 10, inequalities (94) and (98). Note that

$$\begin{aligned}\zeta_{ib}^{(t)} &= \theta(i)\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \mathbb{1} \left\{ z^{(t+1)}(i) = b, \frac{1}{2}\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \leq (\theta(i)m)^{-1}F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\} \\ &\leq \theta(i)\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \mathbb{1} \left\{ z^{(t+1)}(i) = b, \frac{1}{4}\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \leq (\theta(i)m)^{-1}F_{ib}^{(t)} + G_{ib}^{(t)} \right\} \\ &\leq 64\mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} \left( \frac{(F_{ib}^{(t)})^2}{cm^2\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} + \frac{\theta(i)(G_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \right)\end{aligned}$$

where the first inequality follows from the inequality (85) in Lemma 10, and the last inequality follows from the assumption that  $\min_{i \in [p]} \theta(i) \geq c > 0$ . Following Han et al. (2020, Step 4, Proof of Theorem 2) and Lemma 10, we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} \frac{(F_{ib}^{(t)})^2}{cm^2\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{C_0\bar{C}}{cm^2\bar{C}^2} L^{(t)},$$

for a positive universal constant  $C$  and

$$\begin{aligned}\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} \frac{\theta(i)(G_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} &\leq \frac{1}{512} \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} \mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} (\Delta_{\min}^2 + L^{(t)}) \\ &\leq \frac{1}{512} (L^{(t+1)} + L^{(t)}),\end{aligned}$$

where the last inequality follows from the definition of  $L^{(t)}$  and the constraint of  $\boldsymbol{\theta}$  in parameter space (3). For  $\tilde{C}$  also satisfies

$$\frac{C_0\bar{C}}{cm^2\bar{C}^2} \leq \frac{1}{512}, \quad (70)$$

we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta_{ib}^{(t)} \leq \frac{1}{8} L^{(t+1)} + \frac{1}{4} L^{(t)}. \quad (71)$$

Plugging the inequality (71) into the decomposition (69), we obtain the contraction inequality

$$L^{(t+1)} \leq \frac{3}{2}\xi + \frac{1}{2}L^{(t)}, \quad (72)$$

where  $\frac{1}{2}$  is the contraction parameter.

Therefore, with  $\tilde{C}$  satisfying inequalities (70), (94) and (98), we obtain the conclusion in Theorem 5 via inequality (73) combining the inequality (60) in Condition 1 and Lemma 2.

#### We also have the contraction inequality for MLE.

Following the same derivation of (73) with the upper bound of  $\hat{F}_{ib}, \hat{G}_{ib}, \hat{H}_{ib}$  in Lemma 11, we also have

$$L(\hat{z}) \leq \frac{3}{2}\xi' + \frac{1}{2}L(\hat{z}), \quad (73)$$

which indicates the conclusion  $\ell(\hat{z}, z) \lesssim \Delta_{\min}^2 \exp\left(-\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right)$ .

**Step 4 (Verification of Condition 1).** Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (57), (58), and (59) describe the property of the sub-Gaussian noise tensor  $\mathcal{E}$ , and the readers can find the proof directly in Han et al. (2020, Step 5, Proof of Theorem 2). The initial condition (61) for MLE is satisfied by Lemma 12. Here, we include only the verification of inequalities (60) and (61) for algorithm estimators.

Now, we verify the oracle loss condition (60). Recall the definition of  $\xi$ ,

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbf{1} \left\{ \left\langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i)}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right\} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2,$$

Let  $e_i = \mathbf{E}_{i:} \mathbf{V}$  denote the aggregated noise vector for all  $i \in [p]$ , and  $e_i$ 's are independent zero-mean sub-Gaussian vector in  $\mathbb{R}^{r^{K-1}}$ . The entries in  $e_i$  are independent zero-mean sub-Gaussian variables with sub-Gaussian norm upper bounded by  $m_1 \sqrt{r^{K-1}/p^{K-1}}$  with some positive constant  $m_1$ . We have the probability inequality

$$\mathbb{P} \left( \left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right) \leq P_1 + P_2 + P_3,$$

where

$$\begin{aligned} P_1 &= \mathbb{P} \left( \left\langle e_i, [\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s \right\rangle \leq -\frac{\theta(i)m}{8} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right), \\ P_2 &= \mathbb{P} \left( \left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s \right\rangle \leq -\frac{\theta(i)m}{16} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right), \\ P_3 &= \mathbb{P} \left( \left\langle e_i, [\mathbf{S}_{b:}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq -\frac{\theta(i)m}{16} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right). \end{aligned}$$

For  $P_1$ , notice that the inner product  $\left\langle e_j, \mathbf{S}_{z(j)}^s - \mathbf{S}_{b:}^s \right\rangle$  is a sub-Gaussian variable with sub-Gaussian norm bounded by  $m_2 \sqrt{r^{K-1}/p^{K-1}} \|\mathbf{S}_{z(i)}^s - \mathbf{S}_{b:}^s\|$  with some positive constant  $m_2$ . Then, by Chernoff bound, we have

$$P_1 \lesssim \exp \left( -\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(j)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right). \quad (74)$$

For  $P_2$  and  $P_3$ , we only need to derive the upper bound of  $P_2$  due to the symmetry. By the law of total probability, we have

$$P_2 \leq P_{21} + P_{22}, \quad (75)$$

where with some positive constant  $t > 0$ ,

$$\begin{aligned} P_{21} &= \mathbb{P} \left( t \leq \|[\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s\| \right), \\ P_{22} &= \mathbb{P} \left( \left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s \right\rangle \leq -\frac{\theta(i)m}{16} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \mid \|[\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s\| < t \right). \end{aligned}$$

For  $P_{21}$ , note that the term  $\mathbf{W}_{:z(i)}^T \mathbf{EV} = \frac{\sum_{j \neq i, j \in [p]} \mathbf{1}\{z(j)=z(i)\} e_j}{\sum_{j \in [p]} \mathbf{1}\{z(j)=z(i)\}}$  is a sub-Gaussian vector with sub-Gaussian norm bounded by  $m_3 \sqrt{r^K/p^K}$  with some positive constant  $m_3$ . This implies

$$P_{21} \leq \mathbb{P} \left( t \|\mathbf{S}_{z(i)}\| \leq \|\tilde{\mathbf{S}}_{z(i)} - \mathbf{S}_{z(i)}\| \right) \leq \mathbb{P} \left( c_3 t \leq \|\mathbf{W}_{:z(i)}^T \mathbf{EV}\| \right) \lesssim \exp \left( -\frac{p^K t^2}{r^K} \right), \quad (76)$$

where the first inequality follows from the basic inequality in Lemma 4, the second inequality follows from the assumption that  $\min_{a \in [r]} \|\mathbf{S}_{z(i)}\| \geq c_3 > 0$  in (3), and the last inequality follows from the Bernstein inequality.

For  $P_{22}$ , the inner product  $\left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s \right\rangle$  is also a sub-Gaussian variable with sub-Gaussian norm  $m_4 \sqrt{r^{K-1}/p^{K-1}} t$ , conditioned on  $\|[\tilde{\mathbf{S}}_{z(i)}]^s - [\mathbf{S}_{z(i)}]^s\| < t$  with some positive constant  $m_4$ . Then, by Chernoff bound, we have

$$P_{22} \lesssim \exp \left( -\frac{p^{K-1}}{r^{K-1} t^2} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^4 \right). \quad (77)$$

We take  $t = \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|$  in  $P_{21}$  and  $P_{22}$ , and plug the inequalities (76) and (77) into the upper bound for

$P_2$  in (75). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right). \quad (78)$$

Combining the upper bounds (74) and (78) gives

$$\mathbb{P}\left(\left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right) \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right). \quad (79)$$

Hence, we have

$$\begin{aligned} \mathbb{E}\xi &= \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{P}\left\{\left\langle \mathbf{E}_i \mathbf{V}, [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right\} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \\ &\lesssim \frac{1}{p} \sum_{i \in [p]} \theta(i) \max_{i \in [p], b \in [r]} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right) \\ &\leq \exp\left(-M \frac{p^{K-1}}{r^{K-1}} \Delta_{\min}^2\right), \end{aligned} \quad (80)$$

where  $M$  is a positive constant, the first inequality follows from the constraint that  $\sum_{i \in [p]} \theta(i) = p$ , and the last inequality follows from (79).

By Markov's inequality, we have

$$\mathbb{P}\left(\xi \lesssim \mathbb{E}\xi + \exp\left(-\frac{Mp^{K-1}}{2r^{K-1}} \Delta_{\min}^2\right)\right) \geq 1 - C \exp\left(-\frac{Mp^{K-1}}{2r^{K-1}} \Delta_{\min}^2\right),$$

and thus the condition (60) holds with probability at least  $1 - C \exp\left(-\frac{Mp^{K-1}}{2r^{K-1}} \Delta_{\min}^2\right)$  for some constant  $C > 0$ .

#### The initialization condition for MLE also holds.

For  $\xi'$ , notice that  $\langle \mathbf{E}_i, \mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s \rangle$  is a sub-Gaussian vector with variance bounded by  $\|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|^2$  and

$$\mathbb{P}(t \leq \|[\bar{\mathbf{A}}_{a:}]_s - \mathbf{A}_{a:}^s\|) \leq (t \leq \|[\mathbf{P}_{:a}^T \mathbf{YQ}]^s - [\mathbf{P}_{:a}^T \mathbf{XQ}]^s\|) \leq \mathbb{P}(t \min_{a \in [r]} \|\mathbf{S}_{a:}\| \leq \|\mathbf{P}_{:a}^T \mathbf{EQ}\|) \lesssim \exp\left(-\frac{p^K t^2}{r^K}\right),$$

where the first inequality follows from the property in later inequality (101). We also have

$$\xi' \lesssim \left(-\frac{p^{K-1}}{r^{K-1}} \Delta_{\min}^2\right).$$

Finally, we verify the bounded loss condition (61) for algorithm estimator by induction. With output  $z^{(0)}$  from Sub-algorithm 2 and the assumption  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$ , by Theorem 4, we have

$$L^{(0)} \leq \frac{\bar{C} \Delta_{\min}^2}{\tilde{C} r \log p}, \quad \text{when } p \text{ is large enough.}$$

Therefore, the condition (61) holds for  $t = 0$ . Assume the condition (61) also holds for all  $t \leq t_0$ . Then, by the decomposition (73), we have

$$\begin{aligned} L^{(t_0+1)} &\leq \frac{3}{2} \xi + \frac{1}{2} L^{(t_0)} \\ &\leq \exp\left(-M \frac{p^{K-1}}{r^{K-1}} \Delta_{\min}^2\right) + \frac{\Delta_{\min}^2}{r \log p} \\ &\leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \end{aligned}$$

where the second inequality follows from the condition (60) and the last inequality follows from the assumption that  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ . Thus, the condition (61) holds for  $t_0 + 1$ , and the condition (61) is proved by induction.  $\square$

## Useful Definitions and Lemmas for the Proof of Theorem 5

**Lemma 9** (Singular-value property of membership matrices). Under the setup of Theorem 5, suppose the condition (61) holds. Then, for all  $a \in [r]$ , we have  $|z^{(t)}(a)| \asymp p/r$ . Moreover, we have

$$\lambda_r(\mathbf{M}) \asymp \|\mathbf{M}\|_\sigma \asymp \sqrt{p/r}, \quad \lambda_r(\mathbf{W}) \asymp \|\mathbf{W}\|_\sigma \asymp \sqrt{r/p}, \quad \lambda_r(\mathbf{P}) \asymp \|\mathbf{P}\|_\sigma \asymp \min_{a \in [r]} \|\theta_{z^{-1}(a)}\|^{-1} \lesssim \sqrt{r/p}. \quad (81)$$

The inequalities (81) also hold by replacing  $\mathbf{M}$  and  $\mathbf{W}$  to  $\mathbf{M}^{(t)}$  and  $\mathbf{W}^{(t)}$  respectively. Further, we have

$$\lambda_r(\mathbf{W}\mathbf{W}^T) \asymp \|\mathbf{W}\mathbf{W}^T\|_\sigma \asymp r/p, \quad (82)$$

which is also true for  $\mathbf{W}^{(t)}\mathbf{W}^{(t),T}$ .

*Proof of Lemma 9.* The proof for the inequality (81) for  $\mathbf{M}, \mathbf{W}$  can be found in Han et al. (2020, Proof of Lemma 4). The inequalities for  $\mathbf{P}$  follows the same derivation with balance assumption on  $\theta$  and  $\min_{i \in [p]} \theta(i) \geq c$ .

For inequality (82), note that for all  $k \in [r]$ ,

$$\lambda_k(\mathbf{W}\mathbf{W}^T) = \sqrt{\text{eigen}_k(\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T)} \asymp \sqrt{\frac{r}{p} \text{eigen}_k(\mathbf{W}\mathbf{W}^T)} = \sqrt{\frac{r}{p} \lambda_k^2(\mathbf{W})} \asymp \frac{r}{p},$$

where  $\text{eigen}_k(\mathbf{A})$  denotes the  $k$ -th largest eigenvalue of the square matrix  $\mathbf{A}$ , the first inequality follows the fact that  $\mathbf{W}^T\mathbf{W}$  is a diagonal matrix with elements of order  $r/p$ , and the second equation follows from the definition of singular value.  $\square$

**Lemma 10** (Upper bound for  $F_{ib}^{(t)}, G_{ib}^{(t)}$  and  $H_{ib}^{(t)}$ ). Under the Condition 1 and the setup of Theorem 5, assume the constant  $\tilde{C}$  in the condition  $\text{SNR} \geq \tilde{C}p^{-K/2} \log p$  is large enough to satisfy the inequalities (94) and (98). We have

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(F_{ib}^{(t)}\right)^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2} \|\mathbf{E}_{i:}\mathbf{V}\|^2 + \left(1 + \frac{rL^{(t)}}{\Delta_{\min}^2}\right) \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2, \quad (83)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(G_{ib}^{(t)}\right)^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{512} (\Delta_{\min}^2 + L^{(t)}), \quad (84)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{|H_{ib}^{(t)}|}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{4}. \quad (85)$$

Similarly, when the  $\text{SNR} \geq \tilde{C}p^{-(K-1)} \log p$  with large constant  $\tilde{C}$ , we have

$$\begin{aligned} \max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(\hat{F}_{ib}\right)^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} &\lesssim p^{K-1} \frac{rL(\hat{z})}{\Delta_{\min}^2} \\ \max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(\hat{G}_{ib}\right)^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} &\leq \frac{1}{512} (\Delta_{\min}^2 + L(\hat{z})), \\ \max_{i \in [p]} \max_{b \neq z(i)} \frac{|\hat{H}_{ib}|}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} &\leq \frac{1}{4}. \end{aligned}$$

*Proof of Lemma 10.* We prove the the first three inequalities in Lemma 10 separately.

1) Upper bound for  $F_{ib}^{(t)}$ , i.e., inequality (83). Recall the definition of  $F_{ib}^{(t)}$ ,

$$F_{ib}^{(t)} = 2 \left\langle \mathbf{E}_{i:}\mathbf{V}^{(t)}, \left([\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s\right) - \left([\tilde{\mathbf{S}}_b]_s - [\mathbf{S}_b]_s\right) \right\rangle + 2 \left\langle \mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle.$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\left( F_{ib}^{(t)} \right)^2 &\leq 8 \left( \left\langle \mathbf{E}_{i:} \mathbf{V}^{(t)}, \left[ [\tilde{\mathbf{S}}_{z(i):}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right] - \left[ [\tilde{\mathbf{S}}_{b:}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right] \right\rangle \right)^2 \\
&\quad + 8 \left( \left\langle \mathbf{E}_{i:} (\mathbf{V} - \mathbf{V}^{(t)}), \left[ [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right] \right\rangle \right)^2 \\
&\leq 8 \left( \|\mathbf{E}_{i:} \mathbf{V}\|^2 + \|\mathbf{E}_{i:} (\mathbf{V} - \mathbf{V}^{(t)})\|^2 \right) \max_{a \in [r]^s} \|[\tilde{\mathbf{S}}_{a:}]^s - [\mathbf{S}_{a:}^{(t)}]^s\| \\
&\quad + \|\mathbf{E}_{i:} (\mathbf{V} - \mathbf{V}^{(t)})\|^2 \|[\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s\|. \tag{86}
\end{aligned}$$

Note that for all  $a \in [r]$ ,

$$\begin{aligned}
\|[\tilde{\mathbf{S}}_{a:}]^s - [\mathbf{S}_{a:}^{(t)}]^s\|^2 &= \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\
&\leq 2 \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]^s\|^2 + 2 \|[\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\
&\lesssim \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\
&\lesssim rL^{(t)} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\
&\lesssim rL^{(t)}, \tag{87}
\end{aligned}$$

where the second inequality follows from the inequalities (104) and (105) in Lemma 11, the third inequality follows from the condition (61) in Condition 1, and the last inequality follows from the assumption that  $\Delta_{\min}^2 \geq \tilde{C}p^{-K/2} \log p$ .

Note that

$$\begin{aligned}
\|[\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s\|^2 &= \|[\tilde{\mathbf{S}}_{z(i):}]^s - [\mathbf{S}_{z(i):}]^s + [\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s + [\mathbf{S}_{b:}]^s - [\tilde{\mathbf{S}}_{b:}]^s\|^2 \\
&\lesssim \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 + \max_{a \in [r]} \|[\mathbf{S}_{a:}]^s - [\tilde{\mathbf{S}}_{a:}]^s\|^2 \\
&\lesssim \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 + \max_{a \in [r]} \frac{1}{\|\mathbf{S}_{a:}\|^2} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}\|^2 \\
&\lesssim \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \tag{88}
\end{aligned}$$

where the second inequality follows from Lemma 4, and the last inequality follows from the assumptions on  $\|\mathbf{S}_{a:}\|$  in the parameter space (3), the inequality (57) in Condition 1 and the assumption  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ .

Therefore, we finish the proof of inequality (83) by plugging the inequalities (87) and (88) into the upper bound (86).

2) Upper bound for  $G_{ib}^{(t)}$ , i.e., inequality (84). By definition of  $G_{ib}^{(t)}$ , we rearrange terms and obtain

$$\begin{aligned}
G_{ib}^{(t)} &= \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\
&\quad - \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\
&= 2 \left\langle [\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s, \left( [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right) - \left( [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right) \right\rangle \\
&= G_1 + G_2 - G_3, \tag{89}
\end{aligned}$$

where

$$\begin{aligned}
G_1 &= \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 - \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2, \\
G_2 &= 2 \left\langle [\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s, [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right\rangle, \\
G_3 &= 2 \left\langle [\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s, [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right\rangle.
\end{aligned}$$

For  $G_1$ , we have

$$\begin{aligned}
|G_1|^2 &\leq \left| \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 - \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \right|^2 \\
&\leq \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^4 \\
&\leq C^4 \frac{r^4}{\Delta_{\min}^4} (L^{(t)})^4 + \frac{r^2 r^{4K} + p^2 r^{2K+4}}{p^{2K}} \frac{(L^{(t)})^2}{\Delta_{\min}^4} \\
&\leq C^4 \frac{\bar{C}}{\tilde{C}^3} \left( \Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)} \right), \tag{90}
\end{aligned}$$

where the third inequality follows from the inequality (106) in Lemma 9 and the last inequality follows from the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (61) in Condition 1.

For  $G_2$ , noticing that  $[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s = [\mathbf{W}_{z(i):}^T \mathbf{X} \mathbf{V}^{(t)}]^s$ , we have

$$\begin{aligned}
|G_2|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \\
&\leq \frac{2}{\|\mathbf{W}_{z(i):}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\
&\leq C' \frac{r^{2K-1} + K p r^{K+1}}{p^K} \left( \frac{r^2}{\Delta_{\min}^2} (L^{(t)})^2 + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) \\
&\leq \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)}, \tag{91}
\end{aligned}$$

where  $C'$  is a positive universal constant, the second inequality follows from Lemma 4, the third inequality follows from the inequality (58) in Condition 1, the inequalities (106) and (125) in the proof of Lemma 11, and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (61) in Condition 1.

For  $G_3$ , note that by triangle inequality

$$\begin{aligned}
\|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 &\leq \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}]^s\|^2 \\
&\leq \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2}, \tag{92}
\end{aligned}$$

where the last inequality follows from the inequality (124) in the proof of Lemma 11 and  $C$  is a positive constant. Then we have

$$\begin{aligned}
|G_3|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\
&\leq 2 \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 + \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \right) \\
&\quad \times \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\
&\leq C^2 \left( \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \right) \left( \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) + \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)} \\
&\leq \frac{C^2 \bar{C}^2}{\tilde{C}} \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 (\Delta_{\min}^2 + L^{(t)}) + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} \left( \Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)} \right), \tag{93}
\end{aligned}$$

where the third inequality follows from the same procedure to derive (90) and (91), and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (61) in Condition 1.

Choose the  $\tilde{C}$  such that

$$3 \left( C^4 \frac{\bar{C}}{\tilde{C}^3} + \frac{C'}{\tilde{C}^2} + \frac{C^2 \bar{C}^2}{\tilde{C}} + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} \right) \leq \frac{1}{512}. \tag{94}$$

Then, we finish the proof of inequality (84) by plugging the inequalities (90), (91), and (93) into the upper bound (89).

3) Upper bound for  $H_{ib}^{(t)}$ , i.e., the inequality (85). By definition of  $H_{ib}$ , we rearrange terms and obtain

$$\begin{aligned} H_{ib} &= \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 - \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 + \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \\ &= \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\quad + \left( \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 - \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\| \right) \\ &\quad - \left( \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\| - \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\| \right) \\ &= H_1 + H_2 + H_3, \end{aligned}$$

where

$$\begin{aligned} H_1 &= \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 - \|[\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2, \\ H_2 &= \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 - \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2, \\ H_3 &= 2 \left\langle [\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s, [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s \right\rangle. \end{aligned}$$

For  $H_1$ , we have

$$|H_1| \leq \frac{4 \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2}{\|\mathbf{W}_{z(i):}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \leq \frac{r^{2K-1} + K p r^{K+1}}{p^K} \leq \tilde{C}^{-2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \quad (95)$$

following the derivation of  $G_2$  in inequality (91) and the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$ .

For  $H_2$ , by the inequality (92), we have

$$|H_2| \lesssim 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}]\|^2 \lesssim \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \leq C \frac{\bar{C}^2}{\tilde{C}^2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \quad (96)$$

where the last inequality follows from the condition (61) in Condition 1.

For  $H_3$ , by Cauchy-Schwartz inequality, we have

$$|H_3| \lesssim \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\| |H_1|^{1/2} \leq 2 \tilde{C}^{-1} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \quad (97)$$

following the inequalities (92) and (95).

Choose  $\tilde{C}$  such that

$$\tilde{C}^{-2} + C \frac{\bar{C}^2}{\tilde{C}^2} + \tilde{C}^{-1} \leq \frac{1}{4}. \quad (98)$$

Therefore, we finish the proof of inequality (85) combining inequalities (95), (96), and (97).

**Next, we show the upper bounds for  $\hat{F}_{ib}$ ,  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ .**

By Lemma 1, we have

$$\|\mathbf{S}_{a:}^s - \mathbf{S}_{b:}^s\| = (1 + o(1)) \|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|.$$

Also, notice that the matrix product of  $\mathbf{B}^T$  corresponding to the padding operation in Lemma 5, and the padding weights are balanced such that  $\|\mathbf{v} \mathbf{B}\| = (1 + o(1)) \max_a \|\boldsymbol{\theta}_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{v}\|$  for all  $\mathbf{v} \in \mathbb{R}^{r(K-1)}$ . For two vectors  $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{r^{K-1}}$ , we have

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = (1 + o(1)) \|[\mathbf{v}_1 \mathbf{B}^T]^s - [\mathbf{v}_2 \mathbf{B}^T]^s\|. \quad (99)$$

The equation (99) also holds for  $\hat{\mathbf{B}}^T$ .

Note that for all  $i \in [p]$  we have

$$\|\mathbf{A}_{i:} \hat{\mathbf{Q}}\| = \|\mathbf{S}_{z(i):} \mathbf{B}^T \hat{\mathbf{Q}}\| = \|\mathbf{S}_{z(i):} \hat{\mathbf{D}}^{\otimes(K-1)}\| = (1 + o(1)) \|\mathbf{S}_{z(i):}\| = (1 + o(1)) \max_a \|\boldsymbol{\theta}_{z^{-1}(a)}\|^{-(K-1)/2} \|\mathbf{A}_{i:}\|, \quad (100)$$

where the third inequality follows from the singular property of MLE confusion matrix (131) and the last inequality follows from the fact that  $\mathbf{A}_i = \mathbf{S}_{z(i)} \mathbf{B}^T$  and Lemma 9. Above equation indicates that  $\mathbf{A}_{i:}$  is the span space of the singular values as  $p \rightarrow \infty$ . Also, notice that the row space of  $\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T$  is equal to the column space of  $\hat{\mathbf{Q}}$ , and  $\mathbf{A}_{i:} \neq \mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T$  in noisy case.

Hence, for all  $a \in [r]$ , we have

$$\begin{aligned} \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\| &= \left\| \frac{\mathbf{A}_{z(i)} \hat{\mathbf{Q}}}{\|\mathbf{A}_{z(i)} \hat{\mathbf{Q}}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|} \right\| \\ &= (1 + o(1)) \left\| \frac{\mathbf{A}_{z(i)}}{\|\mathbf{A}_{z(i)}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\|} \right\| \\ &= (1 + o(1)) \|[\mathbf{X}_i]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\| \end{aligned} \quad (101)$$

where the second equation follows from (100),  $\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\| = (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|$ , and singular property of  $\hat{\mathbf{B}}^T$ . Similar result holds after replacing  $\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}$  by  $\mathbf{P}_{:a}^T \mathbf{Y} \mathbf{Q}$  or  $\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}$ .

We are now ready to show the upper bounds for  $\hat{F}_{ib}$ ,  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ .

For  $\hat{F}_{ib}$ , we have

$$\begin{aligned} (\hat{F}_{ib})^2 &\leq \|\mathbf{E}_i\|^2 \|[\bar{\mathbf{A}}_{a:}]^s - [\hat{\mathbf{A}}_{a:}]^s\|^2 \\ &\leq \|\mathbf{E}_i\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T]^s - [\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| + \|[\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s - [\hat{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| \right]^2 \\ &\lesssim \|\mathbf{E}_i\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T \hat{\mathbf{Q}}]^s - [\bar{\mathbf{S}}_{a:}]^s\| + \|[\bar{\mathbf{S}}_{a:}]^s - [\hat{\mathbf{S}}_{a:}]^s\| \right]^2. \end{aligned}$$

Following similar derivations in inequalities (87), (88), and the upper bound for  $J_1$ , respectively, we have

$$\|[\bar{\mathbf{S}}_{a:}]^s - [\hat{\mathbf{S}}_{a:}]^s\| \lesssim rL(\hat{z}), \quad \|[\bar{\mathbf{S}}_{a:}]^s - [\bar{\mathbf{S}}_{b:}]^s\| \lesssim \|\mathbf{S}_{a:}^s - \mathbf{S}_{b:}^s\|^2, \quad \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T \hat{\mathbf{Q}}]^s - [\bar{\mathbf{S}}_{a:}]^s\| \lesssim L(\hat{z}).$$

We then obtain the upper bound for  $\hat{F}_{ib}$  by noticing that  $\|\mathbf{E}_i\|^2 \lesssim p^{K-1}$ .

For  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ , by the property (101), we have

$$\begin{aligned} (1 + o(1)) \hat{G}_{ib} &= \left( \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\hat{\mathbf{S}}_{a:}]^s\|_F^2 - \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 \right) \\ &\quad - \left( \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\hat{\mathbf{S}}_{b:}]^s\|_F^2 - \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 \right). \\ (1 + o(1)) \hat{H}_{ib} &= \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 - \|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 + \|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|_F^2. \end{aligned}$$

We obtain the upper bounds following the proof for inequalities (84) and (85).

□

**Lemma 11** (Relationship between misclustering loss and intermediate parameters). Under the Condition 1 and the setup of Theorem 5, we have

$$\|\mathbf{V} - \mathbf{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}} \frac{r}{\Delta_{\min}^2} L^{(t)}, \quad (102)$$

$$\|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}}} \frac{r}{\Delta_{\min}^2} L^{(t)}, \quad (103)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}]^s\| \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}} \right), \quad (104)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^{(t),T} \mathbf{YV}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}]^s\| \leq C \left( \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{rL^{(t)}}{\Delta_{\min}} \right), \quad (105)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^T \mathbf{YV}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}]^s\| \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}} \right), \quad (106)$$

for some positive universal constant  $C$ . In addition, the inequality (105) also holds by replacing  $\mathbf{W}_{:b}^{(t)}$  to  $\mathbf{W}_{:b}$ . Further, the above inequalities holds after replacing  $\mathbf{W}$  to  $\mathbf{P}$ ,  $\mathbf{V}$  to  $\mathbf{Q}$ , and  $L^{(t)}$  to  $L(\hat{z})$ .

*Proof of Lemma 11.* We follow and use several intermediate conclusions in Han et al. (2020, Proof of Lemma 5). We prove each inequality separately.

1) Inequality (102). By Han et al. (2020, Proof of Lemma 5), we have

$$\|\mathbf{V} - \mathbf{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}} r\ell^{(t)}.$$

Then, we complete the proof of inequality (102) by applying Lemma 2 to the above inequality.

2) Inequality (103). By Han et al. (2020, Proof of Lemma 5), we have

$$\|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}}} r\ell^{(t)}.$$

Also, we complete the proof of inequality (102) by applying Lemma 2 to the above inequality.

3) Inequality (104). We upper bound the desired quantity by triangle inequality,

$$\|[\mathbf{W}_{:b}^T \mathbf{YV}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{YV}]^s\| \leq I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 &= \left\| \frac{\mathbf{W}_{:b}^T \mathbf{YV}}{\|\mathbf{W}_{:b}^T \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{YV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right\|, \\ I_2 &= \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^T \mathbf{YV}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right) \mathbf{W}_{:b}^T \mathbf{YV} \right\|, \\ I_3 &= \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{YV}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{YV} \right\|. \end{aligned}$$

Next, we upper bound the quantities  $I_1, I_2, I_3$  separately.

For  $I_1$ , we further bound  $I_1$  by triangle inequality,

$$I_1 \leq I_{11} + I_{12},$$

where

$$I_{11} = \left\| \frac{\mathbf{W}_{:b}^T \mathbf{XV}}{\|\mathbf{W}_{:b}^T \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{XV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right\|, \quad \text{and} \quad I_{12} = \left\| \frac{\mathbf{W}_{:b}^T \mathbf{EV}}{\|\mathbf{W}_{:b}^T \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{EV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right\|.$$

We first consider  $I_{11}$ . Define the confusion matrix  $\mathbf{D} = \mathbf{M}^T \Theta^T \mathbf{W}^{(t)} = \llbracket D_{ab} \rrbracket \in \mathbb{R}^{r \times r}$  where

$$D_{ab} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = a, z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}}, \quad \text{for all } a, b \in [r].$$

By Lemma 9, we have  $\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} \gtrsim p/r$ . Then, we have

$$\sum_{a \neq b, a, b \in [r]} D_{ab} \lesssim \frac{r}{p} \sum_{i: z^{(t)}(i) \neq z(i)} \theta(i) \lesssim \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \frac{1}{\log p}, \quad (107)$$

and for all  $b \in [r]$ ,

$$D_{bb} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \geq \frac{c(\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} - p\ell^{(t)})}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \gtrsim 1 - \frac{1}{\log p}, \quad (108)$$

under the inequality (61) in Condition 1. By the definition of  $\mathbf{W}, \mathbf{W}^{(t)}, \mathbf{V}$ , we have

$$\frac{\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} = [\mathbf{S}_{b:}]^s, \quad \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} = [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}]^s.$$

Let  $\alpha$  denote the angle between  $\mathbf{S}_{b:}$  and  $D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}$ . To roughly estimate the range of  $\alpha$ , we consider the inner product

$$\begin{aligned} \left\langle \mathbf{S}_{b:}, D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \right\rangle &= D_{bb} \|\mathbf{S}_{b:}\|^2 + \sum_{a \neq b} D_{ab} \langle \mathbf{S}_{b:}, \mathbf{S}_{a:} \rangle \\ &\geq D_{bb} \|\mathbf{S}_{b:}\|^2 - \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{b:}\| \max_{a \in [r]} \|\mathbf{S}_{a:}\| \\ &\geq C, \end{aligned}$$

where  $C$  is a positive constant, and the last inequality holds when  $p$  is large enough following the constraint of  $\|\mathbf{S}_{b:}\|$  in parameter space (3) and the bounds of  $\mathbf{D}$  in (107) and (108).

The positive inner product between  $\mathbf{S}_{b:}$  and  $D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}$  indicates  $\alpha \in [0, \pi/2]$ , and thus  $2 \sin \frac{\alpha}{2} \leq \sqrt{2} \sin \alpha$ . Then, by the geometry property of trigonometric function, we have

$$\begin{aligned} \| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \| &= \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \mathbf{S}_{a:} \| \\ &= \sum_{a \neq b, a \in [r]} D_{ab} \| \mathbf{S}_{a:} \sin(\mathbf{S}_{b:}, \mathbf{S}_{a:}) \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\|, \end{aligned} \quad (109)$$

where the first inequality follows from the triangle inequality, and the last inequality follows from Lemma 4. Note that with bounds (107) and (108), when  $p$  is large enough, we have

$$\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| = \|D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}\| \geq D_{bb} \|\mathbf{S}_{b:}\| - \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \geq C_1, \quad (110)$$

for some positive constant  $C_1$ . Notice that  $I_{11} = \sqrt{1 - \cos \alpha} = 2 \sin \frac{\alpha}{2}$ . Therefore, we obtain

$$\begin{aligned} I_{11} &\leq \sqrt{2} \sin \alpha \\ &= \frac{\| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \|}{\|D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}\|} \\ &\leq \frac{1}{C_1} \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \end{aligned}$$

$$\begin{aligned}
&\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ z^{(t)}(i) = b \right\} \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \\
&\leq \frac{rL^{(t)}}{\Delta_{\min}}, 
\end{aligned} \tag{111}$$

where the second inequality follows from (109) and (110), and the last two inequalities follow by the definition of  $D_a$  and  $L^{(t)}$ , and the constraint of  $\|\mathbf{S}_{b:}\|$  in parameter space (3).

We now consider  $I_{12}$ . By triangle inequality, we have

$$I_{12} \leq \frac{1}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} \|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V}\| + \frac{\|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V}\|}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\|.$$

By Han et al. (2020, Proof of Lemma 5), we have

$$\|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V}\| \lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{112}$$

Notice that

$$\|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V}\| \leq \|\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}\| \|\mathbf{X} \mathbf{V}\|_F \lesssim \frac{r^{3/2} L^{(t)}}{\sqrt{p} \Delta_{\min}^2} \|\mathbf{S}\| \|\Theta \mathbf{M}\|_\sigma \lesssim \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}}, \tag{113}$$

where the second inequality follows from Han et al. (2020, Inequality (121), Proof of Lemma 5) and the last inequality follows from Lemma 6 and (61) in Condition 1. Note that  $\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\| = \|\mathbf{S}_{b:}\| \geq c_3$  and  $\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| \geq C_1$  by inequality (110). Therefore, we have

$$\begin{aligned}
I_{12} &\lesssim \|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V}\| + \|(\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\| \\
&\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}} \sqrt{\frac{r^{2K}}{p^K}} \\
&\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}},
\end{aligned} \tag{114}$$

where second inequality follows from the inequalities (112), (113), and (57) in Condition 1.

Hence, combining inequalities (111) and (114) yields

$$I_1 \lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{115}$$

For  $I_2$  and  $I_3$ , recall that  $\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\| = \|\mathbf{S}_{b:}\| \geq c_3$  and  $\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| \geq C_1$  by inequality (110). By triangle inequality and (57) in Condition 1, we have

$$I_2 \leq \frac{\|\mathbf{W}_{:b}^T \mathbf{E} \mathbf{V}\|}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^T \mathbf{E} \mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{116}$$

and

$$I_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{117}$$

Therefore, combining the inequalities (115), (116), and (117), we finish the proof of inequality (104).

4) Inequality (105). Here we only show the proof of inequality (105) with  $\mathbf{W}_{:b}^{(t)}$ . The proof also holds by replacing  $\mathbf{W}_{:b}^{(t)}$  to  $\mathbf{W}_{:b}$ , and we omit the repeated procedures.

We upper bound the desired quantity by triangle inequality

$$\|[\mathbf{W}_{:b}^{(t),T} \mathbf{YV}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}]^s\| \leq J_1 + J_2 + J_3,$$

where

$$\begin{aligned} J_1 &= \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{YV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}\|} \right\|, \\ J_2 &= \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{YV}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{YV} \right\|, \\ J_3 &= \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)} \right\|. \end{aligned}$$

Next, we upper bound the quantities  $J_1, J_2, J_3$  separately.

For  $J_1$ , by triangle inequality, we have

$$J_1 \leq J_{11} + J_{12},$$

where

$$J_{11} = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{XV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}\|} \right\| \quad \text{and} \quad J_{12} = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{EV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{EV}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}\|} \right\|.$$

We first consider  $J_{11}$ . Define the matrix  $\mathbf{V}^k := \mathbf{W}^{\otimes(k-1)} \otimes \mathbf{W}^{(t),\otimes(K-k)}$  for  $k = 2, \dots, K-1$ , and denote  $\mathbf{V}^1 = \mathbf{V}^{(t)}, \mathbf{V}^K = \mathbf{V}$ . Also, define the quantity

$$J_{11}^k = \|[\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1}]^s\|,$$

for  $k = 1, \dots, K-1$ . Let  $\beta_k$  denote the angle between  $\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k$  and  $\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1}$ . With the same idea to prove  $I_{11}$  in inequality (111), we bound  $J_{11}^k$  by the trigonometric function of  $\beta_k$ .

To roughly estimate the range of  $\beta_k$ , we consider the inner product between  $\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k$  and  $\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1}$ . Before the specific derivation of the inner product, note that

$$\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k = \text{Mat}_1(\mathcal{T}_k), \quad \mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1} = \text{Mat}_1(\mathcal{T}_{k+1}),$$

where

$$\begin{aligned} \mathcal{T}_k &= \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T \times_{k+1} \mathbf{W}^{(t),T} \times_{k+1} \cdots \times_K \mathbf{W}^{(t),T} \\ \mathcal{T}_{k+1} &= \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T \times_{k+1} \mathbf{W}^T \times_{k+1} \cdots \times_K \mathbf{W}^{(t),T}. \end{aligned}$$

Recall the definition of confusion matrix  $\mathbf{D} = \mathbf{M}^T \boldsymbol{\Theta}^T \mathbf{W}^{(t)} = \llbracket D_{ab} \rrbracket \in \mathbb{R}^{r \times r}$ . We have

$$\begin{aligned} \langle \mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k, \mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1} \rangle &= \langle \text{Mat}_{k+1}(\mathcal{T}_k), \text{Mat}_{k+1}(\mathcal{T}_{k+1}) \rangle \\ &= \langle \mathbf{D}^T \mathbf{S} \mathbf{Z}^k, \mathbf{S} \mathbf{Z}^k \rangle \\ &= \sum_{b \in [r]} \left( D_{bb} \|\mathbf{S}_{b:} \mathbf{Z}^k\|^2 + \sum_{a \neq b, a \in [r]} D_{ab} \langle \mathbf{S}_{a:} \mathbf{Z}^k, \mathbf{S}_{b:} \mathbf{Z}^k \rangle \right) \\ &\gtrsim (1 - \log p^{-1}) \min_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2 - \log p^{-1} \max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2, \end{aligned} \tag{118}$$

where  $\mathbf{Z}^k = \mathbf{D}_{:b} \otimes \mathbf{I}_r^{\otimes(k-1)} \otimes \mathbf{D}^{\otimes(K-k-1)}$ , the equations follow by the tensor algebra and definitions, and the last inequality follows from the bounds of  $\mathbf{D}$  in (107) and (108).

Note that

$$\|\mathbf{D}\|_\sigma \leq \|\mathbf{D}\|_F \leq \sqrt{\sum_{b \in [r]} D_{bb}^2 + (\sum_{a \neq b, a, b \in [r]} D_{ab})^2} \lesssim \sqrt{r + \log^2 p^{-1}} \lesssim 1, \quad (119)$$

where the second inequality follows from inequality (107), and the fact that for all  $b \in [r]$ ,

$$D_{bb} \lesssim \frac{r}{p} \sum_{i: z(i)=b} \theta(i) \lesssim 1.$$

Also, we have

$$\lambda_r(\mathbf{D}) \geq \lambda_r(\mathbf{W}^{(t)}) \lambda_r(\Theta \mathbf{M}) \gtrsim 1, \quad (120)$$

following the Lemma 6 and Lemma 9. Then, for all  $k \in [K]$ , we have

$$1 \lesssim \|\mathbf{D}_{:b}\| \lambda_r(\mathbf{D})^{K-k-1} \leq \lambda_{r^{K-2}}(\mathbf{Z}^k) \leq \|\mathbf{Z}^k\|_\sigma \leq \|\mathbf{D}_{:b}\| \|\mathbf{D}\|_\sigma^{K-k-1} \lesssim 1. \quad (121)$$

Thus, we have bounds

$$\max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\| \leq \max_{a \in [r]} \|\mathbf{S}_{a:}\| \|\mathbf{Z}^k\|_\sigma \lesssim 1, \quad \min_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\| \geq \min_{a \in [r]} \|\mathbf{S}_{a:}\| \lambda_{r^{K-2}}(\mathbf{Z}^k) \gtrsim 1.$$

Hence, when  $p$  is large enough, the inner product (118) is positive, which implies  $\beta_k \in [0, \pi/2)$  and thus  $2 \sin \frac{\beta_k}{2} \leq \sqrt{2} \sin \beta_k$ .

Next, we upper bound the trigonometric function  $\sin \beta_k$ . Note that

$$\begin{aligned} \sin \beta_k &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}, \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}) \\ &\leq \sin \beta_{k1} + \sin \beta_{k2}, \end{aligned}$$

where

$$\begin{aligned} \sin \beta_{k1} &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}, \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \tilde{\mathbf{D}} \otimes \mathbf{D}^{\otimes K-k-1}), \\ \sin \beta_{k2} &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \tilde{\mathbf{D}} \otimes \mathbf{D}^{\otimes K-k-1}, \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}), \end{aligned}$$

and  $\tilde{\mathbf{D}}$  is the normalized confusion matrix with entries  $\tilde{\mathbf{D}}_{ab} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z^{(t)}=b, z(i)=a\}}{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z^{(t)}=b\}}$ .

To bound  $\sin \beta_{k1}$ , recall Definition 2 that for any cluster assignment  $\bar{z}$  in the  $\varepsilon$ -neighborhood of true  $z$ ,

$$\mathbf{p}(\bar{z}) = (|\bar{z}^{-1}(1)|, \dots, |\bar{z}^{-1}(r)|)^T, \quad \mathbf{p}_\theta(\bar{z}) = (\|\theta_{\bar{z}^{-1}(1)}\|_1, \dots, \|\theta_{\bar{z}^{-1}(r)}\|_1)^T.$$

With the locally linear stability assumption, we have

$$\sin(\mathbf{p}(z^{(t)}), \mathbf{p}_\theta(z^{(t)})) \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

Note that  $\text{diag}(\mathbf{p}(z^{(t)})) \mathbf{D} = \text{diag}(\mathbf{p}_\theta(z^{(t)})) \tilde{\mathbf{D}}$ , and  $\sin(\mathbf{a}, \mathbf{b}) = \min_{c \in \mathbb{R}} \frac{\|\mathbf{a} - c\mathbf{b}\|}{\|\mathbf{a}\|}$  for vectors  $\mathbf{a}, \mathbf{b}$  of same dimension. Let  $c_0 = \arg \min_{c \in \mathbb{R}} \frac{\|\mathbf{p}(z^{(t)}) - c\mathbf{p}_\theta(z^{(t)})\|}{\|\mathbf{p}(z^{(t)})\|}$ . Then, we have

$$\begin{aligned} \min_{c \in \mathbb{R}} \|\mathbf{D} - c\tilde{\mathbf{D}}\|_F &\leq \|\mathbf{I}_r - c_0 \text{diag}(\mathbf{p}(z^{(t)})) \text{diag}^{-1}(\mathbf{p}_\theta(z^{(t)}))\|_F \|\mathbf{D}\|_F \\ &\lesssim \frac{\|\mathbf{p}(z^{(t)}) - c_0 \mathbf{p}_\theta(z^{(t)})\|}{\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1} \\ &= \frac{\|\mathbf{p}(z^{(t)})\|}{\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1} \sin(\mathbf{p}(z^{(t)}), \mathbf{p}_\theta(z^{(t)})) \\ &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned}$$

where the last inequality follows from Lemma 9, the constraint  $\min_{i \in [p]} \theta(i) \geq c > 0$ ,  $\|\mathbf{p}(z^{(t)})\| \lesssim p$  and

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{(t)}, -1}(a)\|_1 \gtrsim p.$$

By the geometry property of trigonometric function, we have

$$\begin{aligned} \sin \beta_{k1} &= \min_{c \in \mathbb{R}} \frac{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{D} - c\tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1}\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}\|} \\ &\leq \frac{\|\mathbf{D}_{:b}^T \mathbf{S}\| \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_\sigma \|\mathbf{D}\|_\sigma^{K-k-1}}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k}(\mathbf{D})} \\ &\lesssim \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_F \\ &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{122}$$

where the second inequality follows from the singular property of  $\mathbf{D}$  in (119), (120) and the constraint of  $\mathbf{S}$  in (3).

To bound  $\sin \beta_{k2}$ , let  $\mathbf{C} = \text{diag}(\{\|\mathbf{S}_{a:}\|\}_{a \in [r]})$ . We have

$$\begin{aligned} \sin \beta_{k2} &\lesssim \frac{\left\| \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{I}_r - \tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1} \right\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}\|} \\ &\lesssim \frac{\|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{Z}^k\|_F}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k-1}(\mathbf{D})} \\ &\lesssim \|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{C}^{-1}\|_F \|\mathbf{C} \mathbf{Z}^k\|_\sigma \\ &\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbf{1}\{z^{(t)}(i) = b\} \|\mathbf{S}_{b:}^s - \mathbf{S}_{z(i):}^s\| \\ &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{123}$$

where the third inequality follows from the singular property of  $\mathbf{D}$  and the boundedness of  $\mathbf{S}$ , and the fourth inequality follows from the definition of  $\tilde{\mathbf{D}}$ , boundedness of  $\mathbf{S}$ , the lower bound of  $\boldsymbol{\theta}$ , and the singular property of  $\mathbf{Z}^k$  in inequality (121), and the last line follows from the definition of  $L^{(t)}$ .

Combining (122) and (123) yields

$$\sin \beta_k \leq \sin \beta_{k1} + \sin \beta_{k2} \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

Finally, by triangle inequality, we obtain

$$J_{11} \leq \sum_{k=1}^{K-1} J_{11}^k \lesssim \sum_{k=1}^{K-1} \sin \beta_k \lesssim (K-1) \frac{rL^{(t)}}{\Delta_{\min}}. \tag{124}$$

We now consider  $J_{12}$ . By triangle inequality, we have

$$J_{12} \leq \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| + \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}^{(t)}\|.$$

Note that

$$\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\| = \|\mathbf{D}^T \mathbf{S} \mathbf{Z}^1\| \geq \lambda_r(\mathbf{D}) \|\mathbf{S}\| \lambda_{r,K-2}(\mathbf{Z}^1) \gtrsim 1, \tag{125}$$

where the inequality follows from the bounds (120) and (121).

By Han et al. (2020, Proof of Lemma 5), we have

$$\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| \lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}} \frac{(K-1)\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{126}$$

Notice that

$$\begin{aligned}
\|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F &\leq \|(\mathbf{I} - \mathbf{D}^T)\mathbf{S}(\mathbf{I}_r^{\otimes(k-1)} \otimes \mathbf{D}^{\otimes(K-k-1)})\|_F \\
&\leq \|(\mathbf{W}^T - \mathbf{W}^{(t),T})\Theta\mathbf{M}\|_F \|\mathbf{S}\|_F \|\mathbf{D}\|_{\sigma}^{K-k-1} \\
&\lesssim \|\mathbf{W}^T - \mathbf{W}^{(t),T}\| \|\Theta\mathbf{M}\|_{\sigma} \\
&\lesssim \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}, 
\end{aligned} \tag{127}$$

where the first inequality follows from the tensor algebra in inequality (118), the second inequality follows from the fact that  $\mathbf{I} = \mathbf{W}^T\Theta\mathbf{M}$ , and the last inequality follows from Han et al. (2020, Proof of Lemma 5). It follows from (127) and Lemma 9 that

$$\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\| \leq \|\mathbf{W}_{:b}^{(t),T}\| \sum_{k=1}^{K-1} \|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F \lesssim \frac{\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}}. \tag{128}$$

Note that  $\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|$  and  $\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}^{(t)}\|$  are lower bounded by inequalities (110) and (125), respectively. We have

$$\begin{aligned}
J_{12} &\lesssim \|\mathbf{W}_{:b}^{(t),T}\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| + \|\mathbf{W}_{:b}^{(t),T}\mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\| \|\mathbf{W}_{:b}^{(t),T}\mathbf{EV}^{(t)}\| \\
&\lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}} \sqrt{\frac{r^{2K}}{p^K}} \\
&\lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}},
\end{aligned}$$

where the second inequality follows from inequalities (126), (128), and the inequality (57) in Condition 1.

For  $J_2$  and  $J_3$ , recall that  $\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|$  and  $\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}^{(t)}\|$  are lower bounded by inequalities (110) and (125), respectively. By triangle inequality and inequality (57) in Condition 1, we have

$$J_2 \leq \frac{\|\mathbf{W}_{:b}^{(t),T}\mathbf{EV}\|}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T}\mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{129}$$

and

$$J_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T}\mathbf{EV}^{(t)}\|}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}^{(t)}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T}\mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{130}$$

Therefore, combining the inequalities (124), (129), and (130), we finish the proof of inequality (105).

5) Inequality (106). By triangle inequality, we upper bound the desired quantity

$$\begin{aligned}
\|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}^{(t)}]^s\| &\leq \|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}]^s\| + \|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}]^s\| \\
&\quad + \|[\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}^{(t)}]^s\| \\
&\lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K}} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}},
\end{aligned}$$

following the inequalities (104) and (105). Therefore, we finish the proof of inequality (106).

**Next, we show the intermediate inequalities holds with  $P, Q$  and  $L(\hat{z})$ .**

Consider the MLE confusion matrix  $\hat{\mathbf{D}} = \mathbf{M}^T\Theta^T\hat{\mathbf{P}} = \llbracket \hat{D}_{ab} \rrbracket \in \mathbb{R}^{r \times r}$  with entries

$$\hat{D}_{ab} = \frac{\sum_{i \in [p]} \theta(i)\hat{\theta}(i)\mathbf{1}\{z(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2} = \frac{\sum_{i \in [p]} (1 + o(p^{K-2}))(\hat{\theta}(i))^2 \mathbf{1}\{z(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2}, \tag{131}$$

where the second equation follows from Lemma 12, and thus  $\sum_{a \in [r]} \hat{D}_{ab} = 1 + o(1)$ . By the derivation of (107), (108), (120), and (119), we have

$$\sum_{a \neq b \in [r]} \hat{D}_{ab} \lesssim \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\hat{\theta}(i))^2 \lesssim \frac{1}{\log p}, \quad \hat{D}_{bb} \gtrsim 1 - \frac{1}{\log p}, \quad \lambda_{\min}(\hat{D}) \asymp \|\hat{D}\|_{\sigma} = (1 + o(1)).$$

for all  $a \neq b \in [r]$ .

Now, we are ready to show the intermediate inequalities. First, by Lemma 1 and  $\min_{i \in [p]} \theta(i) \geq c$ , we have

$$\|\mathbf{S}_{a:}^s - \mathbf{S}_{b:}^s\| \asymp \|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|.$$

Then we can replace the  $L^{(t)}$  by  $L(\hat{z})$  in the proof of Lemma 11. The analogies of inequalities (102), (103), (104), (105), and (106) hold by using the MLE confusion matrix and the definition of  $L(\hat{z})$ .

Particularly, for the analogy of (105), the usage of MLE confusion matrix avoids the stability condition on  $\theta$ . Let  $\bar{D}$  be the normalized version of  $\hat{D}$ . The angle in inequality (122) decays to 0 at speed  $p^{-(K-2)} \lesssim \Delta_{\min}$  when  $K \geq 3$ , and the inequality (123) holds by the fact that

$$\|(\mathbf{I}_r - \bar{D})\mathbf{S}\mathbf{C}^{-1}\|_F \lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{S}_{b:}^s - \mathbf{S}_{z(i):}^s\| \lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{A}_{b:}^s - \mathbf{A}_{z(i):}^s\|.$$

□

**Lemma 12** (Polynomial estimation error of MLE). Let  $(\hat{z}, \hat{\theta}, \hat{\mathcal{S}})$  denote the MLE in (10), and  $\hat{\mathcal{X}}$  denote the mean tensor consisting of parameter  $(\hat{z}, \hat{\theta}, \hat{\mathcal{S}})$ . We have

$$\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \lesssim \sigma^2 (r^K + Kpr),$$

with probability going to 1. When  $\Delta_{\min}^2 \gtrsim p^{-(K-1)} \log p$ , the MLE satisfies

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\theta(i))^2 \lesssim \frac{1}{r \log p}, \quad \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\hat{\theta}(i))^2 \lesssim \frac{1}{r \log p}, \quad L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p},$$

Further, we have

$$\theta(i)^2 = (1 + o(p^{-(K-2)})) \hat{\theta}(i)^2.$$

*Proof of Lemma 12.* Without loss of generality, we assume  $\sigma^2 = 1$  and identity mapping minimizes the misclustering error for MLE. For arbitrary two sets of parameters  $(z, \theta, \mathcal{S}), (z'_k, \theta'_k, \mathcal{S}') \in \mathcal{P}_b(\gamma)$  and corresponding mean tensors  $\mathcal{X}, \mathcal{X}'$ , we have

$$\text{rank}(\text{Mat}_k(\mathcal{X}) - \text{Mat}_k(\mathcal{X}')) \leq \text{rank}(\text{Mat}_k(\mathcal{X})) + \text{rank}(\text{Mat}_k(\mathcal{X}')) \leq 2r_k, \quad k \in [K].$$

Hence, we have

$$\mathcal{X} - \mathcal{X}' \in \mathcal{Q}(2r_1, \dots, 2r_K) := \{\text{Tucker tensor with rank } (2r_1, \dots, 2r_K)\}. \quad (132)$$

Then, we obtain that

$$\begin{aligned} \mathbb{P}(\|\mathcal{X} - \hat{\mathcal{X}}_{ML}\|_F \geq t) &\leq 2\mathbb{P}\left(\sup_{\mathcal{X}, \mathcal{X}' \in \mathcal{P}(r_1, \dots, r_K)} \left\langle \frac{\mathcal{X} - \mathcal{X}'}{\|\mathcal{X} - \mathcal{X}'\|_F}, \mathcal{E} \right\rangle \geq t\right) \\ &\leq 2\mathbb{P}\left(\sup_{\mathcal{T} \in \mathcal{Q}(2r_1, \dots, 2r_K) \cap \{\|\mathcal{T}\|_F=1\}} \langle \mathcal{T}, \mathcal{E} \rangle \geq t\right) \\ &\lesssim \exp(-Kpr), \end{aligned}$$

with the choice  $t \asymp \sigma\sqrt{(Kpr + r^K)}$ . Here the first inequality follows from Wang and Zeng (2019, Lemma 1), the second inequality follows from (132), and the last inequality follows from Han et al. (2022, Lemma E5).  $\square$

When  $\Delta_{\min}^2 \gtrsim p^{-(K-1)} \log p$ , we replace the vector  $\hat{x}_{\hat{z}(i)}$  and  $\hat{\mathbf{X}}$  by our MLE estimator in the proof of Theorem 4. With estimation error  $\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \lesssim (r^K + Kpr)$  and  $\Delta_{\min}^2 \gtrsim p^{-(K-1)} \log p$ , we have

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\theta(i))^2 \lesssim \frac{r^{K-1}}{\Delta_{\min}^2 p^K} \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \lesssim \frac{r^{K-2}}{p^{K-1} \Delta_{\min}^2} \lesssim \frac{1}{r \log p}, \quad \text{and} \quad L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p}.$$

Above result holds for  $\hat{\theta}(i)$  after switching the parameters  $\mathbf{X}$  with  $\hat{\mathbf{X}}$  and switch  $\boldsymbol{\theta}$  with  $\hat{\boldsymbol{\theta}}$  in the proof.

Last, notice that for all  $a \in [r]$

$$(1 - O(1)) \frac{p^2}{r^2} \|\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}}\|_F^2 \leq \left\| \sum_{\hat{z}(i)=z(i)=a} (\theta(i) \mathbf{W}_{:a}^T \mathbf{X} - \hat{\theta}(i) \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}}) \right\|_F^2 \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr,$$

where the first inequality follows from the facts that  $\ell(\hat{z}, z) \lesssim \frac{1}{\log p}$ ,  $|z^{-1}(a)| \asymp p/r$ ,

$$|z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq |\hat{z}^{-1}(a)| \leq |z^{-1}(a)| + C \frac{p}{r} \ell(\hat{z}, z),$$

$$|z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{z(i)=\hat{z}(i)=a} \theta(i) \leq |z^{-1}(a)|, \quad \text{and} \quad |\hat{z}^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{\hat{z}(i)=z(i)=a} \hat{\theta}(i) \leq |\hat{z}^{-1}(a)|.$$

Hence, for all  $i \in [p]$

$$(\theta(i) - \hat{\theta}(i))^2 \|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - O(p) \leq \|(\theta(i) - \hat{\theta}(i)) \mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - \|\hat{\theta}(i) (\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}})\|_F^2 \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr,$$

where the first inequality follows from  $\|\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}}\|_F^2 \lesssim 1/p$  and  $\hat{\theta}(i) \lesssim \frac{p}{r}$ . Notice that for all  $a \in [r]$

$$\|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 \geq \|\mathbf{S}_{a:}\|_F^2 \lambda_{\min}^{2(K-1)}(\boldsymbol{\Theta} \mathbf{M}) \gtrsim p^{K-1}.$$

The inequality indicates that  $\theta(i)^2 = (1 + o(p^{-(K-2)})) \hat{\theta}(i)^2$ .