

# Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jixin Hu<sup>1</sup> and Miaoyan Wang<sup>1</sup>

**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 [1], neuroscience [2], and computer science [3]. Tensors effectively represent the multiway data and serve as the foundation in higher-order data analysis. One data example is from multitissue multiindividual gene expression study [4], [5], where the data tensor consists of expression measurements indexed by (gene, individual, tissue) triplets. Another example is *hypergraph* network [6], [7], [8], [9] 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.

Manuscript received 18 January 2022; revised 23 December 2022; accepted 6 January 2023. The work of Miaoyan Wang was supported in part by NSF under Grant CAREER DMS-2141865, Grant DMS-1915978, Grant DMS-2023239, and Grant EF-2133740; and in part by the Wisconsin Alumni Research Foundation. An earlier version of this paper was presented in part at the 25th International Conference on Artificial Intelligence and Statistics (AISTATS). (Corresponding author: Miaoyan Wang.)

The authors are with the Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: jhu267@wisc.edu; miaoyan.wang@wisc.edu).

Communicated by R. Venkataraman, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIT.2023.3239521>.

Digital Object Identifier 10.1109/TIT.2023.3239521

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. Fig. 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) [10], which extends the usual matrix stochastic block model [11] to tensors. The matrix analysis tools, however, are suboptimal for higher-order clustering. Developing tensor tools for solving block models has received increased interest recently [10], [12], [13].

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 nonrealistic. 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 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 [14] without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity [9]. The dTBM combines a higher-order checkerboard structure with degree parameter  $\theta = (\theta(1), \dots, \theta(p))^T$  to allow heterogeneity among  $p$  nodes. Fig. 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.

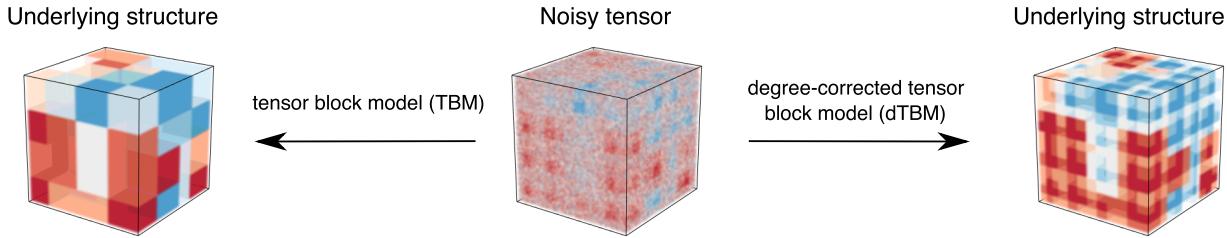


Fig. 1. Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

### 82 A. Our Contributions

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

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

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

### 103 B. Related Work

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

- 107 • *Block model for clustering.* The block model such as  
108 stochastic block model (SBM) and degree-corrected SBM  
109 has been widely used for matrix clustering problems.  
110 The theoretical properties and algorithm performance for  
111 matrix block models have been well-studied [15]; see the  
112 review paper [11] and the references therein. However,  
113 The tensor counterparts are relatively less understood.
- 114 • *Tensor block model.* The (nondegree) tensor block model  
115 (TBM) is a higher-order extension of SBM, and its  
116 statistical-computational properties are investigated in  
117 recent literatures [7], [10], [13]. Some works [16] study  
118 the TBM with sparse observations, while, others [10],  
119 [13] and our work focus on the dense regime. Extending  
120 results from nondegree to degree-corrected model  
121 is highly challenging. Our dTBM parameter space is  
122 equipped with angle-based similarity and nuisance degree  
123 parameters. The extra complexity makes the Cartesian  
124 coordinates based analysis [13] nonapplicable to our  
125 setting. Towards this goal, we have developed a new polar

126 coordinates based analysis to control the model complexity  
127 We have also developed a new angle-based iteration  
128 algorithm to achieve optimal clustering rates *without the*  
129 *need of estimating nuisance degree parameters.*

- 130 • *Degree-corrected block model.* The hypergraph  
131 degree-corrected block model (hDCBM) and its  
132 variant have been proposed in the literature [9],  
133 [17]. For this popular model, however, the optimal  
134 statistical-computational rates remain an open problem.  
135 Our main contribution is to provide a sharp statistical  
136 and computational critical phase transition in dTBM  
137 literature. In addition, our algorithm results in a faster  
138 *exponential* error rate, in contrast to the *polynomial*  
139 rate in [9]. The original hDCBM [9] is designed for  
140 binary observations only, and we extend the model to  
141 both continuous and binary observations. We believe  
142 our results are novel and helpful to the community. See  
143 Fig. 2 for overview of our results.
- 144 • *Global-to-local algorithm strategy.* Our methods gen-  
145 eralize the recent global-to-local strategy for matrix  
146 learning [15], [18], [19] to tensors [13], [16], [20].  
147 Despite the conceptual similarity, we address several  
148 fundamental challenges associated with this nonconvex,  
149 noncontinuous problem. We show the insufficiency of  
150 the conventional tensor HOSVD [21], and we develop  
151 a weighted higher-order initialization that relaxes the  
152 singular-value gap separation condition. Furthermore,  
153 our local iteration leverages the angle-based clustering  
154 in order to avoid explicit estimation of degree heteroge-  
155 neity. Our bounds reveal the interesting interplay between  
156 the computational and statistical errors. We show that  
157 our final estimate *provably* achieves the exact clustering  
158 within only polynomial-time complexity.

### 159 C. Notation

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

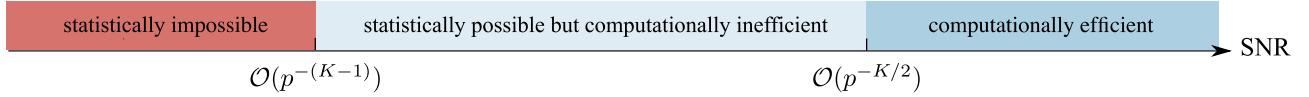


Fig. 2. SNR thresholds for statistical and computational limits in order- $K$  dTBM with dimension  $(p, \dots, p)$  and  $K \geq 2$ . The SNR gap between statistical possibility and computational efficiency exists only for tensors with  $K \geq 3$ .

TABLE I  
COMPARISON BETWEEN PREVIOUS METHODS WITH OUR METHOD

|                                       | Gao et al. (2018)[15] | Ahn et al. (2018)[16]   | Han et al. (2022)[13] | Ghoshdastidar et al. (2019)[7] | Ke et al. (2019)[9] | Ours             |
|---------------------------------------|-----------------------|-------------------------|-----------------------|--------------------------------|---------------------|------------------|
| Allow tensors of arbitrary order      | ✗                     | ✓                       | ✓                     | ✓                              | ✓                   | ✓                |
| Allow degree heterogeneity            | ✓                     | ✗                       | ✗                     | ✓                              | ✓                   | ✓                |
| Singular-value gap-free clustering    | ✓                     | ✓                       | ✓                     | ✗                              | ✗                   | ✓                |
| Misclustering rate (for order $K^1$ ) | -                     | $p^{-(K-1)\alpha^{-1}}$ | $\exp(-p^{K/2})$      | $p^{-1}$                       | $p^{-2}$            | $\exp(-p^{K/2})$ |
| Consider sparse observation           | ✗                     | ✓                       | ✗                     | ✗                              | ✗                   | ✗                |

<sup>1</sup>We list the result for order- $K$  tensors with  $K \geq 3$  and general number of communities  $r = \mathcal{O}(1)$ .

<sup>2</sup>The parameter  $\alpha = f(p) > 0$  denotes the sparsity level which is some function of dimension  $p$ .

171 *a.* For two vector  $\mathbf{a}, \mathbf{b}$  of the same dimension, we denote the  
172 angle between  $\mathbf{a}, \mathbf{b}$  by

$$173 \cos(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|},$$

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

177 Let  $\mathcal{Y} \in \mathbb{R}^{p_1 \times \dots \times p_K}$  be an order- $K$   $(p_1, \dots, p_K)$ -  
178 dimensional tensor. We use  $\mathcal{Y}(i_1, \dots, i_K)$  to denote the  
179  $(i_1, \dots, i_K)$ -th entry of  $\mathcal{Y}$ . The multilinear multiplication of a  
180 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  
181 an order- $K$   $(p_1, \dots, p_K)$ -dimensional tensor  $\mathcal{X}$ , denoted

$$182 \mathcal{X} = \mathcal{S} \times_1 \mathbf{M}_1 \times \dots \times_K \mathbf{M}_K,$$

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

$$184 \mathcal{X}(i_1, \dots, i_K) \\ 185 = \sum_{(j_1, \dots, j_K)} \mathcal{S}(j_1, \dots, j_K) \mathbf{M}_1(i_1, j_1) \dots \mathbf{M}_K(i_K, j_K).$$

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

#### 200 D. Organization

201 The rest of this paper is organized as follows. Section II  
202 introduces the degree-corrected tensor block model (dTBM)  
203 with three motivating examples and presents the identifiability  
204 of dTBM under the angle gap condition. We show the phase  
205 transition and the existence of statistical-computational gaps  
206 for the higher-order dTBM in Section III. In Section IV,  
207 we provide a polynomial-time two-stage algorithm with

208 misclustering rate guarantees. Extension to Bernoulli models  
209 is also presented. In Section V, we compare our work with  
210 nondegree tensor block models. Numerical studies including  
211 the simulation, comparison with other methods, and two real  
212 dataset analyses are in Sections VI-VII. The main technical  
213 ideas we develop for addressing main theorems are provided  
214 in Section VIII. Detailed proofs and extra theoretical results  
215 are provided in Appendix.

## II. MODEL FORMULATION AND MOTIVATIONS

### A. Degree-Corrected Tensor Block Model

216 Suppose that we have an order- $K$  data tensor  $\mathcal{Y} \in \mathbb{R}^{p \times \dots \times p}$ .  
217 Assume that there exist  $r \geq 1$  disjoint communities among the  
218  $p$  nodes. We represent the community assignment by a function  
219  $z: [p] \mapsto [r]$ , where  $z(i) = a$  for  $i$ -th node that belongs to  
220 the  $a$ -th community. Then,  $z^{-1}(a) = \{i \in [p]: z(i) = a\}$   
221 denotes the set of nodes that belong to the  $a$ -th community,  
222 and  $|z^{-1}(a)|$  denotes the number of nodes in the  $a$ -th community.  
223 Let  $\boldsymbol{\theta} = (\theta(1), \dots, \theta(p))^T$  denote the degree heterogeneity for  
224  $p$  nodes. We consider the order- $K$  dTBM [7], [9],  
225

$$226 \mathcal{Y}(i_1, \dots, i_K) = \mathcal{S}(z(i_1), \dots, z(i_K)) \prod_{k=1}^K \theta_{i_k} + \mathcal{E}(i_1, \dots, i_K),$$

227 where  $\mathcal{S} \in \mathbb{R}^{r \times \dots \times r}$  is an order- $K$  tensor collecting the block  
228 means among communities, and  $\mathcal{E} \in \mathbb{R}^{p \times \dots \times p}$  is a noise tensor  
229 consisting of independent zero-mean sub-Gaussian entries  
230 with variance bounded by  $\sigma^2$ . The unknown parameters are  $z$ ,  
231  $S$ , and  $\boldsymbol{\theta}$ . The dTBM can be equivalently written in a compact  
232 form of tensor-matrix product:  
233

$$234 \mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \dots \times_K \boldsymbol{\Theta} \mathbf{M}, \quad (1)$$

235 where  $\boldsymbol{\Theta} = \text{diag}(\theta(1), \dots, \theta(p)) \in \mathbb{R}^{p \times p}$  is a diagonal matrix,  
236  $\mathbf{M} \in \{0, 1\}^{p \times r}$  is the membership matrix associated with  
237 community assignment  $z$  such that  $\mathbf{M}(i, j) = \mathbb{1}\{z(i) = j\}$ .  
238 By definition, each row of  $\mathbf{M}$  has one copy of 1’s and  
239 0’s elsewhere. Note that the discrete nature of  $\mathbf{M}$  renders  
240 our model (1) more challenging than Tucker decomposition.  
241 We call a tensor  $\mathcal{Y}$  an  $r$ -block tensor with degree  $\boldsymbol{\theta}$  if  $\mathcal{Y}$  admits  
242 dTBM (1) and let  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  denote the mean tensor. The goal  
243 of clustering is to estimate  $z$  from a single noisy tensor  $\mathcal{Y}$ .  
244 We are particularly interested in the high-dimensional regime  
245 where  $p$  grows whereas  $r = \mathcal{O}(1)$ .  
246

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  $(\Theta, M, z)$ ; the noise tensor  $\mathcal{E}$  and the data tensor  $\mathcal{Y}$  are still possibly asymmetric. In general, we allow asymmetric mean tensors with  $\{(\Theta_k, M_k, z_k)\}_{k=1}^K$ , one for each mode. The extension can be found in Appendix B.

### B. Motivating Examples

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

1) *Tensor Block Model*: Consider the model (1). Let  $\theta(i) = 1$  for all  $i \in [p]$ . The model (1) reduces to the tensor block model, which is widely used in previous clustering algorithms [10], [12], [13]. The theoretical results in TBM serve as benchmarks for dTBM.

2) *Community Detection in Hypergraphs*: The hypergraph network is a powerful tool to represent the complex entity relations with higher-order interactions [9]. 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 that there exist  $r$  disjoint communities among  $p$  nodes, and the connection probabilities depend on the community assignments and node-specific parameters. Then, the equation (1) models  $\mathbb{E}\mathcal{Y}$  with unknown degree heterogeneity  $\theta$  and sub-Gaussianity parameter  $\sigma^2 = 1/4$ .

3) *Multilayer Weighted Network*: Multilayer weighted network data consists of multiple networks over the same set of nodes. One representative example is the brain connectome data [22]. The multilayer 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 that there exist  $r$  disjoint communities among  $p$  nodes and  $r_l$  disjoint communities among the  $L$  layers. The multilayer network community detection is modeled by the general asymmetric dTBM model (1)

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

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

4) *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 multitissue gene expression dataset, which records different gene expressions 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 that 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 (1)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \Theta_1 M_1 \times_2 \Theta_2 M_2 \times_3 \Theta_3 M_3,$$

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

*Remark 1 (Comparison With Nondegree Models)*: Our dTBM uses fewer block parameters than TBM. In particular, every nondegree  $r_1$ -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 (1). For ease of notation, we focus on symmetric tensors; the extension to asymmetric tensors are similar. We use  $\mathcal{P}$  to denote the following parameter space for  $(z, \mathcal{S}, \theta)$ ,

$$\begin{aligned} \mathcal{P} = & \left\{ (z, \mathcal{S}, \theta) : \theta \in \mathbb{R}_+^p, \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, \right. \\ & \left. c_3 \leq \|\text{Mat}(\mathcal{S})_{:a}\| \leq c_4, \|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r] \right\} \end{aligned} \quad (2)$$

where  $c_i > 0$ 's are universal constants. We briefly describe the rationale of the constraints in (2). First, the entrywise positivity constraint on  $\theta \in \mathbb{R}_+^p$  is imposed to avoid sign ambiguity between entries in  $\theta_{z^{-1}(a)}$  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 (1); see Example 1 below. Second, recall that the quantity  $|z^{-1}(a)|$  denotes the number of nodes in the  $a$ -th community. The constants  $c_1, c_2$  in the  $|z^{-1}(a)|$  bounds assume the roughly balanced size across  $r$  communities. Third, the constant  $c_3$  requires that all slides in  $\mathcal{S}$  have nondegenerate norm. Particularly, the lower bound  $c_3$  excludes the purely zero slide to avoid trivial nonidentifiability of model (1); see Example 2 below. The upper bound  $c_4$  is a technical constraint to avoid the slides with diverging norm as dimension grows. Lastly, the  $\ell_1$  normalization  $\|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$  is imposed to avoid the scalar ambiguity between  $\theta_{z^{-1}(a)}$  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.

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

$$357 \quad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \boldsymbol{\Theta} \mathbf{M} \times_3 \boldsymbol{\Theta} \mathbf{M},$$

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

$$361 \quad \mathcal{X}_{::1} = \mathcal{X}_{::2} = -\mathcal{X}_{::3} = -\mathcal{X}_{::4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

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

$$365 \quad \mathcal{X} = \mathcal{S}' \times_1 \boldsymbol{\Theta}' \mathbf{M}' \times_2 \boldsymbol{\Theta}' \mathbf{M}' \times_3 \boldsymbol{\Theta}' \mathbf{M}',$$

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

$$370 \quad \mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{M}' = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

371 The triplet  $(z', \mathcal{S}', \boldsymbol{\theta}')$  lies in our parameter space (2). In general, we can always reparameterize an  $r$ -block tensor with mixed-signed degree using a  $2r$ -block 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.

377 *Example 2 (Nonidentifiability With Purely Zero Core Slice):*  
 378 Consider an order-2 dTBM with core tensor  $\mathcal{S} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$   
 379 degree matrices  $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \text{diag}(1, 1, 1, 1)$ , and mean tensor

$$380 \quad \mathcal{X} = \boldsymbol{\Theta}_1 \mathbf{M} \mathcal{S} \mathbf{M}^T \boldsymbol{\Theta}_2, \quad \text{with } \mathbf{M} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

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

383 We now provide the identifiability conditions for our model  
 384 before estimation procedures. When  $r = 1$ , the decomposition  
 385 in (1) is always unique (up to cluster label permutation) in  $\mathcal{P}$ ,  
 386 because dTBM is equivalent to the rank-1 tensor family under  
 387 this case. When  $r \geq 2$ , the Tucker rank of signal tensor  $\mathbb{E}\mathcal{Y}$   
 388 in (1) is bounded by, but not necessarily equal to, the number  
 389 of blocks  $r$  [10]. Therefore, one can not apply the classical  
 390 identifiability conditions for low-rank tensors to dTBM. Here,  
 391 we introduce a key separation condition on the core tensor.

392 *Assumption 1 (Angle Gap):* Let  $\mathbf{S} = \text{Mat}(\mathcal{S})$ . Assume that  
 393 the minimal gap between normalized rows of  $\mathbf{S}$  is bounded

away from zero; i.e.,

$$394 \quad \Delta_{\min} := \min_{a \neq b \in [r]} \left\| \frac{\mathbf{S}_{a:}}{\|\mathbf{S}_{a:}\|} - \frac{\mathbf{S}_{b:}}{\|\mathbf{S}_{b:}\|} \right\| > 0, \quad \text{for } r \geq 2. \quad (3)$$

We make the convention  $\Delta_{\min} = 1$  for  $r = 1$ . Equivalently, (3) 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 nonredundancy among clusters measured by angle separation. The denominators involved in definition (3) are well posed because of the lower bound on  $\|\mathbf{S}_{a:}\|$  in (2).

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  $r \geq 2$  and  $K \geq 2$ . The parameterization (1) 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),

$$420 \quad \text{SNR} := \Delta_{\min}^2 / \sigma^2 = p^\gamma, \quad (4)$$

422 with varying  $\gamma \in \mathbb{R}$  that quantifies different regimes of  
 423 interest. We call  $\gamma$  the *signal exponent*. Intuitively, a larger  
 424 SNR, or equivalently a larger  $\gamma$ , benefits the clustering in the  
 425 presence of noise. With quantification (4), we consider the  
 426 following parameter space,

$$427 \quad \mathcal{P}(\gamma) = \mathcal{P} \cap \{\mathcal{S} \text{ satisfies SNR condition (4) with } \gamma\}. \quad (5)$$

428 The 1-block dTBM does not belong to the space  $\mathcal{P}(\gamma)$  when  
 429  $\gamma < 0$ , due to the convention in Assumption 1. Our goal is to  
 430 characterize the clustering accuracy with respect to  $\gamma$  under  
 431 the space  $\mathcal{P}(\gamma)$ .

432 In our algorithmic development, we often refer to the  
 433 regime of balanced degree heterogeneity. We call the degree  
 434  $\boldsymbol{\theta}$  *balanced* if

$$435 \quad \min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|. \quad (6)$$

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

440 *Lemma 1 (Angle Gaps in  $\mathcal{X}$  and  $\mathcal{S}$ ):* Consider the dTBM  
 441 model (1) under the parameter space  $\mathcal{P}$  in (2) with  $r \geq 2$ .  
 442 Suppose  $\boldsymbol{\theta}$  is balanced satisfying (6) and  $\min_{i \in [p]} \theta(i) \geq c$

TABLE II  
PARAMETER SPACE COMPARISON BETWEEN PREVIOUS WORK WITH OUR ASSUMPTION

| Assumptions in parameter space | Gao et al. (2018)[15]         | Han et al. (2022)[13] | Ke et al. (2019)[9] | 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 |

from some constant  $c > 0$ . Then, as  $p \rightarrow \infty$ , 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 (6). 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

$$\mathbf{S} = \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix}, \quad (7)$$

and  $\boldsymbol{\theta}$  such that  $\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2 = p^m \|\boldsymbol{\theta}_{z^{-1}(2)}\|^2$ , where  $m \in [-1, 1]$  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

$$\Delta_{\mathbf{X}}^2 := \min_{i,j \in [p], z(i) \neq z(j)} \left\| \frac{\mathbf{X}_{i:}}{\|\mathbf{X}_{i:}\|} - \frac{\mathbf{X}_{j:}}{\|\mathbf{X}_{j:}\|} \right\|, \quad (8)$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . Take  $a = p^{-1/4}$  in the model setup (7). We have

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

Based on the Theorem 2 in Section III, the dTBM is impossible to solve when  $\Delta_{\mathbf{X}}^2 \lesssim p^{-1}$  even though  $\Delta_{\min}^2 \asymp p^{-1/2}$ ; 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 (6) 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 [15] that restricts the highest degree in the sublinear 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 [9], 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 degree balance across communities. In [15], 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 (2). Define the misclustering error by

$$\ell(\hat{z}, z) = \frac{1}{p} \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq \pi \circ z(i)\}, \quad (48)$$

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 infimum 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 (2) 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  $(\hat{z}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}, \hat{\boldsymbol{\theta}}_{\text{MLE}})$ , over the estimation space  $\mathcal{P}$ , where

$$(\hat{z}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}, \hat{\boldsymbol{\theta}}_{\text{MLE}}) = \arg \min_{(z, \mathcal{S}, \boldsymbol{\theta}) \in \mathcal{P}} \|\mathcal{Y} - \mathcal{X}(z, \mathcal{S}, \boldsymbol{\theta})\|_F^2. \quad (49)$$

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

**• Impossibility.** Assume  $p \rightarrow \infty$  and  $2 \leq r \lesssim p^{1/3}$ . Let  $\mathcal{P}_{\mathcal{S}}(\gamma) := \{\mathcal{S} : c_3 \leq \|\text{Mat}(\mathcal{S})_{:a}\| \leq c_4, a \in [r]\} \cap \{\mathcal{S} : \Delta_{\min}^2 = p^\gamma\}$  denote the space for valid  $\mathcal{S}$  satisfying SNR condition (4), and  $\mathcal{P}_{z, \boldsymbol{\theta}} := \{\boldsymbol{\theta} \in \mathbb{R}_+^p, \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, \|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r]\}$  denote the space for valid  $(z, \boldsymbol{\theta})$ , where  $c_1, c_2, c_3, c_4$  are the constants in parameter space (2). If the signal exponent satisfies  $\gamma < -(K-1)$ , then, for any true core tensor  $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}(\gamma)$ , no estimator  $\hat{z}_{\text{stat}}$  achieves exact recovery in expectation;

526 that is, when  $\gamma < -(K - 1)$ , we have

$$\liminf_{p \rightarrow \infty} \inf_{\mathcal{S} \in \mathcal{P}_S(\gamma)} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \theta) \in \mathcal{P}_{z, \theta}} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1. \quad (10)$$

528 Further, we define the parameter space  $\mathcal{P}'(\gamma') := \mathcal{P} \cap$   
529  $\{\Delta_X^2 = p^{\gamma'}\}$ , where  $\Delta_X^2$  is the mean tensor minimal gap  
530 in (8). When  $\gamma' < -(K - 1)$ , we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \mathcal{S}, \theta) \in \mathcal{P}'(\gamma')} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1.$$

- 532 • **MLE achievability.** Suppose that the signal exponent  
533 satisfies  $\gamma > -(K - 1) + c_0$  for an arbitrary constant  
534  $c_0 > 0$ . Furthermore, assume that  $\theta$  is balanced and  
535  $\min_{i \in [p]} \theta(i) \geq c$  from some constant  $c > 0$ . Then, when  
536  $p \rightarrow \infty$ , for fixed  $r \geq 1$ , the MLE in (9) achieves exact  
537 recovery in high probability; that is,

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

539 with probability going to 1.

540 The proofs for the two parts in Theorem 2 are in the  
541 Appendices B-D and B-G, respectively. The first part of Theorem 2 demonstrates impossibility of exact recovery whenever  
542 the core tensor  $\mathcal{S}$  satisfies SNR condition (4) with exponent  
543  $\gamma < -(K - 1)$ . The proof is information-theoretical, and  
544 therefore the results apply to all statistical estimators, including  
545 but not limited to MLE and trace maximization [6]. The  
546 minimax bound (10) indicates the worst case impossibility for  
547 a particular core tensor  $\mathcal{S}$  with signal exponent  $\gamma < -(K - 1)$ ;  
548 i.e., under the assumptions of Theorem 2, when  $\gamma < -(K - 1)$ ,  
549 we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \mathcal{S}, \theta) \in \mathcal{P}(\gamma)} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1.$$

552 Such worst case impossibility is studied in related works [13],  
553 [15] while our lower bound (10) provides a stronger impossibility  
554 statement for arbitrary core tensors with weak signals.  
555 The second part of Theorem 2 shows the exact recovery of  
556 MLE when  $\gamma > -(K - 1) + c_0$  for an arbitrary constant  
557  $c_0 > 0$ . Combining the impossibility and achievability results,  
558 we conclude that the boundary  $\gamma_{\text{stat}} := -(K - 1)$  is the critical  
559 value for statistical performance of dTBM with respect to our  
560 SNR.

### 561 C. Computational Critical Value

562 The computational critical value means the minimal SNR  
563 required for exact recovery with *polynomial-time* computa-  
564 tional cost. An important ingredient to establish the computa-  
565 tional limits is the *hypergraphic planted clique (HPC) conjec-*  
566 *ture* [23], [24]. The HPC conjecture indicates the impossibility  
567 of fully recovering the planted cliques with polynomial-time  
568 algorithm when the clique size is less than the number of ver-  
569 tices in the hypergraph. The formal statement of HPC detection  
570 conjecture is provided in Definition 1 and Conjecture 1 as  
571 follows.

572 *Definition 1 (Hypergraphic Planted Clique (HPC) Detec-*  
573 *tion):* Consider an order- $K$  hypergraph  $H = (V, E)$  where

574  $V = [p]$  collects vertices and  $E$  collects all the order- $K$   
575 edges. Let  $\mathcal{H}_k(p, 1/2)$  denote the Erdős-Rényi  $K$ -hypergraph  
576 where the edge  $(i_1, \dots, i_K)$  belongs to  $E$  with probability  
577  $1/2$ . Further, we let  $\mathcal{H}_K(p, 1/2, \kappa)$  denote the hypergraph  
578 with planted cliques of size  $\kappa$ . Specifically, we generate a  
579 hypergraph from  $\mathcal{H}_k(p, 1/2)$ , pick  $\kappa$  vertices uniformly from  
580  $[p]$ , denoted  $K$ , and then connect all the hyperedges with  
581 vertices in  $K$ . Note that the clique size  $\kappa$  can be a function of  
582  $p$ , denoted  $\kappa_p$ . The order- $K$  HPC detection aims to identify  
583 whether there exists a planted clique hidden in an Erdős-  
584 Rényi  $K$ -hypergraph. The HPC detection is formulated as the  
585 following hypothesis testing problem

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

587 *Conjecture 1 (HPC Conjecture):* Consider the HPC detec-  
588 tion problem in Definition 1 with  $K \geq 2$ . Suppose the  
589 sequence  $\{\kappa_p\}$  such that  $\limsup_{p \rightarrow \infty} \log \kappa_p / \log \sqrt{p} \leq (1 - \tau)$   
590 for any  $\tau > 0$ . Then, for every sequence of polynomial-time  
591 test  $\{\varphi_p\} : H \mapsto \{0, 1\}$  we have

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

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

596 *Theorem 3 (Computational Critical Value):* Consider gen-  
597 eral Gaussian dTBMs under the parameter space  $\mathcal{P}$  with  
598  $K \geq 2$ . Then, we have the following computational phase  
599 transition.

- 600 • **Impossibility.** Assume HPC conjecture holds and  $r \geq$   
601 2. If the signal exponent satisfies  $\gamma < -K/2$ , then,  
602 no *polynomial-time estimator*  $\hat{z}_{\text{comp}}$  achieves exact recov-  
603 ery in expectation as  $p \rightarrow \infty$ ; that is, when  $\gamma < -K/2$ ,  
604 we have

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

- 605 • **Polynomial-time algorithm achievability.** Suppose that  
606 we have fixed  $r \geq 1$ , and the signal exponent satisfies  
607  $\gamma > -K/2 + c_0$  for an arbitrary constant  $c_0 > 0$ .  
608 Furthermore, assume that the degree  $\theta$  is balanced, lower  
609 bounded in that  $\min_{i \in [p]} \theta_i \geq c$  for some constant  $c > 0$ ,  
610 and satisfies the locally linear stability in Definition 2 in  
611 the neighborhood  $\mathcal{N}(z, \varepsilon)$  for all  $\varepsilon \leq E_0$  and some  $E_0 \gtrsim$   
612  $\log^{-1} p$ . Then, as  $p \rightarrow \infty$ , there exists a polynomial-time  
613 algorithm  $\hat{z}_{\text{poly}}$  that achieves exact recovery in high prob-  
614 ability; that is,

$$\ell(\hat{z}_{\text{poly}}, z) \lesssim \text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right) \rightarrow 0,$$

617 with probability going to 1.

618 The proofs for the two parts in Theorem 3 are in the  
619 Appendices B-E and B-G, respectively. The first part of  
620 Theorem 3 indicates the impossibility of exact recovery by  
621 polynomial-time algorithms when  $\gamma < -K/2$ , and the sec-  
622 ond part shows the existence of such algorithm when  $\gamma >$   
623  $-K/2 + c_0$  for an arbitrary constant  $c_0 > 0$  under extra  
624 technical assumptions. In Section IV, we will present an

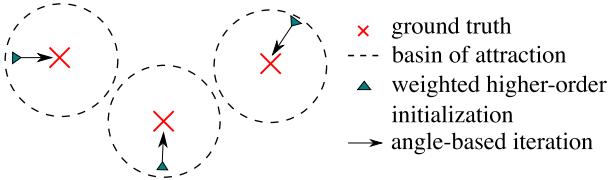


Fig. 3. Illustration of our global-to-local algorithm.

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. Fig. 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

$$Y = \Theta M S + E,$$

where  $Y \in \mathbb{R}^{p \times d}$  collects  $n$  data points in  $\mathbb{R}^d$ ,  $S \in \mathbb{R}^{r \times d}$  collects the  $d$ -dimensional centroids for  $r$  clusters, and  $\Theta \in \mathbb{R}^{p \times p}$ ,  $M \in \{0, 1\}^{p \times r}$ ,  $E \in \mathbb{R}^{p \times d}$  have the same meaning as in dTBM. [25] implies that polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Therefore, we conclude that the statistical-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 Fig. 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 dTBMs with symmetric mean tensor and independent sub-Gaussian noises such as Gaussian and uniform observations. The extensions for Bernoulli observations and other practical issues are in Sections IV-C and IV-D.

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

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

where the superscript  $\cdot^s$  denotes the normalized vector; i.e.,  $\mathbf{a}^s := \mathbf{a}/\|\mathbf{a}\|$  if  $\mathbf{a} \neq 0$  and  $\mathbf{a}^s = 0$  if  $\mathbf{a} = 0$  for any vector  $\mathbf{a}$ . 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 (Relationship Between Misclustering Error and Loss):* Consider the dTBM under the parameter space  $\mathcal{P}$ . Suppose  $\min_{i \in [p]} \theta(i) > c$  for some constant  $c > 0$ . 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 (1), for all  $i \in [p]$ , we have

$$\theta(i)^{-1} \mathbf{X}_{i:} = [\text{Mat}(\mathcal{S} \times_2 \Theta \mathbf{M} \times_3 \Theta \mathbf{M})]_{z(i)} : . \quad (688)$$

This implies that, all node  $i$  belonging to the  $a$ -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 the denoising step and the 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 [21] 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.,

$$\hat{\mathbf{U}}_1 = \text{SVD}_r(\text{Mat}_1(\mathcal{Y} \times_2 \mathbf{U}_{\text{pre},2} \mathbf{U}_{\text{pre},2}^T \times_3 \mathbf{U}_{\text{pre},3} \mathbf{U}_{\text{pre},3}^T)) . \quad (702)$$

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

$$\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\mathbf{U}}_1 \hat{\mathbf{U}}_1^T \times_2 \hat{\mathbf{U}}_2 \hat{\mathbf{U}}_2^T \times_3 \hat{\mathbf{U}}_3 \hat{\mathbf{U}}_3^T . \quad (705)$$

The double projection improves usual matrix spectral methods in order to alleviate the noise effects for  $K \geq 3$  [13].

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 [9], we apply the clustering on the unfolded tensor  $\hat{\mathbf{X}}$  rather than on the factors  $\hat{\mathbf{U}}_k$ . This strategy relaxes the singular-value gap condition [13], [15]. We assign

718 degenerate rows with purely zero entries to an arbitrarily  
 719 random cluster; these nodes are negligible in high-dimensions  
 720 because of the lower bound on  $\|\text{Mat}(\mathcal{S})_{a:}\|$  in (2). The final  
 721 result gives the initial cluster assignment  $z^{(0)}$ . Full procedures  
 722 for clustering are provided in Sub-algorithm 1.

723 We now establish the misclustering error rate of initialization.  
 724

725 *Theorem 4 (Error for Weighted Higher-Order Initialization):*  
 726 Consider the general sub-Gaussian dTBM with fixed  $r \geq 1$ ,  
 727  $K \geq 2$ , i.i.d. noise under the parameter space  $\mathcal{P}$ , and  
 728 Assumption 1. Assume  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  
 729  $c > 0$ . Let  $\Delta_X$  denote the minimal gap in mean tensor defined  
 730 in (8), and let  $z_k^{(0)}$  denote the output of Sub-algorithm 1.  
 731 With probability going to 1, as  $p \rightarrow \infty$ , we have

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

732 Further, assume that  $\boldsymbol{\theta}$  is balanced as (6). We have

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

733 with probability going to 1 as  $p \rightarrow \infty$ .

734 *Remark 5 (Comparison to Previous Results):* For fixed  
 735 SNR, our initialization error rate with  $K = 2$  agrees with  
 736 the initialization error rate  $\mathcal{O}(p^{-1})$  in matrix models [15].  
 737 Furthermore, in the special case of nondegree TBMs with  
 738  $\boldsymbol{\theta} = \mathbf{1}_p$ , we achieve the same initial misclustering error  
 739  $\mathcal{O}(p^{-K/2})$  as in nondegree models [13]. Theorem 4 implies  
 740 the advantage of our algorithm in achieving both accuracy  
 741 and model flexibility.

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

747 *Remark 7 (Singular-Value Gap-Free Clustering):* Note  
 748 that our clustering directly applies to the estimated mean  
 749 tensor  $\hat{\mathcal{X}}$  rather than the leading tensor factors  $\hat{\mathbf{U}}_k$ .  
 750 Applying clustering to the tensor factors suffers from the  
 751 nonidentifiability issue due to the infinitely many orthogonal  
 752 rotations when the number of blocks  $r \geq 3$  in the absence  
 753 of singular-value gaps. Such ambiguity causes the trouble  
 754 for effective clustering [26]. In contrast, our initialization  
 755 algorithm applies the clustering to the overall mean tensor  $\hat{\mathcal{X}}$ .  
 756 This strategy avoids the nonidentifiability issue regardless of  
 757 the number of blocks and singular-value gaps.

## 760 B. Angle-Based Iteration

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

770 algorithm estimates the assignment  $z$  by minimizing the angle  
 771 between data vectors and centroids; i.e.,  
 772

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

773 The classical Euclidean-distance based clustering [13] fails  
 774 to recover  $z$  in the presence of degree heterogeneity, even  
 775 under noiseless case. In contrast, the proposed angle-based  
 776  $k$ -means algorithm achieves accurate recovery without the  
 777 explicit estimation of  $\boldsymbol{\theta}$ .

778 Our Sub-algorithm 2 shares the same spirit as in the angle-  
 779 based  $k$ -means. We still take the order-3 tensor for illustration.  
 780 Specifically, Sub-algorithm 2 updates estimated core tensor  
 781 and cluster assignment in each iteration. We use superscript  
 782  $(t)$  to denote the estimate from the  $t$ -th iteration, where  $t =$   
 783  $1, 2, \dots$ . For core tensor, we consider the following update  
 784 strategy

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

785 Intuitively,  $\mathcal{S}^{(t)}$  becomes closer to the true core  $\mathcal{S}$  as  $z_k^{(t)}$  is  
 786 more precise. For cluster assignment, we first aggregate the  
 787 slices of  $\mathcal{Y}$  and obtain the reduced tensor  $\mathcal{Y}_1^d \in \mathbb{R}^{p \times r \times r}$  on  
 788 the first mode with given  $z_k^{(t)}$ , where

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

789 Similarly, we also obtain  $\mathcal{Y}_2^d, \mathcal{Y}_3^d$ . We use  $\mathbf{Y}_k^d$  and  $\mathbf{S}_k^{(t)}$  to  
 790 denote the  $\text{Mat}_k(\mathcal{Y}^d)$  and  $\text{Mat}_k(\mathcal{S}^{(t)})$ . The rows  $\mathbf{Y}_{k,i}^d$  and  
 791  $\mathbf{S}_{k,a}^{(t)}$  correspond to the  $\mathbf{x}_i$  and  $\mathbf{s}_a$  in the one-dimensional  
 792 clustering (13). Then, we obtain the updated assignment by

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

793 provided that  $\mathbf{S}_{k,a}^{(t)}$  is a nonzero vector. Otherwise, if  $\mathbf{S}_{k,a}^{(t)}$  is  
 794 a zero vector, then we make the convention to assign  $z_k^{(t+1)}(i)$   
 795 randomly in  $[r]$ . Full procedures for our angle-based iteration  
 796 are described in Sub-algorithm 2.

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

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

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

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

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

819 Roughly speaking, the vector  $\mathbf{p}(\bar{z})$  represents the raw cluster  
 820 sizes, and  $\mathbf{p}_\theta(\bar{z})$  represents the relative cluster sizes weighted  
 821 by degrees. The local stability holds trivially for  $\varepsilon = 0$  based  
 822 on the construction of parameter space (2). The condition (14)  
 823 controls the impact of node degree to the  $\mathbf{p}_\theta(\cdot)$  with respect  
 824 to the misclustering rate  $\varepsilon$  and angle gap. Intuitively, the  
 825 condition (14) controls the skewness of degree so that the  
 826 angle between raw cluster size and degree-weighted cluster  
 827

**Algorithm 1** 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 matrices  $\mathbf{U}_{\text{pre},k} = \text{SVD}_r(\text{Mat}_k(\mathcal{Y}))$ ,  $k \in [K]$  and the  $(K-1)$ -mode projections

$$\mathcal{X}_{\text{pre},k} = \mathcal{Y} \times_1 \mathbf{U}_{\text{pre},1} \mathbf{U}_{\text{pre},1}^T \times_2 \dots \times_{k-1} \mathbf{U}_{\text{pre},k-1} \mathbf{U}_{\text{pre},k-1}^T \times_{k+1} \mathbf{U}_{\text{pre},k+1} \mathbf{U}_{\text{pre},k+1}^T \times_{k+2} \dots \times_K \mathbf{U}_{\text{pre},K} \mathbf{U}_{\text{pre},K}^T.$$

2: Compute factor matrices  $\hat{\mathbf{U}}_k = \text{SVD}_r(\text{Mat}_k(\mathcal{X}_{\text{pre},k}))$ ,  $k \in [K]$  and the 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  $\{\hat{x}_j\}_{j \in [r]}$  using weighted  $k$ -means, such that

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

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 the reduced tensor  $\mathcal{Y}_k^d \in \mathbb{R}^{r \times \dots \times r \times p \times r \times \dots \times r}$  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  $J_0 = \{i \in [p] : \|\mathbf{Y}_{i:\}\| = 0\}$ . Set  $z_k^{(t+1)}(i)$  randomly in  $[r]$  for  $i \in J_0$ .

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

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

14: **end for**

15: **end for**

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

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 fixed  $r \geq 1$ ,  $K \geq 2$ , independent noise under the parameter space  $\mathcal{P}$ , and Assumption 1. Assume that the locally linear stability of degree holds in the neighborhood  $\mathcal{N}(z, \varepsilon)$  for all  $\varepsilon \leq E_0$  and some  $E_0 \gtrsim \log^{-1} p$ . 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 the  $k$ -th mode. Suppose  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$ , the SNR  $\geq \tilde{C} p^{-(K-1)} \log p$  for some sufficiently large positive 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 as  $p \rightarrow \infty$ , 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 all parameter assumptions and the results in Theorems 4 and 5, with probability going to 1 as  $p \rightarrow \infty$ , our estimate  $z_k^{(T)}$  achieves exact recovery within

846 polynomial iterations; more precisely,

$$847 \quad z_k^{(T)} = \pi_k \circ z, \quad \text{for all } T \gtrsim \log_{1/\rho} p \text{ and } k \in [K].$$

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

849 Therefore, our combined algorithm is *computationally efficient*  
850 as long as  $\text{SNR} \gtrsim p^{-K/2} \log p$ . Note that, ignoring  
851 the logarithmic term, the minimal SNR requirement,  $p^{-K/2}$ ,  
852 coincides with the computational critical value in Theorem 3.  
853 Therefore, our algorithm is optimal regarding the signal  
854 requirement and lies in the sharpest *computationally efficient*  
855 regime in Fig. 2.

### 856 C. Extension to Bernoulli Observations

857 Bernoulli or network observations are common in multiple  
858 fields. Our iteration Theorem 5 holds for Bernoulli models,  
859 but our initialization Theorem 4 does not. Moreover, our  
860 current dTBM is insufficient to address sparsity with decaying  
861 mean tensor. Here, we provide extra discussions for Bernoulli  
862 initialization and strategies under sparse settings.

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

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

$$872 \quad \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \quad (16)$$

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

877 One possible remedy is to apply singular value decom-  
878 position to the *square unfolding* [27],  $\text{Mat}_{sq}(\cdot)$ , of Bernoulli  
879 tensor  $\mathcal{Y} \in \{0, 1\}^{p_1 \times \dots \times p_K}$ . Specifically, the square  
880 matricization  $\text{Mat}_{sq}(\mathcal{Y}) \in \{0, 1\}^{p^{\lfloor K/2 \rfloor} \times p^{\lceil K/2 \rceil}}$  has entries  
881  $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2) = \mathcal{Y}(i_1, \dots, i_K)$ , where

$$882 \quad \begin{aligned} j_1 &= i_1 + p_1(i_2 - 1) + \dots + p_{\lfloor K/2 \rfloor - 1}(i_{\lfloor K/2 \rfloor} - 1), \\ j_2 &= i_{\lceil K/2 \rceil} + p_{\lceil K/2 \rceil}(i_{\lceil K/2 \rceil + 1} - 1) + \dots \\ &\quad + p_{\lceil K/2 \rceil} \cdot p_{K-1}(i_K - 1). \end{aligned}$$

885 The matrix  $\text{Mat}_{sq}(\mathcal{Y})$  is asymmetric. We interpret  
886  $\text{Mat}_{sq}(\mathcal{Y})$  as the adjacency matrix for a bipartite net-  
887 work with connections between two groups of nodes.  
888 The two groups of nodes in the bipartite network have  
889  $p_1 \cdots p_{\lfloor K/2 \rfloor}$  and  $p_{\lceil K/2 \rceil} \cdots p_K$  nodes, respectively. The  
890 entry  $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2)$  refers to the presence of con-  
891 nection between the nodes indexed by combinations  
892  $(i_1, \dots, i_{\lfloor K/2 \rfloor})$  and  $(i_{\lceil K/2 \rceil}, \dots, i_K)$ . We summarize the  
893 procedure in Algorithm 2.

894 *Proposition 1 (Error for Bernoulli Initialization):*

895 Consider the Bernoulli dTBM in the parameter space  $\mathcal{P}$   
896 with fixed  $r \geq 1, K \geq 2$ . Assume that Assumption 1  
897 holds,  $\boldsymbol{\theta}$  is balanced, and  $\min_{i \in [p]} \theta(i) \geq c$  for some

---

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

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

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

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

---

constant  $c > 0$ . Let  $z_k^{(0)}$  denote the output of Algorithm 2.  
With probability going to 1 as  $p \rightarrow \infty$ , we have

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

*Remark 8 (Comparison with Gaussian model):* 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 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.

*Remark 9 (Comparison With Previous Methods):*

Previous work [9] develops a spectral clustering method for Bernoulli dTBM. [9] adopts a different signal notion based on the singular gap in the core tensor, denoted as  $\Delta_{\text{singular}}$ . By [9, Theorem 1], the spectral method achieves exact recovery with  $\Delta_{\text{singular}} \gtrsim p^{-1/2}$ . However, we are not able to infer the exact recovery of spectral method by our angle-base SNR condition. Consider an order-2 dTBM with  $p > 2, \sigma^2 = 1, \boldsymbol{\theta} = \mathbf{1}_p$ , equal size assignment  $|z^{-1}(a)| = p/r$  for all  $a \in [r]$ , and core matrix equal to the 2-dimensional identity matrix  $\mathbf{S} = \mathbf{I}_2$ . The singular gap under this setting is  $\Delta_{\text{singular}} = \min\{\lambda_1 - \lambda_2, \lambda_2\} = 0$ , where  $\lambda_1 \geq \lambda_2$  are singular values of  $\mathbf{S}$ . In contrast, our angle gap  $\Delta_{\text{min}}^2 = 2$  satisfies the SNR condition in Theorem 5. Then, our algorithm achieves the exact recovery, but the spectral method in [9] fails.

Hence, for fair comparison, we compare the best performance of our algorithm and [9] under the strongest signal setting of each model. Since both methods contain an iteration procedure, we set the iteration number to infinity to avoid the computational error. Considering the largest angle-based SNR  $\asymp 1$  in Theorem 5, our Bernoulli clustering achieves exponential error rate of order  $\exp(-p^{(K-1)})$ ; considering the largest singular

gap  $\Delta_{\text{singular}} \asymp 1$  in Theorem 1 of [9], the spectral clustering has a polynomial error rate of order  $p^{-2}$ . Our algorithm still shows a better theoretical accuracy than the competitive work for Bernoulli observations.

- *Extension to sparse binary dTBMs.* The sparsity is often a popular feature in hypergraphs [9], [16], [28]. Specifically, the sparse binary dTBM assumes that, the entries of  $\mathcal{Y}$  follow independent Bernoulli distributions with the mean

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

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 [28]. Inspired by [28], we adopt the diagonal-deleted HOSVD (D-HOSVD) [9] 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 directions. 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 A.

#### D. Practical Issues

1) *Computational Complexity:* Our two-stage algorithm has a computational cost polynomial in tensor dimension  $p$ . Specifically, the complexity of Sub-algorithm 1 is  $\mathcal{O}(Kp^{K+1} + Kp^K)$ , 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  $\mathcal{Y}_k^d$ , and  $pr^K$  comes from the normalization of  $\mathcal{Y}_k^d$ , the calculation of  $\mathcal{S}^{(t)}$ , and the cluster assignment update in Step 13.

2) *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 (1),

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

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

$$\hat{\theta}(r) = (\hat{\theta}(1, r), \dots, \hat{\theta}(p, r))^T, \quad 1000$$

with  $\hat{\theta}(i, r) = \|\hat{\mathcal{Y}}^d(r)_{i:}\| / \sum_{j: \hat{z}(j, r) = \hat{z}(i, r)} \|\hat{\mathcal{Y}}^d(r)_{j:}\|$ ,  $\hat{\mathcal{Y}}^d(r) = \text{Mat}(\hat{\mathcal{Y}}^d(r))$ ,  $\hat{\mathcal{Y}}^d(r)(i, a_2, \dots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \dots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}$ , 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 NONDEGREE TENSOR BLOCK MODEL

We discuss the connections and differences between dTBM and TBM [13] 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 [13] 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 [13] 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 (4) and the Euclidean-based SNR in [13] as  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$ , respectively:

$$\Delta_{\text{ang}}^2 = 2(1 - \max_{a \neq b \in [r]} \cos(\mathbf{S}_{a:}, \mathbf{S}_{b:})), \quad 1020$$

$$\Delta_{\text{Euc}}^2 = \min_{a \neq b \in [r]} \|\mathbf{S}_{a:} - \mathbf{S}_{b:}\|^2. \quad 1021$$

By Lemma 4 in the Appendix B, we have

$$\Delta_{\text{ang}}^2 \max_{a \in [r]} \|\mathbf{S}_{a:}\|^2 \leq \Delta_{\text{Euc}}^2. \quad 1023$$

The above inequality indicates that 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

$$\mathbf{S} = \begin{pmatrix} p^{(\gamma+1)/2} + 2 & 2 \\ 2 & 4 \end{pmatrix}, \quad \text{with } \gamma \leq -1. \quad 1035$$

The core matrix  $\mathbf{S}$  lies in the parameter spaces of TBM and our dTBM. Here, the constraint  $\gamma \leq -1$  is added to ensure the bounded condition of  $\mathbf{S}$  in our parameter

space in (2). The angle-based and Euclidean-based signal levels of  $S$  are

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

We conclude that TBM with  $S$  achieves exact recovery with a polynomial-time algorithm; see [13, 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 [13] and compare with ours.

*Statistical critical value:*

Ours:  $\Delta_{\text{ang}}^2 \lesssim p^{-(K-1)} \Rightarrow$  statistically impossible;  
 $\Delta_{\text{ang}}^2 \gtrsim p^{-(K-1)} \Rightarrow$  MLE achieves exact recovery;

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

*Computational critical value:*

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

The above comparison reveals four major differences.

First, none of our results in Section III are corollaries of [13]. Both models show the similar conclusion but under different conditions. While the TBM impossibility [13] 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 [13] 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 dTBM-specific techniques to handle the extra degree heterogeneity. In our Theorem 2, we construct a special nontrivial degree heterogeneity to establish the lower bound for arbitrary core tensor with small angle gap, while, TBM [13] considers the constructions without degree parameter. In our Theorem 3, we construct a rank-2 tensor to relate HPC conjecture to  $\Delta_{\text{ang}}^2$ , while TBM [13] constructs a rank-1 tensor to relate HPC conjecture to  $\Delta_{\text{Euc}}^2$ . The asymptotic nonequivalence between

$\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$  renders our proof technically more involved.

Last, we discuss the statistical impossibility statements. Our Theorem 2 implies the statistical impossibility whenever the core tensor  $S$  leads to an angle-based SNR below the critical value, while, Theorem 6 in [13] implies the worst case statistical impossibility for a particular core tensor  $S$  with Euclidean-based SNR below the statistical limit. Hence, our Theorem 2 shows a stronger statistical impossibility for dTBM than that presented in TBM [13, Theorem 6]. However, inspecting the proof of [13], the proof of Theorem 6 indeed implies a stronger TBM impossibility statement for arbitrary core tensor; i.e., when  $\gamma < -(K - 1)$

$$\liminf_{p \rightarrow \infty} \inf_{S \in \mathcal{P}_{S,\text{TBM}} \cap \{\Delta_{\text{Euc}}^2 = p^\gamma\}} \inf_{\hat{z}_{\text{stats}}} \sup_{z \in \mathcal{P}_{z,\text{TBM}}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z)] \geq 1,$$

where  $\mathcal{P}_{S,\text{TBM}}$  and  $\mathcal{P}_{z,\text{TBM}}$  refer to the space for core tensor  $S$  and assignment  $z$  under TBM, respectively. Again, in terms of the strong statistical impossibility, both models show the similar conclusion but under different conditions. Since two impossibilities consider different core tensor regimes with nonequivalent  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$ , we emphasize that different proof techniques are required to obtain these similar conclusions. See our proof sketch in Section VIII-A, Appendices B-D and B-E for detail technical differences.

- *Algorithms.* Both [13] 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 [29], [30]. 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). The CER between  $(\hat{z}, z)$  is equivalent to misclustering error  $\ell(\hat{z}, z)$  up to constant multiplications [31], and a lower CER indicates a better performance.

We generate order-3 tensors with *assortative* [15] 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 (4). Hence, we easily adjust SNR via varying  $\alpha$ . 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

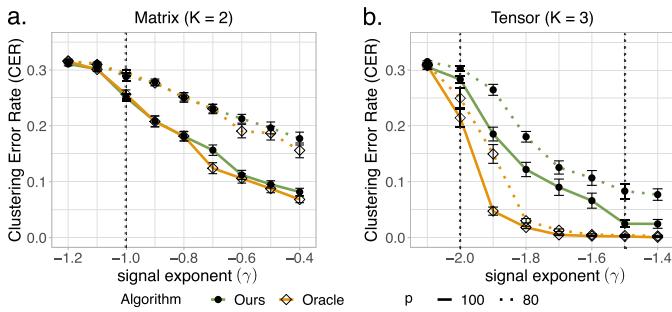


Fig. 4. SNR phase transitions for clustering in dTBM with  $p = \{80, 100\}$ ,  $r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

probability across  $r$  clusters for each mode. Without further explanation, we generate degree heterogeneity  $\boldsymbol{\theta}$  from absolute normal distribution by  $\theta(i) = |X_i| + 1 - 1/\sqrt{2\pi}$  with  $|X_i| \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ ,  $i \in [p]$  and normalize  $\boldsymbol{\theta}$  to satisfy (2). 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. Fig. 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, Fig. 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  $\gamma_{\text{comp}} = -1.5$ . Fig. 4 confirms the existence of the statistical-computational gap in our Theorems 2 and 3.

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  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$ . Fig. 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.

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 ( $p = 50, \sigma^2 = 1$ ), and the accuracy immediately increases with larger dimension  $p = 80$ . The improvement follows from the fact that a larger dimension  $p$  indicates

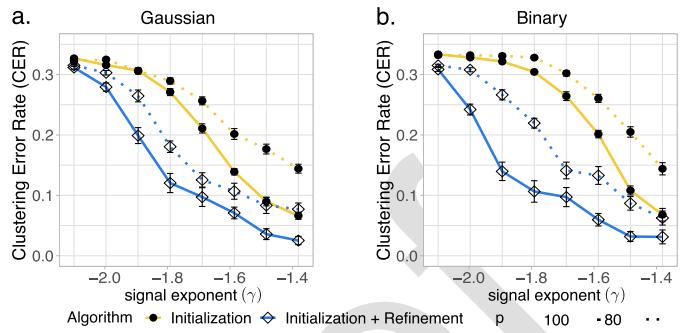


Fig. 5. CER versus signal exponent ( $\gamma$ ) for initialization only and for combined algorithm. We set  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$  under (a) Gaussian models and (b) Bernoulli models.

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** [13]: High-order clustering algorithm developed for nondegree tensor block models;
- **SCORE** [9]: 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 [9]. The methods **SCORE** and **HOSVD+** are designed for degree models, whereas **HOSVD** and **HLloyd** are designed for nondegree 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  $\gamma \in [-1.5, -1.1]$ . Fig. 6 shows that our method **dTBM** outperforms all other algorithms. The suboptimality of **SCORE** and **HOSVD+** indicates the necessity of local iterations on the clustering. Furthermore, Fig. 6 shows the inadequacy of nondegree algorithms in the presence of mild degree heterogeneity. The experiment demonstrates the benefits of addressing heterogeneity in higher-order clustering tasks.

The only exception in Fig. 6 is the slightly better performance of **HLloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLloyd** 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  $\boldsymbol{\theta}$  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

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

| Settings                           | $p = 50, \sigma^2 = 0.25$ |           | $p = 50, \sigma^2 = 1$ |           | $p = 80, \sigma^2 = 0.25$ |      | $p = 80, \sigma^2 = 1$ |           |
|------------------------------------|---------------------------|-----------|------------------------|-----------|---------------------------|------|------------------------|-----------|
|                                    | 2                         | 4         | 2                      | 4         | 2                         | 4    | 2                      | 4         |
| True cluster number $r$            |                           |           |                        |           |                           |      |                        |           |
| 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) |

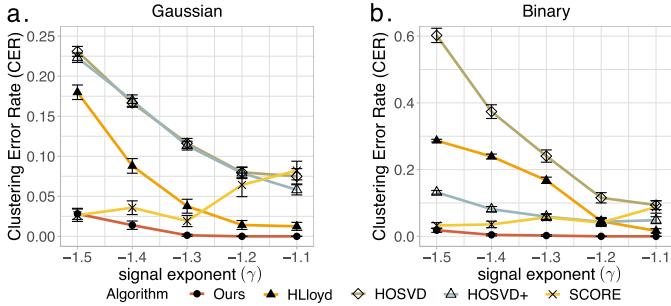


Fig. 6. CER versus signal exponent (denoted  $\gamma$ ) for different methods. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under (a) Gaussian and (b) Bernoulli models.

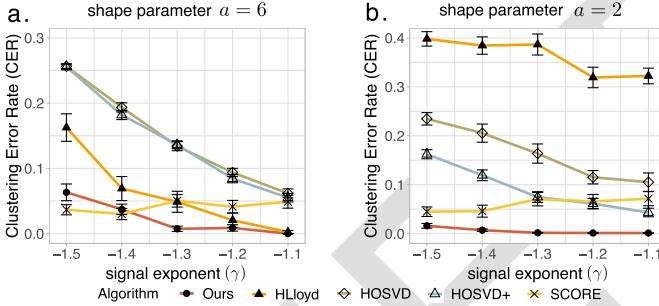


Fig. 7. CER comparison versus signal exponent (denoted  $\gamma$ ) under (a) low (shape parameter  $a = 6$ ) (b) high (shape parameter  $a = 2$ ) degree heterogeneity. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under Gaussian model.

1229  $a$  is called *shape parameter*. We vary  $a \in \{2, 6\}$  and choose  $b$   
1230 such that  $\mathbb{E}X = a(a-1)^{-1}b = 1$  for  $X$  following  $\text{Pareto}(a, b)$ .  
1231 Note that a smaller  $a$  leads to a larger variance in  $\theta$  and hence a  
1232 larger degree heterogeneity. We consider the Gaussian model  
1233 under low ( $a = 6$ ) and high ( $a = 2$ ) degree heterogeneity.  
1234 Fig. 7 shows that the errors for nondegree algorithms (**HLlloyd**,  
1235 **HOSVD**) increase with degree heterogeneity. In addition, the  
1236 advantage of **HLlloyd** over **HOSVD+** disappears with higher  
1237 degree heterogeneity.

1238 The last experiment investigates the effects of degree hetero-  
1239 geneity to clustering performance. We fix the signal exponent  
1240  $\gamma = -1.2$  and vary the extent of degree heterogeneity.  
1241 In this experiment, we generate  $\theta$  from Pareto distribution  
1242 prior to normalization. We vary the shape parameter  $a \in$   
1243  $[3, 6]$  in the Pareto distribution to investigate a range of  
1244 degree heterogeneities. Fig. 8 demonstrates the stability of  
1245 degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over  
1246 the entire range of degree heterogeneity under considera-  
1247 tion. In contrast, nondegree algorithms (**HLlloyd**, **HOSVD**)  
1248 show poor performance with large heterogeneity, especially in  
1249 Bernoulli cases. This experiment, again, highlights the benefit  
1250 of addressing degree heterogeneity in higher-order clustering.

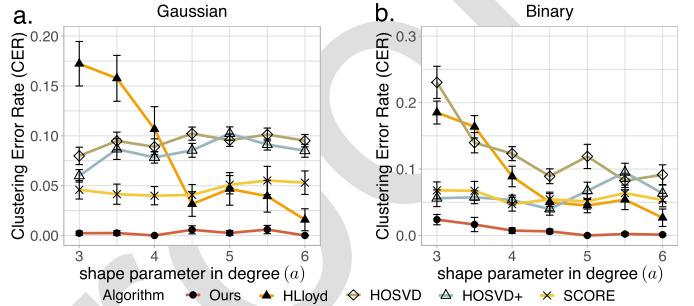


Fig. 8. CER versus shape parameter in degree ( $a \in [3, 6]$ ) for different methods. We set  $p = 100, r = 5, \gamma = -1.2$  under (a) Gaussian and (b) Bernoulli models.

## 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 [32]. We preprocess the original dataset following [33] 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 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 [34]. Because of the symmetry in the data, the estimated brain node clustering results are the same on the first and second modes. Fig. 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.

Fig. 10 illustrates the estimated core tensor  $\hat{\mathcal{S}}$  with estimated clustering, and Fig. 11 visualizes the average brain connections and the connection enrichment in contrast to average networks in each group. In general, we find that the

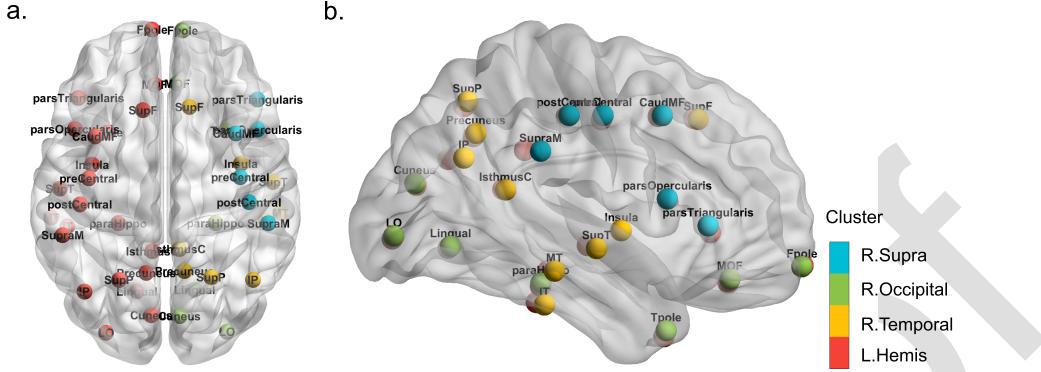


Fig. 9. Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

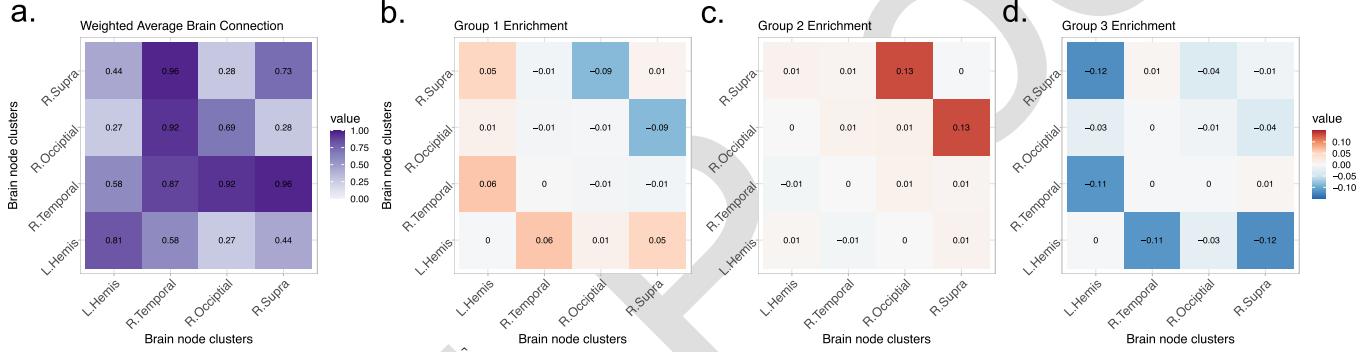


Fig. 10. Mode 3 slices of estimated core tensor  $\hat{S}$ . (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the difference between each slice of  $\hat{S}$  and the averaged slice.

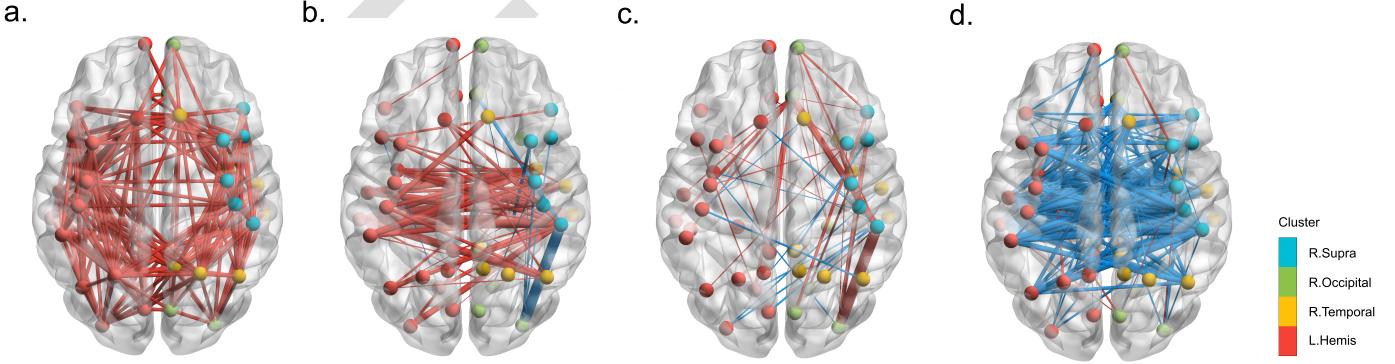


Fig. 11. Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain network enrichments in Groups 1-3. Red edges represent the positive enrichment and blue edges represent the negative enrichment.

inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Fig. 10a). 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 (Fig. 10b), while group 4 with only 36% females exhibits a reduction (Fig. 10d). This result agrees with previous findings in [34]. The enrichment on the back-front connection is also recognized in group 3 (Fig. 10c). The interpretive patterns in our results demonstrate the usefulness of our clustering methods 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 [35]. 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

1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309

TABLE IV  
CLUSTERING ERRORS (MEASURED BY CER) FOR VARIOUS METHODS IN  
THE ANALYSIS OF PERU LEGISLATION DATASET

| 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        |

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.

### VIII. PROOF SKETCHES

In this section, we provide the proof sketches for the main Theorem 2 (Impossibility), Theorem 3 (Impossibility), and Theorems 4-5. Detail proofs and extra theoretical results are provided in Appendix B.

#### A. Proof Sketch of Theorem 2 (Impossibility) and Theorem 3 (Impossibility)

The proofs of impossibility in Theorems 2 and 3 share the same proof idea with [13, Theorems 6 and 7] and [15, Theorem 2]. In both proofs of statistical and computational impossibilities, the key idea is to construct a particular set of parameters to lower bound the minimax rate. Specifically, for statistical impossibility in Theorem 2, we construct a particular  $(z_{\text{stats}}^*, \theta_{\text{stats}}^*) \in \mathcal{P}_{z,\theta}$  such that for all  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$

$$\begin{aligned} & \inf_{\hat{z}_{\text{stats}}} \sup_{(z,\theta) \in \mathcal{P}_{z,\theta}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z)] \\ & \geq \inf_{\hat{z}_{\text{stats}}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z_{\text{stats}}^*) | (z_{\text{stats}}^*, \mathcal{S}^*, \theta_{\text{stats}}^*)] \geq 1; \end{aligned} \quad (20)$$

for computational impossibility in Theorem 3, we construct a particular  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*) \in \mathcal{P}(\gamma)$  such that

$$\begin{aligned} & \inf_{\hat{z}_{\text{comp}}} \sup_{(z,\mathcal{S},\theta) \in \mathcal{P}(\gamma)} \mathbb{E}[p\ell(\hat{z}_{\text{comp}}, z)] \\ & \geq \inf_{\hat{z}_{\text{comp}}} \mathbb{E}[p\ell(\hat{z}_{\text{comp}}, z_{\text{comp}}^*) | (z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)] \geq 1. \end{aligned}$$

The constructions of  $(z_{\text{stats}}^*, \theta_{\text{stats}}^*)$  and  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)$  are the most critical steps. With good constructions, the lower bound “ $\geq 1$ ” can be verified by classical statistical conclusions (e.g. Neyman-Pearson Lemma) or prior work (e.g. HPC Conjecture).

A notable detail in the proof of statistical impossibility is the arbitrariness of  $\mathcal{S}^*$ . The first infimum over  $\mathcal{P}_S(\gamma)$  in the minimax rate (10) requires that the lower bound (20) holds for any  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$ . The arbitrary choice of  $\mathcal{S}^*$  brings extra difficulties in the parameter construction, and consequently a nontrivial  $\theta_{\text{stats}}^* \neq 1$  is chosen to address the arbitrariness. Previous TBM construction in the proof of [13, Theorem 6] with  $\theta_{\text{stats}}^* = 1$  is no longer applicable in our case. Meanwhile, our construction  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)$  leads to a rank-2 mean tensor to relate the HPC Conjecture while TBM [13, Theorem

7] constructs a rank-1 mean tensor. Hence, we emphasize that dTBM-specific techniques are required to obtain our impossibility results, though the proof idea is common for minimax lower bound analysis.

#### B. Proof Sketch of Theorem 4

The proof of Theorem 4 is inspired by the proof idea of [15, 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

$$\begin{aligned} \ell(z^{(0)}, z) & \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ & \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K} \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ & \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \end{aligned} \quad (21)$$

where  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  is the true mean. The first inequality in (21) holds with the assumption  $\min_{i \in [p]} \theta(i) \geq c > 0$  in Theorem 4. The second inequality relies on the key Lemma 1, which indicates

$$\min_{z(i) \neq z(j)} \|[\mathbf{X}_{i:}]^s - [\mathbf{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \quad (22)$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (22) (or the proof of Lemma 1), in which the proof of [15] is no longer applicable due to different angle gap assumption in our dTBM. To address the angle gap notion, we develop the extra padding technique in Lemma 5 and balance assumption (6). Last, we finish the proof of Theorem 4 by showing the third inequality of (21) using [13, Proposition 1].

#### C. Proof Sketch of Theorem 5

The proof of Theorem 5 is inspired by the proof idea of [13, 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)

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

We show that  $L^{(t)}$  provides an upper bound for the misclustering error of interest via the inequality  $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2}$  in Lemma 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

1397  $t$ -th iteration quantities  $\mathbf{W}^{(t)}, \mathbf{V}^{(t)}$  corresponding to  $\mathbf{M}^{(t)}$  (or  
 1398 equivalently  $z^{(t)}$ ). To evaluate  $L^{(t+1)}$ , we prove the bound

$$\begin{aligned} & \mathbb{1}\left\{z^{(t+1)}(i) = b\right\} \\ &= \mathbb{1}\left\{\|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_{b:}^{(t)}]_s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_{z(i):}^{(t)}]_s\|^2\right\} \\ &\leq A_{ib} + B_{ib}, \end{aligned} \quad (23)$$

1402 where  $\mathbf{Y} = \text{Mat}(\mathcal{Y})$ ,  $\mathbf{S} = \text{Mat}(\mathcal{S})$ ,  $\mathbf{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$  and

$$\begin{aligned} 1403 \quad A_{ib} &= \mathbb{1}\left\{\left\langle \mathbf{E}_{i:} \mathbf{V}, \left[\tilde{\mathbf{S}}_{z(i):}\right]^s - \left[\tilde{\mathbf{S}}_{b:}\right]^s\right\rangle \lesssim -\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2\right\}, \\ 1404 \quad B_{ib} &= \mathbb{1}\left\{\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \lesssim F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}\right\}. \end{aligned}$$

1405 The terms  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$  are controlled by  $z^{(t)}, \mathcal{S}^{(t)}$ ; see the  
 1406 detailed definitions in (68), (69), (70). Note that the event  $A_{ib}$   
 1407 only involves the oracle estimator independent of  $t$ , while all  
 1408 the terms related to the  $t$ -th iteration are in  $B_{ib}$ . Thus, the  
 1409 inequality (23) decomposes the misclustering loss in the  $(t+1)$ -th iteration  
 1410 into the oracle loss and the loss in  $t$ -th iteration.  
 1411 This decomposition leads to the separation of statistical error  
 1412 and computational error in the final upper bound of Theorem 5.  
 1413 Specifically, we prove the contraction inequality

$$\begin{aligned} 1414 \quad L^{(t+1)} &\leq M\xi + \rho L^{(t)}, \\ 1415 \quad \text{with } \xi &= \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \end{aligned} \quad (24)$$

1416 where  $M$  is a positive constant,  $\rho \in (0, 1)$  is the contraction  
 1417 parameter, and we call  $\xi$  the oracle loss. Controlling the  
 1418 probability of event  $B_{ib}$  and obtaining the  $\rho L^{(t)}$  term in the  
 1419 right hand side of (24) are the most challenging parts in  
 1420 the proof of Theorem 5. Note that the true and estimated  
 1421 core tensors are involved via their normalized rows such  
 1422 as  $\mathbf{S}_{a:}^s, \tilde{\mathbf{S}}_{a:}^s, [\mathbf{S}_{a:}^{(t)}]^s$ . The Cartesian coordinate based analysis  
 1423 in [13] is no longer applicable in our case. Instead, we use  
 1424 the polar-coordinate based analysis and the geometry property  
 1425 of trigonometric functions to derive the high probability upper  
 1426 bounds for  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ .

1427 Further, by sub-Gaussian concentration, we prove the high  
 1428 probability upper bound for oracle loss

$$\xi \lesssim \text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right). \quad (25)$$

1430 Combining the decomposition (24) and the oracle bound (25),  
 1431 we finish the proof of Theorem 5.

1432 The proof of MLE error shares the similar idea as Theorems 4-5. We first show a weaker polynomial rate for MLE  
 1433 and then improve the rate from polynomial to exponential  
 1434 through the iterations. The only difference is that the MLE  
 1435 remains the same over iterations due to its global optimality.  
 1436 See Appendix B-G for the detailed proof.

## APPENDIX A ADDITIONAL NUMERICAL EXPERIMENTS

### A. Bernoulli Phase Transition

1441 The first additional experiment verifies the  
 1442 statistical-computational gap in Section III under the Bernoulli  
 1443 model. Consider the Bernoulli model with  $p = \{80, 100\}$ ,

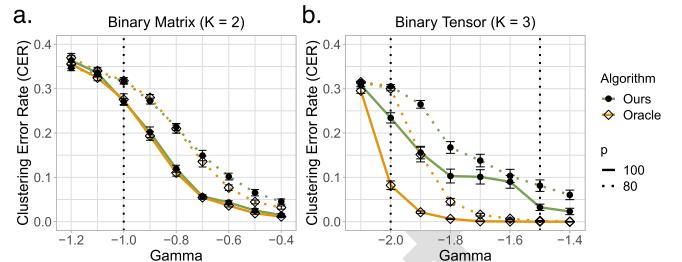


Fig. 12. SNR phase transitions for Bernoulli dTBM with  $p = \{80, 100\}$ ,  $r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

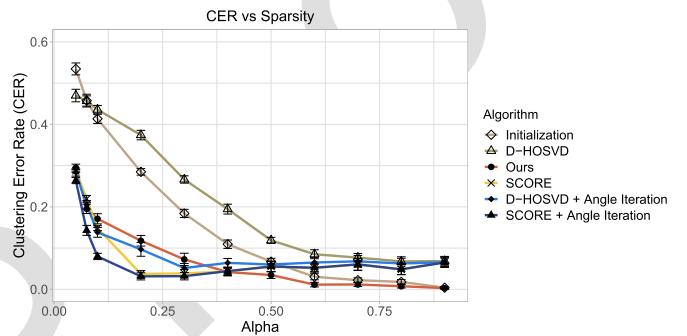


Fig. 13. CER comparison versus sparsity parameter  $\alpha_p$  in  $[0.05, 0.9]$ . We set  $p = 100, r = 5$  and  $\gamma = -1.2$  under sparse binary dTBM.

1444  $r = 5$ . We vary  $\gamma$  in  $[-1.2, -0.4]$  and  $[-2.1, -1.4]$  for  
 1445 matrix ( $K = 2$ ) and tensor ( $K = 3$ ) clustering, respectively.  
 1446 We approximate MLE using an oracle estimator, i.e., the  
 1447 output of Sub-algorithm 2 initialized from the true assignment.  
 1448 Fig. 12 shows a similar pattern as Fig. 4. The algorithm and  
 1449 oracle estimators have no gap in the matrix case, while an  
 1450 error gap emerges between the critical values  $\gamma_{\text{stat}} = -2$  and  
 1451  $\gamma_{\text{comp}} = -1.5$  in the tensor case. Fig. 4 suggests the  
 1452 statistical-computational gap in Bernoulli models.

### B. Sparsity

1453 The second additional experiment evaluates the algorithm  
 1454 performances under the sparse binary dTBM (18). We fix the  
 1455 signal exponent  $\gamma = -1.2$  and vary the sparsity parameter  
 1456  $\alpha_p \in [0.05, 0.9]$ . A smaller  $\alpha_p$  leads to a higher probability  
 1457 of zero entries in the observation. In addition to the three  
 1458 algorithms mentioned in Section VI-B (denoted **Initialization**,  
 1459 **dTBM**, and **SCORE**), we consider other three algorithms based  
 1460 on the discussion in Section IV-C:

- **D-HOSVD**, the diagonal-deleted HOSVD in [9];
- **D-HOSVD + Angle**, the combined algorithm of our  
 1463 angle-based iteration with initialization from **D-HOSVD**;
- **SCORE + Angle**, the combined algorithms of our  
 1465 angle-based iteration with initialization from **SCORE**.

1466 Fig. 13 shows a slightly larger error in **dTBM** than that in  
 1467 **SCORE**, **D-HOSVD + Angle**, and **SCORE + Angle** under the  
 1468 sparse setting with  $\alpha_p < 0.3$ . The small gap between **dTBM**  
 1469 and other sparse-specific methods implies the robustness of our  
 1470 algorithm. In addition, comparing **SCORE** versus **SCORE + Angle**  
 1471 (or **D-HOSVD** versus **D-HOSVD + Angle**) indicates the  
 1472 benefit of our angle iterations under the sparse dTBM. In the  
 1473

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 B PROOFS

We provide the proofs for all the theorems in our main paper. In each subsection, 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.

#### 1) Preliminaries:

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

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

• For a vector  $\mathbf{a}$ , let  $\mathbf{a}^s := \mathbf{a}/\|\mathbf{a}\|$  denote the normalized vector. We make the convention that  $\mathbf{a}^s = \mathbf{0}$  if  $\mathbf{a} = \mathbf{0}$ .

• For a matrix  $\mathbf{A} \in \mathbb{R}^{n \times m}$ , let  $\mathbf{A}^{\otimes K} := \mathbf{A} \otimes \cdots \otimes \mathbf{A} \in \mathbb{R}^{n^K \times m^K}$  denote the Kronecker product of  $K$  copies of matrices  $\mathbf{A}$ .

• For a matrix  $\mathbf{A}$ , let  $\|\mathbf{A}\|_\sigma$  denote the spectral norm of matrix  $\mathbf{A}$ , which is equal to the maximal singular value of  $\mathbf{A}$ ; let  $\lambda_k(\mathbf{A})$  denote the  $k$ -th largest singular value of  $\mathbf{A}$ ; let  $\|\mathbf{A}\|_F$  denote the Frobenius norm of matrix  $\mathbf{A}$ .

2) 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

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where } \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \mathbf{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \mathbf{M}_K, \quad (26)$$

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  $\mathbf{M}_k \in \{0, 1\}^{p_k \times r_k}$  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  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ ,

$$\begin{aligned} \mathcal{P}(\{r_k\}) &= \left\{ (\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) : \boldsymbol{\theta}_k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} |z_k^{-1}(a)| \leq \frac{c_2 p_k}{r_k}, \right. \\ &\quad c_3 \leq \|\mathbf{S}_{k,a,:}\| \leq c_4, \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|_1 = |z_k^{-1}(a)|, \\ &\quad \left. \text{for all } a \in [r_k], k \in [K] \right\}. \end{aligned} \quad (27)$$

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

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|. \quad (28)$$

We also consider the generalized Assumption 1 on angle gap.

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

$$\Delta_{\min} := \min_{k \in [K]} \min_{a \neq b \in [r_k]} \|\mathbf{S}_{k,a,:}^s - \mathbf{S}_{k,b,:}^s\| > 0. \quad (29)$$

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

$$\mathcal{P}(\gamma) = \mathcal{P}(\{r_k\}) \cap \{\mathcal{S} \text{ satisfies } \text{SNR} = p^\gamma \text{ and } p_k \asymp p, k \in [K]\}. \quad (30)$$

### B. Proof of Theorem 1

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

$$\begin{aligned} \mathcal{X} &= \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \mathbf{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \mathbf{M}'_K \\ &= \mathcal{S}' \times_1 \boldsymbol{\Theta}'_1 \mathbf{M}'_1 \times_2 \cdots \times_K \boldsymbol{\Theta}'_K \mathbf{M}'_K, \end{aligned} \quad (31)$$

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 (26) violates Assumption 2. Note that  $\Delta_{\min}^2 = 1$  when there exists  $k \in [K]$  such that  $r_k = 1$ . Hence, we consider the case that  $r_k \geq 2$  for all  $k \in [K]$ . Without loss of generality, we assume  $\|\mathbf{S}_{1,1,:}^s - \mathbf{S}_{1,2,:}^s\| = 0$ .

By constraints in parameter space (27), 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

$$\mathcal{S} = \mathcal{S}_0 \times_1 \mathbf{C} \mathbf{R}, \quad (32)$$

where  $\mathbf{C} = \text{diag}(1, c, 1, \dots, 1) \in \mathbb{R}^{r_1 \times r_1}$  and

$$\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{1}_{r_1-2} \end{pmatrix} \in \mathbb{R}^{r_1 \times (r_1-1)}. \quad (33)$$

Let  $\mathbf{D} = \text{diag}(1 + c, 1, \dots, 1) \in \mathbb{R}^{r_1-1 \times r_1-1}$ . Consider the parameterization  $\mathbf{M}'_1 = \mathbf{M}_1 \mathbf{R}$ ,  $\mathcal{S}' = \mathcal{S}_0 \times_1 \mathbf{D}$ , and

$$\theta'_1(i) = \begin{cases} \frac{1}{1+c} \theta_1(i) & i \in z_1^{-1}(1), \\ \frac{c}{1+c} \theta_1(i) & i \in z_1^{-1}(2), \\ \theta_1(i) & \text{otherwise,} \end{cases} \quad (34)$$

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 (26) satisfies Assumption (2). We show the uniqueness of the three parameters,  $\{M_k\}, \{\mathcal{S}\}, \{\theta_k\}$  separately.

First, we show the uniqueness of  $M_k$  for all  $k \in [K]$ . When  $r_k = 1$ , all possible  $M_k$ 's are equal to the vector  $\mathbf{1}_{p_k}$ , and the uniqueness holds trivially. Hence, we consider the case that  $r_k \geq 2$ . Without loss of generality, we consider  $k = 1$  with  $r_1 \geq 2$  and show the uniqueness of the first mode membership matrix; i.e.,  $M'_1 = M_1 P_1$  where  $P_1$  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  $\|X_{1,z_1(i)}^s - X_{1,z_1(j)}^s\| = 0$  and thus  $\|(S')_{1,z'_1(i)}^s - (S')_{1,z'_1(j)}^s\| = 0$  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  $\|X_{1,i}^s - X_{1,j}^s\| \neq 0$  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  $\theta_k$  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

$$\frac{X_{1,z_1(i)}}{X_{1,z_1(j)}} = \frac{X'_{1,z_1(i)}}{X'_{1,z_1(j)}}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta'_1(j)}{\theta'_1(i)}. \quad (29)$$

Let  $\theta'_1(j) = c\theta_1(j)$  for some positive constant  $c$ . By equation (29), we have  $\theta'_1(i) = c\theta_1(i)$  for all  $i \in z_1^{-1}(a)$ . By the constraint  $(\{z_k\}, \mathcal{S}', \{\theta'_k\}) \in \mathcal{P}(\{r_k\})$ , 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  $\theta_1 = \theta'_1$  provided that  $z_1 = z'_1$ .

Last, we show the uniqueness of  $\mathcal{S}$ ; i.e.,  $\mathcal{S}' = \mathcal{S} \times_1 P_1^{-1} \times_2 \cdots \times_K P_K^{-1}$ , where  $P_k$ 's are permutation matrices for all  $k \in [K]$ . Provided  $z'_k = z_k, \theta'_k = \theta_k$ , we have  $M'_k = M_k P_k$  and  $\Theta'_k = \Theta_k$  for all  $k \in [K]$ .

Let  $D_k = [(\Theta'_k M'_k)^T (\Theta'_k M'_k)]^{-1} (\Theta'_k M'_k)^T, k \in [K]$ . By the parameterization (28), we have

$$\begin{aligned} \mathcal{S}' &= \mathcal{X} \times_1 D_1 \times_2 \cdots \times_K D_K \\ &= \mathcal{S} \times_1 D_1 \Theta_1 M_1 \times_1 \cdots \times_K D_K \Theta_K M_K \\ &= \mathcal{S} \times_1 P_1^{-1} \times_2 \cdots \times_K P_K^{-1}. \end{aligned}$$

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 (26) with  $(\{z_k\}, \mathcal{S}, \{\theta_k\}) \in \mathcal{P}(\{r_k\})$  and  $r_k \geq 2, k \in [K]$ . Then, for any  $k \in [K]$  and index pair  $(i, j) \in [p_k]^2$ , we have

$$\begin{aligned} \|S_{k,z_k(i)}^s - S_{k,z_k(j)}^s\| &= 0 \quad \text{if and only if} \\ \|X_{k,z_k(i)}^s - X_{k,z_k(j)}^s\| &= 0. \end{aligned}$$

*Proof of Lemma 3:* Without loss of generality, we prove  $k = 1$  only and drop the subscript  $k$  in  $X_k, S_k$  for notational convenience. By tensor matricization, we have

$$X_{j:} = \theta_1(j) S_{z_1(j):} [\Theta_2 M_2 \otimes \cdots \otimes \Theta_K M_K]^T. \quad (1620)$$

Let  $\tilde{M} = \Theta_2 M_2 \otimes \cdots \otimes \Theta_K M_K$ . Notice that for two vectors  $a, b$  and two positive constants  $c_1, c_2 > 0$ , we have

$$\|a^s - b^s\| = \|(c_1 a)^s - (c_2 b)^s\|. \quad (1623)$$

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

$$\|S_{z_1(i):}^s - S_{z_1(j):}^s\| = 0 \quad \text{if and only if} \quad (1626)$$

$$\left\| \left[ S_{z_1(i):} \tilde{M}^T \right]^s - \left[ S_{z_1(j):} \tilde{M}^T \right]^s \right\| = 0. \quad (1627)$$

$$(\Leftarrow) \text{ Suppose } \left\| \left[ S_{z_1(i):} \tilde{M}^T \right]^s - \left[ S_{z_1(j):} \tilde{M}^T \right]^s \right\| = 0. \quad (1628)$$

There exists a positive constant  $c$  such that  $S_{z_1(i):} \tilde{M}^T = c S_{z_1(j):} \tilde{M}^T$ . Note that

$$S_{z_1(i):} = S_{z_1(i):} \tilde{M}^T \left[ \tilde{M} \left( \tilde{M}^T \tilde{M} \right)^{-1} \right], \quad (1631)$$

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

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

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  $S_{z(i):}$  can be folded to a tensor  $\mathcal{S}' = [\mathcal{S}'_{a_2, \dots, a_K}] \in \mathbb{R}^{r^{K-1}}$ ; i.e.,  $\text{vec}(\mathcal{S}') = S_{z(i):}$ . Define weight vectors  $w_{a_2, \dots, a_K}$  corresponding to the elements in  $\mathcal{S}'_{a_2, \dots, a_K}$  by

$$\begin{aligned} w_{a_2 \dots a_K} &= [\theta_{z^{-1}(a_2)}^T \otimes \cdots \otimes \theta_{z^{-1}(a_K)}^T] \in \mathbb{R}^{|z^{-1}(a_2)| \times \cdots \times |z^{-1}(a_K)|}, \end{aligned} \quad (1646)$$

for all  $a_k \in [r], k = 2, \dots, K$ , where  $\otimes$  denotes the Kronecker product. Therefore, we have  $X_{i:} = \theta(i) \text{Pad}_w(S_{z(i)})$  where  $w = \{w_{a_2, \dots, a_K}\}_{a_k \in [r], k \in [K]/\{1\}}$ . Specifically, we have  $\|w_{a_2, \dots, a_K}\|^2 = \prod_{k=2}^K \|\theta_{z^{-1}(a_k)}\|^2$ , and by the balanced assumption (6) we have

$$\max_{(a_2, \dots, a_K)} \|w_{a_2, \dots, a_K}\|^2 = (1 + o(1)) \min_{(a_2, \dots, a_K)} \|w_{a_2, \dots, a_K}\|^2. \quad (30) \quad (1652)$$

Consider the inner product of  $X_{i:}$  and  $X_{j:}$  for  $z(i) \neq z(j)$ . By the definition of weighted padding operator (56) and the balanced assumption (30), we have

$$\begin{aligned} &\langle X_{i:}, X_{j:} \rangle \\ &= \theta(i)\theta(j) \langle \text{Pad}_w(S_{z(i)}), \text{Pad}_w(S_{z(j)}) \rangle \\ &= \theta(i)\theta(j) \min_{(a_2, \dots, a_K)} \|w_{a_2, \dots, a_K}\|^2 \langle S_{z(i)}, S_{z(j)} \rangle (1 + o(1)). \end{aligned} \quad (1657)$$

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

Then, we have

$$\begin{aligned} \cos(\mathbf{S}_{z_1(i):}, \mathbf{S}_{z_1(j):}) &= \frac{\langle \mathbf{S}_{z_1(i):}, \mathbf{S}_{z_1(j):} \rangle}{\|\mathbf{S}_{z_1(i):}\| \|\mathbf{S}_{z_1(j):}\|} \\ &= (1 + o(1)) \frac{\langle \mathbf{X}_{i:}, \mathbf{X}_{j:} \rangle}{\|\mathbf{X}_{i:}\| \|\mathbf{X}_{j:}\|} \\ &= (1 + o(1)) \cos(\mathbf{X}_{i:}, \mathbf{X}_{j:}), \end{aligned}$$

where the second inequality follows by the balance assumption on  $\theta$ .

Further, notice that  $\|\mathbf{v}_1^s - \mathbf{v}_2^s\|^2 = 2(1 - \cos(\mathbf{v}_1, \mathbf{v}_2))$ . For all  $i, j$  such that  $z(i) \neq z(j)$ , when  $p \rightarrow \infty$ , we have

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

Combining the inequalities (12) and (12) in the proof of Theorem 2 in [15], we have

$$\begin{aligned} \inf_{\hat{z}_1} \mathbb{E} [\ell(\hat{z}_1, z_1^*) | (z_k^*, \mathcal{S}^*, \theta_k^*)] &\geq \\ \frac{C}{r^3 |T_1^c|} \sum_{i \in T_1^c} \inf_{\hat{z}_1(i)} \{ &\mathbb{P}[\hat{z}_1(i) = 1 | z_1^*(i) = 2, z_k^*, \mathcal{S}^*, \theta_k^*] \\ &+ \mathbb{P}[\hat{z}_1(i) = 2 | z_1^*(i) = 1, z_k^*, \mathcal{S}^*, \theta_k^*]\}, \end{aligned} \quad (31)$$

where  $C$  is some positive constant,  $\hat{z}_1$  on the left hand side denote the generic assignment functions in  $\mathcal{P}(\gamma)$ , and the infimum on the right hand side is taken over the generic assignment function family of  $\hat{z}_1(i)$  for all nodes  $i \in T_1^c$ . Here, the factor  $r^3 = r \cdot r^2$  in (31) 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.

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

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

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):* Consider the general asymmetric dTBM (26) in the special case that  $p_k = p$  and  $r_k = r$  for all  $k \in [K]$  with  $K \geq 2$ ,  $2 \leq r \lesssim p^{1/3}$  as  $p \rightarrow \infty$ . For simplicity, we show the minimax rate for the estimation on the first mode  $\hat{z}_1$ ; the proof for other modes are essentially the same.

To prove the minimax rate (10), it suffices to take an arbitrary  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$  with  $\gamma < -(K-1)$  and construct  $(z_k^*, \theta_k^*)$  such that

$$\inf_{\hat{z}_1} \mathbb{E} [p\ell(\hat{z}_1, z_1^*) | (z_k^*, \mathcal{S}^*, \theta_k^*)] \geq 1.$$

We first define a subset of indices  $T_k \subset [p_k]$ ,  $k \in [K]$  in order to avoid the complication of label permutation. Based on [13, Proof of Theorem 6], we consider the restricted family of  $\hat{z}_k$ 's for which the following three conditions are satisfied:

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

for all  $k \in [K]$ . Now, we consider the construction:

- (i)  $\{z_k^*\}$  satisfies properties (a)-(c) with misclassification sets  $T_k^c$  for all  $k \in [K]$ ;
- (ii)  $\{\theta_k^*\}$  such that  $\theta_k^*(i) \leq \sigma r^{(K-1)/2} p^{-(K-1)/2}$  for all  $i \in T_k^c$ ,  $k \in [K]$  and  $\max_{k \in [K], a \in [r]} \|\theta_{k,z_k^{*, -1}(a)}\|_2^2 \asymp p/r$ .

Next, we need to find the lower bound of the rightmost side in (31). We consider the hypothesis test based on model (26). First, we reparameterize the model under the construction (i)-(ii).

$$\mathbf{x}_a^* = [\text{Mat}_1(\mathcal{S}^* \times_2 \Theta_2^* \mathbf{M}_2^* \times_3 \cdots \times_K \Theta_K^* \mathbf{M}_K^*)]_{a:}, \quad (32)$$

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 (31) for  $i = 1$ . The analysis for other  $i \in T_1^c$  are similar. For notational simplicity, we suppress the subscript  $i$  and write  $\mathbf{y}, \theta^*, z$  in place of  $\mathbf{y}_1, \theta_1^*(1)$  and  $z_1(1)$ , respectively. The equivalent vector problem for assessing the summand in (31) is

$$\mathbf{y} = \theta^* \mathbf{x}_z^* + \mathbf{e}, \quad (32)$$

where  $z \in \{1, 2\}$  is an unknown parameter,  $\theta^* \in \mathbb{R}_+$  is the given heterogeneity degree,  $\mathbf{x}_1^*, \mathbf{x}_2^* \in \mathbb{R}^{p^{K-1}}$  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 (32):

$$H_0 : z = 1, \mathbf{y} = \theta^* \mathbf{x}_1^* + \mathbf{e} \leftrightarrow H_1 : z = 2, \mathbf{y} = \theta^* \mathbf{x}_2^* + \mathbf{e}, \quad (33)$$

The hypothesis testing (33) is a simple versus simple testing, since the assignment  $z$  is the only unknown parameter in the test. By Neyman-Pearson lemma, the likelihood ratio test is optimal with minimal Type I + II error. Under Gaussian model, the likelihood ratio test of (33) is equivalent to the least square estimator  $\hat{z}_{LS} = \arg \min_{a=\{1,2\}} \|\mathbf{y} - \theta^* \mathbf{x}_a^*\|_F^2$ .

Let  $\mathbf{S} = \text{Mat}_1(\mathcal{S})$ . Note that

$$\begin{aligned} &\|\theta^* \mathbf{x}_1^* - \theta^* \mathbf{x}_2^*\|_F \\ &\leq \theta^* \|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \prod_{k=2}^K \lambda_{\max}(\Theta_k^* \mathbf{M}_k^*) \\ &\leq \theta^* \|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \max_{k \in [K]/\{1\}, a \in [r]} \|\theta_{k,z_k^{*, -1}(a)}\|_2^{K-1} \\ &\leq \sigma r^{(K-1)/2} p^{-(K-1)/2} 2 c_4 p^{(K-1)/2} r^{-(K-1)/2} \\ &\leq 2 c_4 \sigma, \end{aligned}$$

where  $\lambda_{\max}(\cdot)$  denotes the maximal singular value, the second inequality follows from Lemma 6, and the third inequality

follows from property (ii) and the boundedness constraint in  $\mathcal{P}_S(\gamma)$  such that  $\|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \leq \|\mathbf{S}_{1:}^*\|_F + \|\mathbf{S}_{2:}^*\|_F \leq 2c_4$ .  
Hence, we have

$$\begin{aligned} & \inf_{\hat{z}_1(1)} \{\mathbb{P}[\hat{z}_1(1) = 1 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ & \quad + \mathbb{P}[\hat{z}_1(1) = 2 | z_1^*(1) = 1, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*]\} \\ &= 2\mathbb{P}[\hat{z}_{LS} = 1 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ &= 2\mathbb{P}[\|\mathbf{y} - \boldsymbol{\theta}^* \mathbf{x}_1^*\|_F^2 \leq \|\mathbf{y} - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F^2 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ &= 2\mathbb{P}[2\langle \mathbf{e}, \boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^* \rangle \geq \|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F^2] \\ &= 2\mathbb{P}[N(0, 1) \geq \boldsymbol{\theta}^* \|\mathbf{x}_1^* - \mathbf{x}_2^*\|_F / (2\sigma)] \\ &\geq 2\mathbb{P}[N(0, 1) \geq c_4] \geq c, \end{aligned} \quad (34)$$

where the first equation holds by symmetry, the third equation holds by rearrangement, the fourth equation holds from the fact that  $\langle \mathbf{e}, \boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^* \rangle \sim N(0, \sigma \|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F)$ , and  $c$  is some positive constant in the last inequality.

Plugging the inequality (34) into the inequality (31) for all  $i \in T_1^c$ , then, we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_1} \mathbb{E}[\rho(\hat{z}_1, z_1^*) | z_k^*, \boldsymbol{\theta}_k^*, \mathcal{S}^*] \geq \liminf_{p \rightarrow \infty} \frac{Ccp}{r^3} \geq Cc,$$

where the last inequality follows by the condition  $r = o(p^{1/3})$ . By the discrete nature of the misclustering error, we obtain our conclusion

$$\liminf_{p \rightarrow \infty} \inf_{\mathcal{S}^* \in \mathcal{P}_S(\gamma)} \inf_{\hat{z}_{\text{stat}}} \sup_{(z^*, \boldsymbol{\theta}^*) \in \mathcal{P}_{z, \boldsymbol{\theta}}} \mathbb{E}[\rho(\hat{z}_{\text{stat}}, z)] \geq 1.$$

Last, with constructed  $z_k^*, \boldsymbol{\theta}_k^*$  satisfying properties (i) and (ii) and  $\gamma' < -(K-1)$ , we construct a core tensor  $\mathcal{S}^*$  such that  $\Delta_{\mathbf{X}^*}^2 \leq p^{-(K-1)}$ . Based on the property (ii) and the boundedness constraint of  $\mathcal{S}^*$  in  $\mathcal{P}$ , we still have  $\|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F \leq 2c_4\sigma$ . Hence, we obtain the desired result

$$\begin{aligned} & \liminf_{p \rightarrow \infty} \inf_{\hat{z}_1} \sup_{(z, \mathcal{S}, \boldsymbol{\theta}) \in \mathcal{P}'(\gamma')} \mathbb{E}[\rho(\hat{z}_1, z_1)] \\ & \geq \liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \mathbb{E}[\rho(\hat{z}_1, z_1^*) | z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \geq 1. \end{aligned}$$

□

### 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  $K \geq 2$  and  $r \geq 2$ . 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

$$\mathcal{S} = \mathcal{S}(\gamma) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}^{\otimes K}. \quad (35)$$

Then, we consider the signal tensor family

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

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 (5). 1789
- (ii) For every  $\gamma \in \mathbb{R}$ ,  $\{\mathcal{X} - \mathbf{1}: \mathcal{X} \in \mathcal{P}_{\text{shifted}}(\gamma)\} \subset \mathcal{P}_{\text{nondegree}}(\gamma)$ , 1790 where  $\mathcal{P}_{\text{nondegree}}(\gamma)$  denotes the subfamily of rank-one tensor 1791 block model constructed in the proof of [13, Theorem 7]. 1792

The verification of the above two properties is provided in the end 1794 of this proof. 1795

Now, following the proof of [13, Theorem 7], when  $\gamma < -K/2$ , 1796 every polynomial-time algorithm estimator  $(\hat{\mathbf{M}}_k)_{k \in [K]}$  obeys 1797

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{nondegree}}(\gamma)} \mathbb{P}(\exists k \in [K], \hat{\mathbf{M}}_k \neq \mathbf{M}_k) \geq 1/2, \quad (36) \quad 1798$$

under the HPC Conjecture 1. The inequality (36) implies 1799

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{nondegree}}(\gamma)} \max_{k \in [K]} \mathbb{E}[\rho(\mathbf{z}_k, \hat{\mathbf{z}}_k)] \geq 1. \quad 1800$$

Based on properties (i)-(ii), we conclude that 1801

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}(\gamma)} \max_{k \in [K]} \mathbb{E}[\rho(\mathbf{z}_k, \hat{\mathbf{z}}_k)] \geq 1. \quad 1802$$

We complete the proof by verifying the properties (i)-(ii). For (i), 1803 we verify that the angle gap for the core tensor  $\mathcal{S}$  in (35) is on the 1804 order of  $\sigma p^{-\gamma/2}$ . Specifically, write  $\mathbf{1} = (1, 1, 1)$  and  $\mathbf{e} = (1, -1, 0)$ . 1805 We have 1806

$$\text{Mat}(\mathcal{S}) = \begin{bmatrix} \text{Vec}(\mathbf{1}^{\otimes K-1}) + \sigma p^{-\gamma/2} \text{Vec}\left(\mathbf{e}^{\otimes(K-1)}\right) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) - \sigma p^{-\gamma/2} \text{Vec}\left(\mathbf{e}^{\otimes(K-1)}\right) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) \end{bmatrix}. \quad 1807$$

Based on the orthogonality  $\langle \mathbf{1}, \mathbf{e} \rangle = 0$ , the minimal angle gap among 1808 rows of  $\text{Mat}(\mathcal{S})$  is 1809

$$\begin{aligned} \Delta_{\min}^2(\mathcal{S}) &\asymp \tan^2(\text{Mat}(\mathcal{S})_{1:}, \text{Mat}(\mathcal{S})_{3:}) \\ &= \left(\frac{\|\mathbf{e}\|_2}{\|\mathbf{1}\|_2}\right)^{2(K-1)} \sigma^2 d^{-\gamma} \\ &\asymp \sigma^2 d^{-\gamma}. \end{aligned} \quad 1810 \quad 1811 \quad 1812$$

Therefore, we have shown that  $\mathcal{P}_{\text{shifted}}(\gamma) = \mathcal{P}(\gamma)$ . Finally, the 1813 property (ii) follows directly by comparing the definition of  $\mathcal{S}$  in (35) 1814 with that in the proof of [13, Theorem 7]. 1815 □

### F. Proof of Theorem 4 and Proposition 1

*Proof of Theorem 4:* We prove Theorem 4 under the dTBM (1) 1817 with symmetric mean tensor, parameters  $(z, \mathcal{S}, \boldsymbol{\theta})$ , fixed  $r \geq 1, K \geq 1818 2$ , and i.i.d. noise. For the case  $r = 1$ , we have  $L(z^{(0)}, z) = 1819 0, \ell(z^{(0)}, z) = 0$  trivially. Hence, we focus on the proof of the first 1820 mode clustering  $z_1^{(0)}$  with  $r \geq 2$ ; the proofs for the other modes can 1821 be extended similarly. We drop the subscript  $k$  in the matricizations 1822  $\mathbf{M}_k, \mathbf{X}_k, \mathbf{S}_k$  and in the estimate  $z_1^{(0)}$ . We firstly show the proof with 1823 balanced  $\boldsymbol{\theta}$ . 1824

We firstly show the upper bound for misclustering error  $\ell(z^{(0)}, z)$ . 1825 First, by Lemma 1, there exists a positive constant such that 1826  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\min}$ . By the balance assumption 1827 on  $\boldsymbol{\theta}$  and Lemma 8, we have 1828

$$\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_I} \theta(i)^2 + 4 \sum_{i \in S} \theta(i)^2, \quad (37) \quad 1829$$

where 1830

$$S_0 = \{i: \|\hat{\mathbf{X}}_{i:}\| = 0\}, S = \{i \in S_0^c: \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\| \geq c_0 \Delta_{\min}/2\}. \quad 1831$$

1832 On one hand, note that for any set  $P \in [p]$ ,

$$\begin{aligned} 1833 \quad \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 &= \sum_{i \in P} \|\theta(i) \mathbf{S}_{z(i)} : (\Theta \mathbf{M})^{T, \otimes(K-1)}\|^2 \\ 1834 \quad &\geq \sum_{i \in P} \theta(i)^2 \min_{a \in [r]} \|\mathbf{S}_{a:}\|^2 \lambda_r^{2(K-1)}(\Theta \mathbf{M}) \\ 1835 \quad &\gtrsim \sum_{i \in P} \theta(i)^2 p^{K-1} r^{-(K-1)}, \end{aligned}$$

1836 where the last inequality follows Lemma 6, the assumption that  
1837  $\min_{i \in [p]} \theta(i) \geq c$ , and the constraint  $\min_{a \in [r]} \|\mathbf{S}_{a:}\| \geq c_3$  in  
1838 the parameter space (2). Thus, we have

$$1839 \quad \sum_{i \in P} \theta(i)^2 \lesssim \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 p^{-(K-1)} r^{K-1}. \quad (38)$$

1840 On the other hand, note that

$$\begin{aligned} 1841 \quad &\sum_{i \in S} \|\mathbf{X}_{i:}\|^2 \\ 1842 \quad &\leq 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 + 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \\ 1843 \quad &\leq \frac{8}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1844 \quad &\quad (40) \end{aligned}$$

$$\begin{aligned} 1845 \quad &\leq \frac{16}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \left[ \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \hat{\mathbf{X}}_{i:}^s\|^2 + \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 \right] \\ 1846 \quad &\quad + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1847 \quad &\leq \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (42) \end{aligned}$$

$$1848 \quad \leq \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (43)$$

$$1849 \quad \lesssim \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (44)$$

1850 where inequalities (39) and (41) follow from the triangle  
1851 inequality, (40) follows from the definition of  $S$ , (42) follows  
1852 from the update rule of  $k$ -means in Step 6 of Sub-algorithm 1,  
1853 (43) follows from Lemma 4, and the last inequality (44)  
1854 follows from Lemma 7. Also, note that

$$\begin{aligned} 1855 \quad \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 &= \sum_{i \in S_0} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \\ 1856 \quad &\leq \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1857 \quad &\lesssim (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (45) \end{aligned}$$

1858 where the equation follows from the definition of  $S_0$ . There-  
1859 fore, combining the inequalities (37), (38), (44), and (45),  
1860 we have

$$\begin{aligned} 1861 \quad &\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ 1862 \quad &\lesssim \left( \sum_{i \in S} \|\mathbf{X}_{i:}\|^2 + \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \right) p^{-(K-1)} r^{K-1} \\ 1863 \quad &\lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}} (p^{K/2} r + pr^2 + r^K). \quad (46) \end{aligned}$$

With the assumption that  $\min_{i \in [p]} \theta(i) \geq c$ , we finally obtain  
1864 the result  
1865

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \quad (46)$$

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

Without the balanced  $\theta$ , we have  
1866  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\mathbf{X}}$ . Replacing the definition  
1867 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  
1868 only  $r$  different values. We let  $\mathbf{X}_a^s = \mathbf{X}_{i:}^s$  for all  $i$  such that  
1869  $z(i) = a, a \in [r]$ .

Notice that

$$\|\mathbf{X}_{i:}\|^2 \gtrsim p^{K-1} r^{-(K-1)} \quad (47)$$

and

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

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

$$\begin{aligned} 1870 \quad &\sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1871 \quad &\lesssim \sum_{i \in [p]} \left( \|\mathbf{X}_{i:}\|^2 - \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 \right) \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1872 \quad &\lesssim \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1873 \quad &\lesssim \eta \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \mathbf{X}_{i:}^s\|^2 \\ 1874 \quad &\lesssim \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1875 \quad &\lesssim p^{K/2} r + pr^2 + r^K. \quad (47) \end{aligned}$$

Hence, we have

$$\begin{aligned} 1876 \quad \sum_{i \in [p]} \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1877 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} \sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1878 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} (p^{K/2} r + pr^2 + r^K), \quad (48) \end{aligned}$$

where the first inequality follows from the assumption  
1879  $\min_{i \in [p]} \theta(i) \geq c > 0$ , the second inequality follows from  
1880 the inequality (38), and the last inequality comes from the  
1881 inequality (47).

Next, we consider the following quantity,

$$\begin{aligned} 1882 \quad &\sum_{i \in [p]} \theta(i) \|\mathbf{X}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1883 \quad &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\mathbf{X}_{i:}^s - \hat{\mathbf{X}}_{i:}^s\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1884 \quad &\lesssim \sum_{i \in [p]} \frac{\theta(i)^2}{\|\mathbf{X}_{i:}\|^2} \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1885 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} (p^{K/2} r + pr^2 + r^K), \quad (49) \end{aligned}$$

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 (48). In addition, with Theorem 4 and the condition  $\text{SNR} \gtrsim p^{-K/2} \log p$ , 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

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

where the last inequality follows from the inequality (48).

Finally, we obtain

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

where the first inequality follows from Lemma 1, the third inequality follows from inequalities (49) and (50), and the last inequality follows from the assumption that  $\text{SNR} \geq \tilde{C}p^{-K/2} \log p$ .  $\square$

*Proof of Proposition 1:* Algorithm 2 shares the same algorithm strategy as Sub-algorithm 1 but with a different estimation of the mean tensor,  $\hat{\mathcal{X}}'$ . Hence, the proof of Proposition 1 follows the same proof idea with the proof of Theorem 4. Replacing the estimation  $\hat{\mathcal{X}}$  by  $\hat{\mathcal{X}}'$  in the proof of Theorem 4, we have

$$\begin{aligned} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ \lesssim \left( \sum_{i \in S} \|\mathbf{X}_{i:}\|^2 + \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \right) p^{-(K-1)} r^{K-1}. \end{aligned} \quad (51)$$

By inequalities (43) and (45), we have

$$\sum_{i \in S} \|\mathbf{X}_{i:}\|^2 \leq \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) \|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2, \quad (52)$$

$$\sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \leq \|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2. \quad (53)$$

Hence, it suffices to find the upper bound of the estimation error  $\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2$  to complete our proof. Note that the matricization  $\text{Mat}_{sq}(\mathcal{X}) \in \mathbb{R}^{p^{\lceil K/2 \rceil} \times p^{\lceil K/2 \rceil}}$  has  $\text{rank}(\text{Mat}_{sq}(\mathcal{X})) \leq r^{\lceil K/2 \rceil}$ , and Bernoulli random variables follow the sub-Gaussian distribution with bounded variance  $\sigma^2 = 1/4$ . Apply Lemma 9 to  $\mathbf{Y} = \text{Mat}_{sq}(\mathcal{Y})$ ,  $\mathbf{X} = \text{Mat}_{sq}(\mathcal{X})$ , and  $\hat{\mathbf{X}} = \text{Mat}_{sq}(\hat{\mathcal{X}}')$ . Then, with probability tending to 1 as  $p \rightarrow \infty$ , we have

$$\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2 = \|\text{Mat}_{sq}(\hat{\mathcal{X}}') - \text{Mat}_{sq}(\mathcal{X})\|_F^2 \lesssim p^{\lceil K/2 \rceil}. \quad (54)$$

Combining the estimation error (54) with inequalities (52), (53), and (51), we obtain

$$\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}} p^{\lceil K/2 \rceil}. \quad (55)$$

Replace the inequality (46) in the proof of Theorem 4 by inequality (55). With the the same procedures to obtain  $\ell(\hat{z}^{(0)}, z)$  and  $L(\hat{z}^{(0)}, z)$  for Theorem 4, we finish the proof of Proposition 1.  $\square$

#### 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

$$\sin(\mathbf{v}_1, \mathbf{v}_2) \leq \|\mathbf{v}_1^s - \mathbf{v}_2^s\| \leq \frac{2 \|\mathbf{v}_1 - \mathbf{v}_2\|}{\max(\|\mathbf{v}_1\|, \|\mathbf{v}_2\|)}. \quad (56)$$

*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

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = \sqrt{2(1 - \cos \alpha)} = 2 \sin \frac{\alpha}{2} \geq \sin \alpha, \quad (57)$$

where the equations follow 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

$$\text{Pad}_{\mathbf{w}}(\mathbf{a}) = [a_1 \circ \mathbf{w}_1, \dots, a_d \circ \mathbf{w}_d]^T, \quad (58)$$

where  $a_i \circ \mathbf{w}_i = [a_i w_{i1}, \dots, a_i w_{ip_i}]^T$ , for all  $i \in [d]$ . 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

$$\min_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2 \leq \|\text{Pad}_{\mathbf{w}}(\mathbf{a})\|^2 \leq \max_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2. \quad (59)$$

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

$$\text{Pad}_{\mathbf{w}}^{-1}(\text{Pad}_{\mathbf{w}}(\mathbf{a})) = \mathbf{a}. \quad (60)$$

**Lemma 5 (Angle for Weighted Padding Vectors):** Suppose that we have two nonzero vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ . Given the weight collection  $\mathbf{w}$ , we have

$$\begin{aligned} \frac{\min_{i \in [d]} \|\mathbf{w}_i\|}{\max_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}) &\stackrel{*}{\leq} \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \\ &\stackrel{**}{\leq} \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}). \quad (58) \end{aligned}$$

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

First, we prove the inequality  $**$  in (58). 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 (57) to vectors  $\mathbf{b}$  and  $\mathbf{a}^\perp$ .

Next, we prove inequality  $*$  in (58). With the decomposition of  $\text{Pad}_{\mathbf{w}}(\mathbf{b})$  and the inverse weighted padding operator, we have

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

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 (57) 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 (2) and assumption that  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$ , the singular values of  $\Theta M$  are bounded as

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

**Proof of Lemma 6:** Note that

$$(\Theta M)^T \Theta M = D,$$

with  $D = \text{diag}(D_1, \dots, D_r)$  where  $D_a = \|\theta_{z^{-1}(a)}\|^2, a \in [r]$ . By the definition of singular values, we have

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

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

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

where the last inequality follows from the constraint in parameter space (2). Finally, notice that

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

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

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

where  $C, c$  are some positive constants.

**Proof of Lemma 7:** See [13, 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  $a \in [r]$  with  $r \geq 2$ . 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 (6) such that  $\frac{\max_{a \in [r]} \|\theta_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\theta_{z^{-1}(a)}\|^2} = 1 + o(1)$ . Consider an arbitrary estimate  $\hat{z}$  with  $\hat{\mathbf{x}}_i = \hat{\mathbf{v}}_{\hat{z}(i)}$  for all  $i \in S$ . Then, if

$$\min_{a \neq b \in [r]} \|\mathbf{v}_a - \mathbf{v}_b\| \geq 2c, \quad (59)$$

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 4 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

$$R_1 = \{u \in [r] : C_u = \emptyset\},$$

$$R_2 = \{u \in [r] : C_u \neq \emptyset, \text{ for all } i, j \in C_u, \hat{z}(i) = \hat{z}(j)\},$$

$$R_3 = \{u \in [r] : C_u \neq \emptyset, \text{ there exist } i, j \in C_u, \hat{z}(i) \neq \hat{z}(j)\}.$$

2056 Note that  $\cup_{u \in [r]} C_u = S_0^c / S^c$  and  $C_u \cap C_v = \emptyset$  for any  $u \neq v$ .  
 2057 Suppose there exist  $i \in C_u$  and  $j \in C_v$  with  $u \neq v \in [r]$  and  
 2058  $\hat{z}(i) = \hat{z}(j)$ . Then we have

2059  $\|\mathbf{v}_{z(i)} - \mathbf{v}_{z(j)}\| \leq \|\mathbf{v}_{z(i)} - \hat{\mathbf{x}}_i\| + \|\mathbf{v}_{z(j)} - \hat{\mathbf{x}}_j\| < 2c,$

2060 which contradicts to the assumption (59). Hence, the estimates  
 2061  $\hat{z}(i) \neq \hat{z}(j)$  for the nodes  $i \in C_u$  and  $j \in C_v$  with  $u \neq v$ .  
 2062 By the definition of  $R_2$ , the nodes in  $\cup_{u \in R_2} C_u$  have the same  
 2063 assignment with  $z$  and  $\hat{z}$ . Then, we have

2064  $\min_{\pi \in \Pi} \sum_{i: \hat{z}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_0} \theta(i)^2 + \sum_{i \in S} \theta(i)^2 + \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2.$

2065 We only need to bound  $\sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2$  to finish the proof.  
 2066 Note that every  $C_u$  with  $u \in R_3$  contains at least two  
 2067 nodes assigned to different clusters by  $\hat{z}$ . Then, we have  
 2068  $|R_2| + 2|R_3| \leq r$ . Since  $|R_1| + |R_2| + |R_3| = r$ , we have  
 2069  $|R_3| \leq |R_1|$ . Hence, we obtain

2070 
$$\begin{aligned} \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2 &\leq |R_3| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ 2071 &\leq |R_1| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ 2072 &\leq \frac{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \sum_{i \in \cup_{u \in R_1} z^{-1}(u)} \theta(i)^2 \\ 2073 &\leq 2 \sum_{i \in S} \theta(i)^2, \end{aligned}$$

2074 where the last inequality holds by the balanced assumption on  
 2075  $\boldsymbol{\theta}$  when  $p$  is large enough, and the fact that  $\cup_{u \in R_1} z^{-1}(u) \subset S$ .  
 2076  $\square$

2077 *Lemma 9 (Low-Rank Matrix Estimation):* Let  $\mathbf{Y} = \mathbf{X} +$   
 2078  $\mathbf{E} \in \mathbb{R}^{m \times n}$ , where  $n > m$  and  $\mathbf{E}$  contains independent mean-  
 2079 zero sub-Gaussian entries with bounded variance  $\sigma^2$ . Suppose  
 2080  $\text{rank}(\mathbf{X}) = r$ . Consider the least square estimator

2081 
$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}' \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{X}') \leq r} \|\mathbf{X}' - \mathbf{Y}\|_F^2.$$

2082 There exist positive constants  $C_1, C_2$  such that

2083 
$$\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 \leq C_1 \sigma^2 nr,$$

2084 with probability at least  $1 - \exp(-C_2 nr)$ .

2085 *Proof of Lemma 9:* Note that  $\|\hat{\mathbf{X}} - \mathbf{Y}\|_F^2 \leq \|\mathbf{X} - \mathbf{Y}\|_F^2$  by  
 2086 the definition of least square estimator.

2087 We have

2088 
$$\begin{aligned} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 \\ 2089 &\leq 2 \langle \hat{\mathbf{X}} - \mathbf{X}, \mathbf{Y} - \mathbf{X} \rangle \\ 2090 &\leq 2 \|\hat{\mathbf{X}} - \mathbf{X}\|_F \sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{T}, \mathbf{Y} - \mathbf{X} \rangle \quad (60) \end{aligned}$$

2091 with probability at least  $1 - \exp(-C_2 nr)$ , where the second  
 2092 inequality follows by re-arrangement.

2093 Consider the SVD for matrix  $\mathbf{T} = \mathbf{U} \Sigma \mathbf{V}^T$  with orthogonal  
 2094 matrices  $\mathbf{U} \in \mathbb{R}^{m \times 2r}, \mathbf{V} \in \mathbb{R}^{n \times 2r}$  and diagonal matrix  $\Sigma \in$

$\mathbb{R}^{2r \times 2r}$ . We have

2095 
$$\begin{aligned} &\sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{T}, \mathbf{Y} - \mathbf{X} \rangle \\ 2096 &= \sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{U} \Sigma, \mathbf{EV} \rangle \\ 2097 &= \sup_{\mathbf{v} \in \mathbb{R}^{2nr}} \mathbf{v}^T \mathbf{e} \leq C \sigma \sqrt{nr}, \end{aligned} \quad (61) \quad 2098$$

2099 with probability  $1 - \exp(-C_2 nr)$ , where  $C, C_2$  are two  
 2100 positive constants, the vectorization  $\mathbf{e} = \text{Vec}(\mathbf{EV}) \in \mathbb{R}^{2nr}$   
 2101 has independent mean-zero sub-Gaussian entries with bounded  
 2102 variance  $\sigma^2$  due to the orthogonality of  $\mathbf{V}$ , and the last  
 2103 inequality follows from [36, Theorem 1.19].  
 2104 Combining inequalities (60) and (61), we obtain the desired  
 2105 conclusion.  $\square$

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

2106 *Proof of Theorem 2 (Achievability) and Theorem 5:* The  
 2107 proofs of Theorem 2 (Achievability) and Theorem 5 share the  
 2108 same idea. We prove the contraction step by step. In each  
 2109 step, we show the specific procedures for the algorithm loss  
 2110 and address the MLE loss by stating the difference.

2111 We consider dTBM (1) with symmetric mean tensor, param-  
 2112 eters  $(z, \mathcal{S}, \boldsymbol{\theta})$ , fixed  $r \geq 1, K \geq 2$ , and i.i.d. noise. Let  
 2113  $(\hat{z}, \hat{\mathcal{S}}, \hat{\boldsymbol{\theta}})$  denote the MLE in (9), and  $(z_k^{(0)}, \mathcal{S}^{(0)}, \boldsymbol{\theta}_k^{(0)})$  denote  
 2114 parameters related to the initialization. For the case  $r = 1$ ,  
 2115  $\ell(z_k^{(t)}, z) = 0$  trivially for all  $t \geq 0, k \in [k]$ . Hence, we focus  
 2116 on the proof of the first mode clustering  $z_1^{(t+1)}$  with  $r \geq 2$ ; the  
 2117 extension for other modes can be obtained similarly. We drop  
 2118 the subscript  $k$  in the matricizations  $\boldsymbol{\Theta}, \mathbf{M}_k, \mathbf{S}_k, \mathbf{X}_k$  and in  
 2119 estimates  $z_k^{(0)}, z_k^{(t+1)}, z_k^{(t)}$  for ease of the notation. Without  
 2120 loss of generality, we assume that the variance  $\sigma = 1$ , and that  
 2121 the identity permutation minimizes the initial misclustering  
 2122 error; i.e.,  $\pi^{(0)} = \arg \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1} \{z^{(0)}(i) \neq \pi \circ z(i)\}$   
 2123 and  $\pi^{(0)}(a) = a$  for all  $a \in [r]$ , and so for  $\hat{z}$ .

2124 *Step 1 (Notation and Conditions):* We first introduce addi-  
 2125 tional notations and the necessary conditions used in the proof.  
 2126 We will verify that the conditions hold in our context under  
 2127 high probability in the last step of the proof.

#### Notation:

2128 1) Projection. We use  $\mathbf{I}_d$  to denote the identity matrix of  
 2129 dimension  $d$ . For a vector  $\mathbf{v} \in \mathbb{R}^d$ , let  $\text{Proj}(\mathbf{v}) \in \mathbb{R}^{d \times d}$  denote  
 2130 the projection matrix to  $\mathbf{v}$ . Then,  $\mathbf{I}_d - \text{Proj}(\mathbf{v})$  is the projection  
 2131 matrix to the orthogonal complement  $\mathbf{v}^\perp$ .

2132 2) We define normalized membership matrices

2133 
$$\mathbf{W} = \mathbf{M} \left( \text{diag}(\mathbf{1}_p^T \mathbf{M}) \right)^{-1}, \quad \mathbf{W}^{(t)} = \mathbf{M}^{(t)} \left( \text{diag}(\mathbf{1}_p^T \mathbf{M}^{(t)}) \right)^{-1},$$

2134 weighted normalized membership matrices

2135 
$$\mathbf{P} = \boldsymbol{\Theta} \mathbf{M} (\text{diag}(\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2, \dots, \|\boldsymbol{\theta}_{z^{-1}(r)}\|^2))^{-1},$$

2136 
$$\hat{\mathbf{P}} = \hat{\boldsymbol{\Theta}} \hat{\mathbf{M}} (\text{diag}(\|\hat{\boldsymbol{\theta}}_{z^{-1}(1)}\|^2, \dots, \|\hat{\boldsymbol{\theta}}_{z^{-1}(r)}\|^2))^{-1},$$

2137 and the dual normalized and dual weighted normalized mem-  
 2138 bership matrices

2139 
$$\mathbf{V} = \mathbf{W}^{\otimes(K-1)}, \quad \mathbf{V}^{(t)} = \left( \mathbf{W}^{(t)} \right)^{\otimes(K-1)},$$

2140 
$$\hat{\mathbf{Q}} = \mathbf{P}^{\otimes K-1}, \quad \hat{\mathbf{Q}} = \hat{\mathbf{P}}^{\otimes K-1}.$$

2141 Also, let  $\mathbf{B} = (\boldsymbol{\Theta} \mathbf{M})^{\otimes(K-1)}, \hat{\mathbf{B}} = (\hat{\boldsymbol{\Theta}} \hat{\mathbf{M}})^{\otimes(K-1)}$ . By the  
 2142 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.,

$$\begin{aligned}\mathcal{S}^{(t)} &= \mathcal{Y} \times_1 (\mathbf{W}^{(t)})^T \times_2 \cdots \times_K (\mathbf{W}^{(t)})^T, \\ \tilde{\mathcal{S}} &= \mathcal{Y} \times_1 \mathbf{W}^T \times_2 \cdots \times_K \mathbf{W}^T, \\ \hat{\mathcal{S}} &= \mathcal{Y} \times_1 \hat{\mathbf{P}}^T \times_2 \cdots \times_K \hat{\mathbf{P}}^T, \\ \bar{\mathcal{S}} &= \mathcal{Y} \times_1 \mathbf{P}^T \times_2 \cdots \times_K \mathbf{P}^T.\end{aligned}$$

4) We define the matricizations of tensors

$$\begin{aligned}\mathbf{S} &= \text{Mat}(\mathcal{S}), \quad \mathbf{Y} = \text{Mat}(\mathcal{Y}), \quad \mathbf{X} = \text{Mat}(\mathcal{X}), \quad \mathbf{E} = \text{Mat}(\mathcal{E}), \\ \mathbf{S}^{(t)} &= \text{Mat}(\mathcal{S}^{(t)}), \quad \hat{\mathcal{S}} = \text{Mat}(\hat{\mathcal{S}}), \quad \tilde{\mathcal{S}} = \text{Mat}(\tilde{\mathcal{S}}), \quad \bar{\mathcal{S}} = \text{Mat}(\bar{\mathcal{S}}).\end{aligned}$$

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 (2), 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

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

We also define the loss for oracle and weighted oracle estimators

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

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

Then we introduce the necessary conditions in Condition 1.

*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

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

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

$$\|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_b]_s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_{z(i)}]_s\|^2. \quad (67)$$

Note that the event (67) 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 (67) yields the decomposition

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

where

$$\begin{aligned}F_{ib}^{(t)} &= 2 \langle \mathbf{E}_{i:} \mathbf{V}^{(t)}, ([\tilde{\mathcal{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s) - ([\tilde{\mathcal{S}}_b]_s - [\mathbf{S}_b]_s) \rangle \\ &\quad + 2 \langle \mathbf{E}_{i:} (\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathcal{S}}_{z(i)}]_s - [\tilde{\mathcal{S}}_b]_s \rangle,\end{aligned}$$

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

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

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

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

Note that

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

where the first inequality follows from the property of eigenvalues; the last inequality follows from Lemma 6, Lemma 10, 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  $\|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|$  (72) into the inequality (71) gives

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

where

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

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

Taking the weighted summation of (73) 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

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2. \quad (74)$$

Similarly to  $\xi$  in (74), 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  $\hat{\theta}$  satisfies  $\hat{\theta}(i) = \langle \mathbf{Y}_{i:}, \hat{\mathbf{A}}_{\hat{z}(i):} \rangle / \|\hat{\mathbf{A}}_{\hat{z}(i):}\|_F^2$  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

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

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

$$\|\mathbf{EV}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}} \left( p^{1/2} + r^{(K-1)/2} \right), \quad \|\mathbf{EV}\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}}, \quad \|\mathbf{W}_a^T \mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}, \text{ for all } a \in [r], \quad (62)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_\sigma \lesssim \left( \sqrt{r^{K-1}} + K\sqrt{pr} \right), \quad (63)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_F \lesssim \left( \sqrt{pr^{K-1}} + K\sqrt{pr} \right), \quad (64)$$

$$\xi \leq \exp \left( -M \frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad \xi' \lesssim \exp \left( -\frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad (65)$$

$$L^{(t)} \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \quad \text{for } t = 0, 1, \dots, T, \quad L(\hat{z}) \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \quad (66)$$

where  $M$  is a positive universal constant in inequality (84),  $\bar{C}, \tilde{C}$  are positive universal constants in the proof of Theorem 4 and assumption  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$ , respectively. Further, inequality (62) holds by replacing  $\mathbf{V}$  to  $\mathbf{V}^{(t)}, \mathbf{Q}, \hat{\mathbf{Q}}$  and  $\mathbf{W}_{:a}$  to  $\mathbf{W}_{:a}^{(t)}, \mathbf{P}_{:a}^T, \hat{\mathbf{P}}_{:a}^T$  when initialization condition (66) holds.

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

$$A_{ib}' = \mathbb{1} \left\{ \hat{z}(i) = b, \langle \mathbf{E}_{i:}, [\tilde{\mathbf{A}}_{z(i):}]^s - [\tilde{\mathbf{A}}_{b:}]^s \rangle \right. \quad (2232)$$

$$\left. \leq -\frac{m'}{4} \sqrt{\frac{p^{K-1}}{r^{K-1}}} \|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2 \right\}, \quad (2233)$$

$$B_{ib}' = \mathbb{1} \left\{ \hat{z}(i) = b, -\frac{1}{2} \|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2 \right. \quad (2234)$$

$$\left. \leq \sqrt{\frac{r^{K-1}}{(m')^2 p^{K-1}}} \hat{F}_{ib} + \hat{G}_{ib} + \hat{H}_{ib} \right\} \quad (2235)$$

with terms

$$\hat{F}_{ib} = 2 \left\langle \mathbf{E}_{i:}, ([\tilde{\mathbf{A}}_{z(i):}]^s - [\hat{\mathbf{A}}_{a:}]^s) - ([\tilde{\mathbf{A}}_{b:}]^s - [\hat{\mathbf{A}}_{b:}]^s) \right\rangle, \quad (2237)$$

$$\hat{G}_{ib} = \left( \|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{z(i):}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \right) \quad (2238)$$

$$- \left( \|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{b:}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \right), \quad (2239)$$

$$\hat{H}_{ib} = \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \quad (2240)$$

$$+ \|\mathbf{A}_{z(i):}^s - \mathbf{A}_{b:}^s\|_F^2. \quad (2241)$$

*Step 3 (Derivation of Contraction Inequality):* In this step we derive the upper bound of  $\zeta_{ib}$  and obtain the contraction inequality (24). We show the analysis in the following one-column box for a better presentation.

*Step 4 (Verification of Condition 1):* Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (62), (63), and (64) describe the property of the sub-Gaussian noise tensor  $\mathcal{E}$ , and the readers can find the proof directly in [13, Step 5, Proof of Theorem 2]. The initial condition (66) for MLE is satisfied by Lemma 13. Here, we include only the verification of inequalities (65) and (66) for algorithm estimators.

Now, we verify the oracle loss condition (65). Recall the definition of  $\xi$ ,

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ \langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \rangle \right. \quad (2256)$$

$$\leq -\frac{\theta(i)m}{4} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \left\{ \cdot \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right\}$$

where

2267

$$P_1 = \mathbb{P} \left( \langle e_i, [S_{z(i)}]_s^s - [S_b]_s^s \rangle \leq -\frac{\theta(i)m}{8} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right), \quad 2268$$

$$P_2 = \mathbb{P} \left( \langle e_i, [\tilde{S}_{z(i)}]_s^s - [S_{z(i)}]_s^s \rangle \leq -\frac{\theta(i)m}{16} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right), \quad 2269$$

$$P_3 = \mathbb{P} \left( \langle e_i, [S_b]_s^s - [\tilde{S}_b]_s^s \rangle \leq -\frac{\theta(i)m}{16} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right). \quad 2270$$

2258 Let  $e_i = E_{i,:} V$  denote the aggregated noise vector for all  $i \in [p]$ , and  $e_i$ 's are independent zero-mean sub-Gaussian vector  
2259 in  $\mathbb{R}^{r^{K-1}}$ . The entries in  $e_i$  are independent zero-mean sub-  
2260 Gaussian variables with sub-Gaussian norm upper bounded by  
2261  $m_1 \sqrt{r^{K-1}/p^{K-1}}$  with some positive constant  $m_1$ . We have  
2262 the probability inequality  
2263

$$\begin{aligned} 2265 \mathbb{P} \left( \langle e_i, [\tilde{S}_{z(i)}]_s^s - [\tilde{S}_b]_s^s \rangle \leq -\frac{\theta(i)m}{4} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right) \\ 2266 \leq P_1 + P_2 + P_3, \end{aligned}$$

2264 For  $P_1$ , notice that the inner product  $\langle e_j, S_{z(j)}^s - S_b^s \rangle$  is  
2265 a sub-Gaussian variable with sub-Gaussian norm bounded  
2266 by  $m_2 \sqrt{r^{K-1}/p^{K-1}} \|S_{z(i)}^s - S_b^s\|$  with some positive  
2267

*Step 3:* Choose the constant  $\tilde{C}$  in the condition  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$  that satisfies the condition of Lemma 11, inequalities (98), and (102). Note that

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

where the first inequality follows from the inequality (89) in Lemma 11, and the last inequality follows from the assumption that  $\min_{i \in [p]} \theta(i) \geq c > 0$ . Following [13, Step 4, Proof of Theorem 2] and Lemma 11, we have

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

for a positive universal constant  $C$  and

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

where the last inequality follows from the definition of  $L^{(t)}$  and the constraint of  $\theta$  in parameter space (2). For  $\tilde{C}$  also satisfies

$$\frac{C_0 \bar{C}}{cm^2 \tilde{C}^2} \leq \frac{1}{512}, \quad (75)$$

we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta_{ib}^{(t)} \leq \frac{1}{8} L^{(t+1)} + \frac{1}{4} L^{(t)}. \quad (76)$$

Plugging the inequality (76) into the decomposition (74), we obtain the contraction inequality

$$L^{(t+1)} \leq \frac{3}{2} \xi + \frac{1}{2} L^{(t)}, \quad (77)$$

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

Therefore, with  $\tilde{C}$  satisfying inequalities (75), (98) and (102), we obtain the conclusion in Theorem 5 via inequality (77) combining the inequality (65) in Condition 1 and Lemma 2.

We also have the contraction inequality for MLE.

Following the same derivation of (77) with the upper bound of  $\hat{F}_{ib}, \hat{G}_{ib}, \hat{H}_{ib}$  in Lemma 12, we also have

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

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

constant  $m_2$ . Then, by Chernoff bound, we have

$$P_1 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(j)}]_s - [\mathbf{S}_b]_s\|^2\right). \quad (78)$$

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

$$P_2 \leq P_{21} + P_{22}, \quad (79)$$

where with some positive constant  $t > 0$ ,

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

For  $P_{21}$ , note that the term  $\mathbf{W}_{:z(i)}^T \mathbf{EV} = \frac{\sum_{j \neq i, j \in [p]} \mathbb{1}\{z(j)=z(i)\} e_j}{\sum_{j \in [p]} \mathbb{1}\{z(j)=z(i)\}}$  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

$$\begin{aligned} P_{21} &\leq \mathbb{P}\left(t \|\mathbf{S}_{z(i)}\| \leq \|\tilde{\mathbf{S}}_{z(i)} - \mathbf{S}_{z(i)}\|\right) \\ &\leq \mathbb{P}\left(c_3 t \leq \|\mathbf{W}_{:z(i)}^T \mathbf{EV}\|\right) \\ &\lesssim \exp\left(-\frac{p^K t^2}{r^K}\right), \end{aligned} \quad (80)$$

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 (2), 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

$$P_{22} \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1} t^2} \|[\mathbf{S}_{z(j)}]_s - [\mathbf{S}_b]_s\|^4\right). \quad (81)$$

We take  $t = \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|$  in  $P_{21}$  and  $P_{22}$ , and plug the inequalities (80) and (81) into to the upper bound for  $P_2$  in (79). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right). \quad (82)$$

Combining the upper bounds (78) and (82) gives

$$\begin{aligned} \mathbb{P}\left(\left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right) \\ \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right). \end{aligned} \quad (83)$$

Hence, we have

$$\begin{aligned} \mathbb{E}\xi &= \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{P}\left\{\left\langle \mathbf{E}_i \mathbf{V}, [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle\right. \\ &\quad \left.\leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right\} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \\ &\lesssim \frac{1}{p} \sum_{i \in [p]} \theta(i) \max_{i \in [p], b \in [r]} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 \\ &\quad \cdot \exp\left(\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2\right) \\ &\leq \exp\left(-M \frac{p^{K-1}}{r^{K-1}} \Delta_{\min}^2\right), \end{aligned} \quad (84)$$

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 (83).

By Markov's inequality, we have

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

and thus the condition (65) 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

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

where the first inequality follows from the property in later inequality (105). We also have

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

Finally, we verify the bounded loss condition (66) 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

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

Therefore, the condition (66) holds for  $t = 0$ . Assume that the condition (66) also holds for all  $t \leq t_0$ . Then, by the decomposition (77), we have

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

where the second inequality follows from the condition (65) and the last inequality follows from the assumption that  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ . Thus, the condition (66) holds for  $t_0+1$ , and the condition (66) is proved by induction.  $\square$

### Useful Lemmas for the Proof of Theorem 5:

*Lemma 10 (Singular-Value Property of Membership Matrices):* Under the setup of Theorem 5, suppose that the condition (66) holds. Then, for all  $a \in [r]$ , we have  $|(\mathbf{z}^{(t)})^{-1}(a)| \asymp p/r$ . Moreover, we have

$$\lambda_r(\mathbf{M}) \asymp \|\mathbf{M}\|_\sigma \asymp \sqrt{p/r}, \quad \lambda_r(\mathbf{W}) \asymp \|\mathbf{W}\|_\sigma \asymp \sqrt{r/p}, \\ \lambda_r(\mathbf{P}) \asymp \|\mathbf{P}\|_\sigma \asymp \min_{a \in [r]} \|\theta_{z^{-1}(a)}\|^{-1} \lesssim \sqrt{r/p}. \quad (85)$$

The inequalities (85) also hold by replacing  $\mathbf{M}$  and  $\mathbf{W}$  to  $\mathbf{M}^{(t)}$  and  $\mathbf{W}^{(t)}$  respectively. Further, we have

$$\lambda_r(\mathbf{W}\mathbf{W}^T) \asymp \|\mathbf{W}\mathbf{W}^T\|_\sigma \asymp r/p, \quad (86)$$

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

*Proof of Lemma 10:* The proof for the inequality (85) for  $\mathbf{M}, \mathbf{W}$  can be found in [13, 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 (86), note that for all  $k \in [r]$ ,

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

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 11 (Upper Bound for  $F_{ib}^{(t)}, G_{ib}^{(t)}$  and  $H_{ib}^{(t)}$ ):* Under the Condition 1 and the setup of Theorem 5 with fixed  $r \geq 2$ , 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 (98) and (102). As  $p \rightarrow \infty$ , we have

$$\begin{aligned} \max_{i \in [p]} \max_{b \neq z(i)} \frac{(F_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \\ \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2} \|\mathbf{E}_{i:}\mathbf{V}\|^2 + \left(1 + \frac{rL^{(t)}}{\Delta_{\min}^2}\right) \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2, \quad (87) \end{aligned}$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(G_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{512} (\Delta_{\min}^2 + L^{(t)}), \quad (88)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{|H_{ib}^{(t)}|}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{4}. \quad (89)$$

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

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(\hat{F}_{ib})^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \lesssim p^{K-1} \frac{rL(\hat{z})}{\Delta_{\min}^2} \quad (2381)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(\hat{G}_{ib})^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \leq \frac{1}{512} (\Delta_{\min}^2 + L(\hat{z})), \quad (2382)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{|\hat{H}_{ib}|}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \leq \frac{1}{4}. \quad (2383)$$

*Proof of Lemma 11:* We prove the the first three inequalities in Lemma 11 separately.

1) Upper bound for  $F_{ib}^{(t)}$ , i.e., inequality (87). Recall the definition of  $F_{ib}^{(t)}$ ,

$$\begin{aligned} F_{ib}^{(t)} &= 2 \left\langle \mathbf{E}_{i:}\mathbf{V}^{(t)}, \left([\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s\right) - \left([\tilde{\mathbf{S}}_b]_s - [\mathbf{S}_b]_s\right) \right\rangle \\ &\quad + 2 \left\langle \mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle. \end{aligned} \quad (2388)$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned} &(F_{ib}^{(t)})^2 \\ &\leq 8 \left( \left\langle \mathbf{E}_{i:}\mathbf{V}^{(t)}, \left([\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s\right) - \left([\tilde{\mathbf{S}}_b]_s - [\mathbf{S}_b]_s\right) \right\rangle \right)^2 \\ &\quad + 8 \left( \left\langle \mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle \right)^2 \\ &\leq 8 \left( \|\mathbf{E}_{i:}\mathbf{V}\|^2 + \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2 \right) \max_{a \in [r]} \|[\tilde{\mathbf{S}}_a]_s - [\mathbf{S}_a]_s\|^2 \\ &\quad + \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2 \|[\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\|. \end{aligned} \quad (90)$$

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

$$\begin{aligned} \|[\tilde{\mathbf{S}}_a]_s - [\mathbf{S}_a]_s\|^2 &= \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 \\ &\leq 2 \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]_s\|^2 \\ &\quad + 2 \|[\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 \\ &\lesssim \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\ &\lesssim rL^{(t)} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\ &\lesssim rL^{(t)}, \end{aligned} \quad (91)$$

where the second inequality follows from the inequalities (108) and (109) in Lemma 12, the third inequality follows from the condition (66) 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

$$\begin{aligned} &\|[\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\|^2 \\ &= \|[\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s + [\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s + [\mathbf{S}_b]_s - [\tilde{\mathbf{S}}_b]_s\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 + \max_{a \in [r]} \|[\mathbf{S}_a]_s - [\tilde{\mathbf{S}}_a]_s\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 + \max_{a \in [r]} \frac{1}{\|\mathbf{S}_a\|^2} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2, \end{aligned} \quad (92)$$

where the second inequality follows from Lemma 4, and the last inequality follows from the assumptions on  $\|S_{a:}\|$  in the parameter space (2), the inequality (62) in Condition 1 and the assumption  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ .

Therefore, we finish the proof of inequality (87) by plugging the inequalities (91) and (92) into the upper bound (90).

2) Upper bound for  $G_{ib}^{(t)}$ , i.e., inequality (88). By definition of  $G_{ib}^{(t)}$ , we rearrange terms and obtain

$$\begin{aligned} G_{ib}^{(t)} &= \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &\quad - \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &= 2 \left\langle [\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s, \left( [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right) \right. \\ &\quad \left. - \left( [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right) \right\rangle \\ &= G_1 + G_2 - G_3, \end{aligned} \tag{93}$$

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

$$\begin{aligned} |G_1|^2 &\leq \left| \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \right|^2 \\ &\leq \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^4 \\ &\leq C^4 \frac{r^4}{\Delta_{\min}^4} (L^{(t)})^4 + \frac{r^2 r^{4K} + p^2 r^{2K+4}}{p^{2K}} \frac{(L^{(t)})^2}{\Delta_{\min}^4} \\ &\leq C^4 \frac{\bar{C}}{\tilde{C}^3} \left( \Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)} \right), \end{aligned} \tag{94}$$

where the third inequality follows from the inequality (110) in Lemma 12 and the last inequality follows from the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) 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

$$\begin{aligned} |G_2|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\quad \cdot \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \\ &\leq \frac{2}{\|\mathbf{W}_{z(i):}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2 \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \end{aligned}$$

$$\leq C' \frac{r^{2K-1} + K p r^{K+1}}{p^K} \tag{2447}$$

$$\cdot \left( \frac{r^2}{\Delta_{\min}^2} (L^{(t)})^2 + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) \tag{2448}$$

$$\leq \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)}, \tag{2449}$$

where  $C'$  is a positive universal constant, the second inequality follows from Lemma 4, the third inequality follows from the inequality (63) in Condition 1, the inequalities (110) and (129) in the proof of Lemma 12, and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) in Condition 1.

For  $G_3$ , note that by triangle inequality

$$\begin{aligned} &\|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq \|[\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]\|^2 \\ &\leq \|[\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2}], \end{aligned} \tag{96}$$

where the last inequality follows from the inequality (128) in the proof of Lemma 12 and  $C$  is a positive constant. Then we have

$$\begin{aligned} |G_3|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq 2 \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \right. \\ &\quad \left. + \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq C^2 \left( \|[\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2}] \right. \\ &\quad \left. \cdot \left( \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) + \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)} \right) \\ &\leq \frac{C^2 \bar{C}^2}{\tilde{C}} \|[\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 (\Delta_{\min}^2 + L^{(t)}) \\ &\quad + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} (\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)})], \end{aligned} \tag{97}$$

where the third inequality follows from the same procedure to derive (94) and (95), and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) in Condition 1.

Choose the  $\tilde{C}$  such that

$$3 \left( C^4 \frac{\bar{C}}{\tilde{C}^3} + \frac{C'}{\tilde{C}^2} + \frac{C^2 \bar{C}^2}{\tilde{C}} + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} \right) \leq \frac{1}{512}. \tag{98}$$

Then, we finish the proof of inequality (88) by plugging the inequalities (94), (95), and (97) into the upper bound (93).

3) Upper bound for  $H_{ib}^{(t)}$ , i.e., the inequality (89). By definition of  $H_{ib}$ , we rearrange terms and obtain

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

where

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

For  $H_1$ , we have

$$\begin{aligned} |H_1| &\leq \frac{4 \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2}{\|\mathbf{W}_{:z(i)}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \\ &\leq \frac{r^{2K-1} + K p r^{K+1}}{p^K} \\ &\leq \tilde{C}^{-2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \end{aligned} \quad (99)$$

following the derivation of  $G_2$  in inequality (95) and the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$ .

For  $H_2$ , by the inequality (96), we have

$$\begin{aligned} |H_2| &\lesssim 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \\ &\lesssim \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \\ &\leq C \frac{\bar{C}^2}{\tilde{C}^2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \end{aligned} \quad (100)$$

where the last inequality follows from the condition (66) in Condition 1.

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

$$\begin{aligned} |H_3| &\lesssim \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\| |H_1|^{1/2} \\ &\leq 2 \tilde{C}^{-1} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \end{aligned} \quad (101)$$

following the inequalities (96) and (99).

Choose  $\tilde{C}$  such that

$$\tilde{C}^{-2} + C \frac{\bar{C}^2}{\tilde{C}^2} + \tilde{C}^{-1} \leq \frac{1}{4}. \quad (102)$$

Therefore, we finish the proof of inequality (89) combining inequalities (99), (100), and (101).

Next, we Show the Upper Bounds for  $\hat{F}_{ib}$ ,  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ . By Lemma 1, we have

$$\|\mathbf{S}_{a:}^s - \mathbf{S}_{b:}^s\| = (1 + o(1)) \|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|. \quad (2517)$$

Also, notice that the matrix product of  $\mathbf{B}^T$  corresponds to the padding operation in Lemma 5, and the padding weights are balanced such that  $\|\mathbf{v} \mathbf{B}\| = (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{v}\|$  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

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = (1 + o(1)) \|[\mathbf{v}_1 \mathbf{B}^T]^s - [\mathbf{v}_2 \mathbf{B}^T]^s\|. \quad (103) \quad (2523)$$

The equation (103) also holds for  $\hat{\mathbf{B}}^T$ .

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

$$\begin{aligned} \|\mathbf{A}_{i:} \hat{\mathbf{Q}}\| &= \|\mathbf{S}_{z(i):} \mathbf{B}^T \hat{\mathbf{Q}}\| \\ &= \|\mathbf{S}_{z(i):} \hat{\mathbf{D}}^{\otimes(K-1)}\| \\ &= (1 + o(1)) \|\mathbf{S}_{z(i):}\| \\ &= (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{-(K-1)/2} \|\mathbf{A}_{i:}\|, \end{aligned} \quad (104) \quad (2526-2529)$$

where the third inequality follows from the singular property of MLE confusion matrix (135) and the last inequality follows from the fact that  $\mathbf{A}_{i:} = \mathbf{S}_{z(i):} \mathbf{B}^T$  and Lemma 10. 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

$$\begin{aligned} &\|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\| \\ &= \left\| \frac{\mathbf{A}_{z(i):} \hat{\mathbf{Q}}}{\|\mathbf{A}_{z(i):} \hat{\mathbf{Q}}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|} \right\| \\ &= (1 + o(1)) \left\| \frac{\mathbf{A}_{z(i):}}{\|\mathbf{A}_{z(i):}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\|} \right\| \\ &= (1 + o(1)) \|[\mathbf{X}_i]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\| \end{aligned} \quad (105) \quad (2538-2541)$$

where the second equation follows from (104),  $\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\| = (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|$ , 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

$$\begin{aligned} (\hat{F}_{ib})^2 &\leq \|\mathbf{E}_{i:}\|^2 \|[\bar{\mathbf{A}}_{a:}]^s - [\hat{\mathbf{A}}_{a:}]^s\|^2 \\ &\leq \|\mathbf{E}_{i:}\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T]^s - [\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| \right. \\ &\quad \quad \left. + \|[\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s - [\hat{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| \right]^2 \\ &\lesssim \|\mathbf{E}_{i:}\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T \hat{\mathbf{Q}}]^s - [\bar{\mathbf{S}}_{a:}]^s\| + \|[\bar{\mathbf{S}}_{a:}]^s - [\hat{\mathbf{S}}_{a:}]^s\| \right]^2. \end{aligned} \quad (2550-2553)$$

Following similar derivations in inequalities (91), (92), and the upper bound for  $J_1$  in the proof of Lemma 12, respectively,

2556 we have

$$2557 \quad \|[\bar{S}_{a:}]^s - [\hat{S}_{a:}]^s\| \lesssim rL(\hat{z}), \quad \|[\bar{S}_{a:}]^s - [\bar{S}_{b:}]^s\| \lesssim \|S_{a:}^s - S_{b:}^s\|^2, \\ 2558 \quad \text{and}$$

$$2559 \quad \|[\bar{S}_a \mathbf{B}^T \hat{\mathbf{Q}}]^s - [\bar{S}_{a:}]^s\| \lesssim L(\hat{z}).$$

2560 We then obtain the upper bound for  $\hat{F}_{ib}$  by noticing that  
2561  $\|\mathbf{E}_i\|^2 \lesssim p^{K-1}$ .

2562 For  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ , by the property (105), we have

$$2563 \quad (1 + o(1))\hat{G}_{ib}$$

$$2564 \quad = \left( \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\hat{S}_{a:}]^s\|_F^2 - \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 \right) \\ 2565 \quad - \left( \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\hat{S}_{b:}]^s\|_F^2 - \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 \right), \\ 2566 \quad (1 + o(1))\hat{H}_{ib} \\ 2567 \quad = \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 - \|[\mathbf{X}_{i:} \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\|_F^2 \\ 2568 \quad + \|A_{a:}^s - A_{b:}^s\|_F^2.$$

2569 We obtain the upper bounds following the proof for inequalities (88) and (89).  $\square$

2572 *Lemma 12 (Relationship Between Misclustering Loss and  
2573 Intermediate Parameters):* Under the Condition 1 and the setup  
2574 of Theorem 5 with fixed  $r \geq 2$ , as  $p \rightarrow \infty$ , we have

$$2575 \quad \|\mathbf{V} - \mathbf{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}} \frac{r}{\Delta_{\min}^2} L^{(t)}}, \quad (106)$$

$$2576 \quad \|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}} \frac{r}{\Delta_{\min}^2} L^{(t)}}, \quad (107)$$

$$2577 \quad \max_{b \in [r]} \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}]^s\| \\ 2578 \quad \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} \right), \quad (108)$$

$$2579 \quad \max_{b \in [r]} \|[\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\| \\ 2580 \quad \leq C \left( \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} + \frac{rL^{(t)}}{\Delta_{\min}} \right), \quad (109)$$

$$2581 \quad \max_{b \in [r]} \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\| \\ 2582 \quad \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} \right), \quad (110)$$

2583 for some positive universal constant  $C$ . In addition, the  
2584 inequality (109) also holds by replacing  $\mathbf{W}_{:b}^{(t)}$  to  $\mathbf{W}_{:b}$ . Further,  
2585 the above inequalities holds after replacing  $\mathbf{W}$  to  $\mathbf{P}$ ,  $\mathbf{V}$  to  $\mathbf{Q}$ ,  
2586 and  $L^{(t)}$  to  $L(\hat{z})$ .

2587 *Proof of Lemma 12:* We follow and use several intermediate  
2588 conclusions in [13, Proof of Lemma 5]. We prove each  
2589 inequality separately.

2590 1) Inequality (106). By [13, Proof of Lemma 5], we have

$$2591 \quad \|\mathbf{V} - \mathbf{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}} r \ell^{(t)}}.$$

Then, we complete the proof of inequality (106) by applying  
2592 Lemma 2 to the above inequality.

2) Inequality (107). By [13, Proof of Lemma 5], we have

$$2593 \quad \|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}} r \ell^{(t)}}. \quad 2594$$

Also, we complete the proof of inequality (106) by applying  
2595 Lemma 2 to the above inequality.

3) Inequality (108). We upper bound the desired quantity  
2596 by triangle inequality,

$$2597 \quad \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}]^s\| \leq I_1 + I_2 + I_3, \quad 2598$$

where

$$2601 \quad I_1 = \left\| \frac{\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right\|, \quad 2602$$

$$2603 \quad I_2 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right) \mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V} \right\|, \quad 2604$$

$$I_3 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V} \right\|. \quad 2604$$

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

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

$$I_1 \leq I_{11} + I_{12}, \quad 2607$$

where

$$I_{11} = \left\| \frac{\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right\|, \quad 2609$$

and

$$I_{12} = \left\| \frac{\mathbf{W}_{:b}^T \mathbf{E} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right\|. \quad 2611$$

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

$$D_{ab} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = a, z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}}, \quad \text{for all } a, b \in [r]. \quad 2614$$

By Lemma 10, we have  $\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} \gtrsim p/r$ . Then,  
2615 we have

$$\sum_{a \neq b, a, b \in [r]} D_{ab} \lesssim \frac{r}{p} \sum_{i: z^{(t)}(i) \neq z(i)} \theta(i) \lesssim \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \frac{1}{\log p}, \quad 2617$$

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

$$D_{bb} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \quad 2619$$

$$\geq \frac{c(\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} - p\ell^{(t)})}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \quad 2620$$

$$\gtrsim 1 - \frac{1}{\log p}, \quad 2621$$

under the inequality (66) in Condition 1. By the definition of  $\mathbf{W}, \mathbf{W}^{(t)}, \mathbf{V}$ , we have

$$\frac{\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} = [\mathbf{S}_{b:}]^s,$$

and

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

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 (2) and the bounds of  $\mathbf{D}$  in (111) and (112).

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

$$\begin{aligned} & \| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \| \\ &= \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \mathbf{S}_{a:} \| \\ &= \sum_{a \neq b, a \in [r]} D_{ab} \| \mathbf{S}_{a:} \sin(\mathbf{S}_{b:}, \mathbf{S}_{a:}) \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\|, \end{aligned} \quad (113)$$

where the first inequality follows from the triangle inequality, and the last inequality follows from Lemma 4. Note that with bounds (111) and (112), when  $p$  is large enough, we have

$$\begin{aligned} \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| &= \|D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}\| \\ &\geq D_{bb} \|\mathbf{S}_{b:}\| - \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \\ &\geq C_1, \end{aligned} \quad (114)$$

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

$$\begin{aligned} I_{11} &\leq \sqrt{2} \sin \alpha \\ &= \frac{\| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \|}{\| D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \|} \\ &\leq \frac{1}{C_1} \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \\ &\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \{ z^{(t)}(i) = b \} \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \\ &\leq \frac{r L^{(t)}}{\Delta_{\min}}, \end{aligned} \quad (115)$$

where the second inequality follows from (113) and (114), 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 (2).

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

$$\begin{aligned} I_{12} &\leq \frac{1}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| \\ &\quad + \frac{\| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \|}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\|. \end{aligned} \quad (116)$$

By [13, Proof of Lemma 5], we have

$$\begin{aligned} \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}. \quad (116) \\ \text{Notice that} \quad & \\ \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \| &\leq \|\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}\| \|\mathbf{X} \mathbf{V}\|_F \quad (116) \\ &\lesssim \frac{r^{3/2} L^{(t)}}{\sqrt{p} \Delta_{\min}^2} \|\mathbf{S}\| \|\Theta \mathbf{M}\|_\sigma \quad (116) \\ &\lesssim \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}}, \end{aligned} \quad (117)$$

where the second inequality follows from [13, Inequality (121), Proof of Lemma 5] and the last inequality follows from Lemma 6 and (66) 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 (114). Therefore, we have

$$\begin{aligned} I_{12} &\lesssim \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| \\ &\quad + \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \| \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\| \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} + \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}} \sqrt{\frac{r^{2K}}{p^K}} \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}, \end{aligned} \quad (118)$$

where second inequality follows from the inequalities (116), (117), and (62) in Condition 1.

Hence, combining inequalities (115) and (118) yields

$$I_1 \lesssim \frac{r L^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}. \quad (119)$$

For  $I_2$  and  $I_3$ , recall that  $\|\mathbf{W}_{:b}^T \mathbf{XV}\| = \|S_{b:}\| \geq c_3$  and  $\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\| \geq C_1$  by inequality (114). By triangle inequality and (62) in Condition 1, we have

$$I_2 \leq \frac{\|\mathbf{W}_{:b}^T \mathbf{EV}\|}{\|\mathbf{W}_{:b}^T \mathbf{XV}\|} \lesssim \|\mathbf{W}_{:b}^T \mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}, \quad (120)$$

and

$$I_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{EV}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}. \quad (121)$$

Therefore, combining the inequalities (119), (120), and (121), we finish the proof of inequality (108).

4) Inequality (109). Here we only show the proof of inequality (109) 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

$$J_1 = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{YV}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{YV}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{(t)}\|} \right\|,$$

$$J_2 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{YV}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{XV}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{YV} \right\|,$$

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

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

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

and

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

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 (115), 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

$$\mathcal{T}_k = \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T$$

$$\times_{k+1} \mathbf{W}^{(t),T} \times_{k+2} \cdots \times_K \mathbf{W}^{(t),T}$$

$$\mathcal{T}_{k+1} = \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T$$

$$\times_{k+1} \mathbf{W}^T \times_{k+2} \cdots \times_K \mathbf{W}^{(t),T}.$$

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

$$\langle \mathbf{W}_{:b}^{(t),T} \mathbf{XV}^k, \mathbf{W}_{:b}^{(t),T} \mathbf{XV}^{k+1} \rangle$$

$$= \langle \text{Mat}_{k+1}(\mathcal{T}_k), \text{Mat}_{k+1}(\mathcal{T}_{k+1}) \rangle$$

$$= \langle \mathbf{D}^T \mathbf{SZ}^k, \mathbf{SZ}^k \rangle$$

$$= \sum_{b \in [r]} \left( D_{bb} \|\mathbf{S}_{b:} \mathbf{Z}^k\|^2 + \sum_{a \neq b, a \in [r]} D_{ab} \langle \mathbf{S}_{a:} \mathbf{Z}^k, \mathbf{S}_{b:} \mathbf{Z}^k \rangle \right)$$

$$\gtrsim (1 - \log p^{-1}) \min_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2 - \log p^{-1} \max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2,$$

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 (111) and (112).

Note that

$$\|\mathbf{D}\|_\sigma \leq \|\mathbf{D}\|_F$$

$$\leq \sqrt{\sum_{b \in [r]} D_{bb}^2 + (\sum_{a \neq b, a \in [r]} D_{ab})^2}$$

$$\lesssim \sqrt{r + \log^2 p^{-1}} \lesssim 1,$$

where the second inequality follows from inequality (111), 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

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

following the Lemma 6 and Lemma 10. Then, for all  $k \in [K]$ , we have

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

$$\leq \|\mathbf{Z}^k\|_\sigma \leq \|\mathbf{D}_{:b}\| \|\mathbf{D}\|_\sigma^{K-k-1} \lesssim 1.$$

Thus, we have bounds

$$\max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\| \leq \max_{a \in [r]} \|\mathbf{S}_{a:}\| \|\mathbf{Z}^k\|_\sigma \lesssim 1,$$

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

Hence, when  $p$  is large enough, the inner product (122) 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

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

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

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

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

Note that we have  $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2} \leq \frac{\bar{C}}{C} r \log^{-1}(p)$  by Condition 1 and Lemma 2. Then, with the locally linear stability assumption, the  $\theta$  is  $\ell^{(t)}$ -locally linearly stable; i.e.,

$$\sin(\mathbf{p}(z^{(t)}), \mathbf{p}_{\theta}(z^{(t)})) \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

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 10, the constraint  $\min_{i \in [p]} \theta(i) \geq c > 0$ ,  $\|\mathbf{p}(z^{(t)})\| \lesssim p$  and  $\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1 \gtrsim p$ .

By the geometry property of trigonometric function, we have

$$\begin{aligned} \sin \beta_{k1} &= \min_{c \in \mathbb{R}} \frac{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{D} - c\tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1}\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}\|} \\ &\leq \frac{\|\mathbf{D}_{:b}^T \mathbf{S}\| \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_{\sigma} \|\mathbf{D}\|_{\sigma}^{K-k-1}}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k}(\mathbf{D})} \\ &\lesssim \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_F \\ &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{126}$$

where the second inequality follows from the singular property of  $\mathbf{D}$  in (123), (124) and the constraint of  $\mathbf{S}$  in (2).

To bound  $\sin \beta_{k2}$ , let  $\mathbf{C} = \text{diag}(\{\|\mathbf{S}_{a:\}\|}_{a \in [r]})$ . We have

$$\begin{aligned} \sin \beta_{k2} &\lesssim \frac{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{I}_r - \tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1}\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}\|} \\ &\lesssim \frac{\|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{Z}^k\|_F}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k-1}(\mathbf{D})} \\ &\lesssim \|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{C}^{-1}\|_F \|\mathbf{C} \mathbf{Z}^k\|_{\sigma} \\ &\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbf{1}\{z^{(t)}(i) = b\} \|\mathbf{S}_{b:\}^s - \mathbf{S}_{z(i):\}^s\| \\ &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{127}$$

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  $\theta$ , and the singular property of  $\mathbf{Z}^k$  in inequality (125), and the last line follows from the definition of  $L^{(t)}$ .

Combining (126) and (127) yields

$$\sin \beta_k \leq \sin \beta_{k1} + \sin \beta_{k2} \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

Finally, by triangle inequality, we obtain

$$J_{11} \leq \sum_{k=1}^{K-1} J_{11}^k \lesssim \sum_{k=1}^{K-1} \sin \beta_k \lesssim (K-1) \frac{r L^{(t)}}{\Delta_{\min}}. \tag{128}$$

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

$$\begin{aligned} J_{12} &\leq \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| \\ &\quad + \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}^{(t)}\|. \end{aligned} \tag{129}$$

Note that

$$\begin{aligned} \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\| &= \|\mathbf{D}^T \mathbf{S} \mathbf{Z}^1\| \\ &\geq \lambda_r(\mathbf{D}) \|\mathbf{S}\| \lambda_{r^{K-2}}(\mathbf{Z}^1) \gtrsim 1, \end{aligned} \tag{129}$$

where the inequality follows from the bounds (124) and (125).

By [13, Proof of Lemma 5], we have

$$\begin{aligned} &\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| \\ &\lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}} \frac{(K-1)\sqrt{L^{(t)}}}{\Delta_{\min}}. \end{aligned} \tag{130}$$

Notice that

$$\begin{aligned} &\|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F \\ &\leq \|(\mathbf{I} - \mathbf{D}^T) \mathbf{S}(\mathbf{I}_r^{\otimes(k-1)} \otimes \mathbf{D}^{\otimes(K-k-1)})\|_F \\ &\leq \|(\mathbf{W}^T - \mathbf{W}^{(t),T}) \Theta \mathbf{M}\|_F \|\mathbf{S}\|_F \|\mathbf{D}\|_{\sigma}^{K-k-1} \\ &\lesssim \|\mathbf{W}^T - \mathbf{W}^{(t),T}\| \|\Theta \mathbf{M}\|_{\sigma} \\ &\lesssim \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}}, \end{aligned} \tag{131}$$

where the first inequality follows from the tensor algebra in inequality (122), the second inequality follows from the fact that  $\mathbf{I} = \mathbf{W}^T \Theta \mathbf{M}$ , and the last inequality follows from [13, Proof of Lemma 5]. It follows from (131) and Lemma 10 that

$$\begin{aligned} \|\mathbf{W}_{:b}^{(t),T} \mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\| &\leq \|\mathbf{W}_{:b}^{(t),T}\| \sum_{k=1}^{K-1} \|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F \\ &\lesssim \frac{\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}}. \end{aligned} \quad (132)$$

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 (114) and (129), respectively. We have

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

where the second inequality follows from inequalities (130), (132), and the inequality (62) 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 (114) and (129), respectively. By triangle inequality and inequality (62) in Condition 1, we have

$$J_2 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}\mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}, \quad (133)$$

and

$$J_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}^{(t)}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}\mathbf{V}^{(t)}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}. \quad (134)$$

Therefore, combining the inequalities (128), (133), and (134), we finish the proof of inequality (109).

5) Inequality (110). By triangle inequality, we upper bound the desired quantity

$$\begin{aligned} &\|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ &\leq \|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}]^s\| \\ &\quad + \|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}]^s\| \\ &\quad + \|[\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ &\lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}, \end{aligned}$$

following the inequalities (108) and (109). Therefore, we finish the proof of inequality (110).

Next, we Show the Intermediate Inequalities Holds With  $\mathbf{P}, \mathbf{Q}$  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

$$\hat{D}_{ab} = \frac{\sum_{i \in [p]} \theta(i)\hat{\theta}(i)\mathbb{1}\{z(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2}$$

$$= \frac{\sum_{i \in [p]} (1 + o(p^{K-2}))(\hat{\theta}(i))^2 \mathbb{1}\{z(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2}, \quad (135)$$

where the second equation follows from Lemma 13, and thus  $\sum_{a \in [r]} \hat{D}_{ab} = 1 + o(1)$ . By the derivation of (111), (112), (124), and (123), we have

$$\sum_{a \neq b \in [r]} \hat{D}_{ab} \lesssim \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\}(\hat{\theta}(i))^2 \lesssim \frac{1}{\log p},$$

$$\hat{D}_{bb} \gtrsim 1 - \frac{1}{\log p}, \quad \lambda_{\min}(\hat{\mathbf{D}}) \asymp \|\hat{\mathbf{D}}\|_{\sigma} = (1 + o(1)).$$

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 12. The analogies of inequalities (106), (107), (108), (109), and (110) hold by using the MLE confusion matrix and the definition of  $L(\hat{z})$ .

Particularly, for the analogy of (109), the usage of MLE confusion matrix avoids the stability condition on  $\theta$ . Let  $\bar{\mathbf{D}}$  be the normalized version of  $\hat{\mathbf{D}}$ . The angle in inequality (126) decays to 0 at speed  $p^{-(K-2)} \lesssim \Delta_{\min}$  when  $K \geq 3$ , and the inequality (127) holds by the fact that

$$\|(\mathbf{I}_r - \bar{\mathbf{D}})\mathbf{S}\mathbf{C}^{-1}\|_F \lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{S}_{b:}^s - \mathbf{S}_{z(i):}^s\|$$

$$\lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{A}_{b:}^s - \mathbf{A}_{z(i):}^s\|.$$

□

**Lemma 13 (Polynomial Estimation Error of MLE):** Let  $(\hat{z}, \hat{\mathcal{S}}, \hat{\theta})$  denote the MLE in (9) with fixed  $K \geq 2$  and symmetric mean tensor, and  $\hat{\mathcal{X}}$  denote the mean tensor consisting of parameter  $(\hat{z}, \hat{\mathcal{S}}, \hat{\theta})$ . With high probability going to 1 as  $p \rightarrow \infty$ , we have

$$\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \lesssim \sigma^2 (r^K + Kpr),$$

with probability going to 1. When SNR  $\gtrsim p^{-(K-1)} \log p$ ,  $\theta$  is balanced, and  $\min_{i \in [p]} \theta(i) \geq c$  for some positive constant  $c$ , the MLE satisfies

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\}(\theta(i))^2 \lesssim \frac{1}{r \log p},$$

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\}(\hat{\theta}(i))^2 \lesssim \frac{1}{r \log p},$$

$$\text{and } L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p},$$

Further, we have

$$\theta(i)^2 = (1 + o(p^{-(K-2)}))\hat{\theta}(i)^2.$$

Proof of Lemma 13: Without loss of generality, we assume  $\sigma^2 = 1$  and identity mapping minimizes the misclustering error for MLE. For arbitrary two sets of parameters  $(z, \mathcal{S}, \boldsymbol{\theta}), (z', \mathcal{S}', \boldsymbol{\theta}') \in \mathcal{P}(\gamma)$  and corresponding mean tensors  $\mathcal{X}, \mathcal{X}'$ , we have

$$\begin{aligned} & \text{rank}(\text{Mat}_k(\mathcal{X}) - \text{Mat}_k(\mathcal{X}')) \\ & \leq \text{rank}(\text{Mat}_k(\mathcal{X})) + \text{rank}(\text{Mat}_k(\mathcal{X}')) \\ & \leq 2r, \quad k \in [K]. \end{aligned}$$

Hence, we have

$$\mathcal{X} - \mathcal{X}' \in \mathcal{Q}(2r, \dots, 2r), \quad (136)$$

where  $\mathcal{Q}(r, \dots, r) := \{\text{Tucker tensor with rank } (r, \dots, r)\}$ . Then, we obtain that

$$\begin{aligned} & \mathbb{P}(\|\mathcal{X} - \hat{\mathcal{X}}_{ML}\|_F \geq t) \\ & \leq 2\mathbb{P}\left(\sup_{\mathcal{X}, \mathcal{X}' \in \mathcal{Q}(r, \dots, r)} \left\langle \frac{\mathcal{X} - \mathcal{X}'}{\|\mathcal{X} - \mathcal{X}'\|_F}, \mathcal{E} \right\rangle \geq t\right) \\ & \leq 2\mathbb{P}\left(\sup_{\mathcal{T} \in \mathcal{Q}(2r, \dots, 2r) \cap \{\|\mathcal{T}\|_F=1\}} \langle \mathcal{T}, \mathcal{E} \rangle \geq t\right) \\ & \lesssim \exp(-Kpr), \end{aligned}$$

with the choice  $t \asymp \sigma\sqrt{(Kpr + r^K)}$ . Here the first inequality follows from [10, Lemma 1], the second inequality follows from (136), and the last inequality follows from [37, Lemma E5].

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

$$\begin{aligned} \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\theta(i))^2 & \lesssim \frac{r^{K-1}}{\Delta_{\min}^2 p^K} \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \\ & \lesssim \frac{r^{K-2}}{p^{K-1} \Delta_{\min}^2} \\ & \lesssim \frac{1}{r \log p}, \end{aligned}$$

and

$$L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p}.$$

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  $a \in [r]$

$$\begin{aligned} & (1 - O(1)) \frac{p^2}{r^2} \|\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}}\|_F^2 \\ & \leq \sum_{\hat{z}(i)=z(i)=a} (\theta(i) \mathbf{W}_{:a}^T \mathbf{X} - \hat{\theta}(i) \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}})^2 \\ & \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr, \end{aligned}$$

where the first inequality follows from the facts that  $\ell(\hat{z}, z) \lesssim \frac{1}{\log p}, |z^{-1}(a)| \asymp p/r$ ,

$$\begin{aligned} & |z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq |\hat{z}^{-1}(a)| \leq |z^{-1}(a)| + C \frac{p}{r} \ell(\hat{z}, z), \\ & |z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{z(i)=\hat{z}(i)=a} \theta(i) \leq |z^{-1}(a)|, \end{aligned}$$

and

$$|\hat{z}^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{\hat{z}(i)=z(i)=a} \hat{\theta}(i) \leq |\hat{z}^{-1}(a)|.$$

Hence, for all  $i \in [p]$

$$\begin{aligned} & (\theta(i) - \hat{\theta}(i))^2 \|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - O(p) \\ & \leq \|(\theta(i) - \hat{\theta}(i)) \mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - \|\hat{\theta}(i)(\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}})\|_F^2 \\ & \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr, \end{aligned}$$

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  $a \in [r]$

$$\|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 \geq \|\mathbf{S}_{:a}\|_F^2 \lambda_{\min}^{2(K-1)} (\Theta \mathbf{M}) \gtrsim p^{K-1}.$$

The inequality indicates that  $\theta(i)^2 = (1 + o(p^{-(K-2)}))\hat{\theta}(i)^2$ .  $\square$

## ACKNOWLEDGMENT

The authors would like to thank Zheng Tracy Ke, Anru Zhang, Rungang Han, and Yuetian Luo for helpful discussions and for sharing software packages.

## REFERENCES

- [1] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, “Tensor decompositions for learning latent variable models,” *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, Aug. 2014.
- [2] L. Wang, D. Durante, R. E. Jung, and D. B. Dunson, “Bayesian network-response regression,” *Bioinformatics*, vol. 33, no. 12, pp. 1859–1866, 2017.
- [3] P. Koniusz and A. Cherian, “Sparse coding for third-order supersymmetric tensor descriptors with application to texture recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5395–5403.
- [4] M. Wang, J. Fischer, and Y. S. Song, “Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition,” *Ann. Appl. Statist.*, vol. 13, no. 2, pp. 1103–1127, Jun. 2019.
- [5] V. Hore et al., “Tensor decomposition for multiple-tissue gene expression experiments,” *Nature Genet.*, vol. 48, no. 9, p. 1094, 2016.
- [6] D. Ghoshdastidar and A. Dukkipati, “Uniform hypergraph partitioning: Provable tensor methods and sampling techniques,” *J. Mach. Learn. Res.*, vol. 18, no. 50, pp. 1–41, 2017.
- [7] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral hypergraph partitioning under planted partition model,” *Ann. Statist.*, vol. 45, no. 1, pp. 289–315, 2017.
- [8] K. Ahn, K. Lee, and C. Suh, “Community recovery in hypergraphs,” *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6561–6579, Oct. 2019.
- [9] Z. T. Ke, F. Shi, and D. Xia, “Community detection for hypergraph networks via regularized tensor power iteration,” 2019, *arXiv:1909.06503*.
- [10] M. Wang and Y. Zeng, “Multiway clustering via tensor block models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 715–725.
- [11] E. Abbe, “Community detection and stochastic block models: Recent developments,” *J. Mach. Learn. Res.*, vol. 18, pp. 1–86, Jan. 2018.
- [12] E. C. Chi, B. J. Gaines, W. W. Sun, H. Zhou, and J. Yang, “Provable convex co-clustering of tensors,” *J. Mach. Learn. Res.*, vol. 21, no. 214, pp. 1–58, 2020.
- [13] R. Han, Y. Luo, M. Wang, and A. R. Zhang, “Exact clustering in tensor block model: Statistical optimality and computational limit,” *J. Roy. Stat. Soc., Ser. B, Stat. Methodol.*, vol. 84, no. 5, pp. 1666–1698, Nov. 2022.

- 2992 [14] P. J. Bickel and A. Chen, "A nonparametric view of network models and  
2993 Newman–Girvan and other modularities," *Proc. Nat. Acad. Sci. USA*,  
2994 vol. 106, no. 50, pp. 21068–21073, 2009.
- 2995 [15] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Community detection in degree-corrected block models," *Ann. Statist.*, vol. 46, no. 5,  
2996 pp. 2153–2185, Oct. 2018.
- 2998 [16] K. Ahn, K. Lee, and C. Suh, "Hypergraph spectral clustering in the  
2999 weighted stochastic block model," *IEEE J. Sel. Topics Signal Process.*,  
3000 vol. 12, no. 5, pp. 959–974, Oct. 2018.
- 3001 [17] M. Yuan, R. Liu, Y. Feng, and Z. Shang, "Testing community structure  
3002 for hypergraphs," *Ann. Statist.*, vol. 50, no. 1, pp. 147–169, Feb. 2022.
- 3003 [18] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-  
3004 rank matrix factorization: An overview," *IEEE Trans. Signal Process.*,  
3005 vol. 67, no. 20, pp. 5239–5269, Oct. 2019.
- 3006 [19] S.-Y. Yun and A. Proutiere, "Optimal cluster recovery in the labeled  
3007 stochastic block model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29,  
3008 2016, pp. 973–981.
- 3009 [20] C. Kim, A. S. Bandeira, and M. X. Goemans, "Stochastic block  
3010 model for hypergraphs: Statistical limits and a semidefinite programming  
3011 approach," 2018, *arXiv:1807.02884*.
- 3012 [21] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear  
3013 singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21,  
3014 no. 14, pp. 1253–1278, 2006.
- 3015 [22] Z. Zhang, G. I. Allen, H. Zhu, and D. Dunson, "Tensor network  
3016 factorizations: Relationships between brain structural connectomes and  
3017 traits," *NeuroImage*, vol. 197, pp. 330–343, Aug. 2019.
- 3018 [23] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational  
3019 limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338,  
3020 Nov. 2018.
- 3021 [24] M. Brennan and G. Bresler, "Reducibility and statistical-computational  
3022 gaps from secret leakage," in *Proc. 33rd Conf. Learn. Theory*, vol. 125,  
3023 2020, pp. 648–847.
- 3024 [25] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of  
Lloyd's algorithm and its variants," 2016, *arXiv:1612.02099*.
- 3025 [26] E. Abbe, J. Fan, K. Wang, and Y. Zhong, "Entrywise eigenvector analysis  
3026 of random matrices with low expected rank," *Ann. Statist.*, vol. 48, no. 3,  
3027 pp. 1452–1474, Jun. 2020.
- 3028 [27] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower  
3029 bounds and improved relaxations for tensor recovery," in *Proc. 31th Int.  
3030 Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. 73–81.
- 3031 [28] L. Floreescu and W. Perkins, "Spectral thresholds in the bipartite stochastic  
3032 block model," in *Proc. 29th Conf. Learn. Theory*, vol. 49, 2016,  
3033 pp. 943–959.
- 3034 [29] C. Gao and A. Y. Zhang, "Iterative algorithm for discrete structure  
3035 recovery," *Ann. Statist.*, vol. 50, no. 2, pp. 1066–1094, Apr. 2022.
- [30] I. E. Chien, C.-Y. Lin, and I.-H. Wang, "On the minimax misclassification  
3037 ratio of hypergraph community detection," *IEEE Trans. Inf. Theory*,  
3038 vol. 65, no. 12, pp. 8095–8118, Dec. 2019.
- [31] M. Meilă, "Local equivalences of distances between clusterings—A  
3039 geometric perspective," *Mach. Learn.*, vol. 86, no. 3, pp. 369–389,  
3040 Mar. 2012.
- [32] D. C. Van Essen et al., "The WU-minn human connectome project:  
An overview," *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.
- [33] R. S. Desikan et al., "An automated labeling system for subdividing  
3041 the human cerebral cortex on MRI scans into gyral based regions of  
3042 interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, Jul. 2006.
- [34] J. Hu, C. Lee, and M. Wang, "Generalized tensor decomposition with  
3043 features on multiple modes," *J. Comput. Graph. Statist.*, vol. 31, no. 1,  
3044 pp. 204–218, Jan. 2022.
- [35] S. H. Lee, J. M. Magallanes, and M. A. Porter, "Time-dependent com-  
3045 munity structure in legislation cosponsorship networks in the congress  
3046 of the republic of Peru," *J. Complex Netw.*, vol. 5, no. 1, pp. 127–144,  
3047 2017.
- [36] P. Rigollet and J.-C. Hütter, "High dimensional statistics," *Lect. Notes  
Course 18S997*, vol. 813, no. 814, p. 46, 2015.
- [37] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and  
3048 computational framework for generalized tensor estimation," *Ann. Statist.*,  
3049 vol. 50, no. 1, pp. 1–29, Feb. 2022.

**Jiaxin Hu** received the B.S. degree from Wuhan University in 2019 and the  
3060 master's degree in statistics from the University of Wisconsin–Madison in  
3061 2020, where she is currently pursuing the Ph.D. degree with the Department  
3062 of Statistics. Her research interests include tensor methods, network analysis,  
3063 and applications in genetics and social science.

**Miaoyan Wang** received the B.S. degree in mathematics from Fudan Uni-  
3065 versity in 2010 and the Ph.D. degree in statistics from the University of  
3066 Chicago in 2015. She is currently an Assistant Professor of statistics at  
3067 the University of Wisconsin–Madison. She is also a Faculty Affiliate with  
3068 the Mathematical Foundations of Machine Learning and the Institute for  
3069 Foundations of Data Science. Her research interests include the intersection  
3070 of statistics, machine learning, and genetics. She was a recipient of NSF  
3071 CAREER Award in 2022, Best Student Paper Awards (with her as an Advisor)  
3072 from American Statistical Association (three times in 2021, 2022, and 2023),  
3073 and New England Statistical Society in 2022.

# Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jixin Hu<sup>✉</sup> and Miaoyan Wang<sup>✉</sup>

**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 [1], neuroscience [2], and computer science [3]. Tensors effectively represent the multiway data and serve as the foundation in higher-order data analysis. One data example is from multitissue multiindividual gene expression study [4], [5], where the data tensor consists of expression measurements indexed by (gene, individual, tissue) triplets. Another example is *hypergraph* network [6], [7], [8], [9] 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.

Manuscript received 18 January 2022; revised 23 December 2022; accepted 6 January 2023. The work of Miaoyan Wang was supported in part by NSF under Grant CAREER DMS-2141865, Grant DMS-1915978, Grant DMS-2023239, and Grant EF-2133740; and in part by the Wisconsin Alumni Research Foundation. An earlier version of this paper was presented in part at the 25th International Conference on Artificial Intelligence and Statistics (AISTATS). (Corresponding author: Miaoyan Wang.)

The authors are with the Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: jhu267@wisc.edu; miaoyan.wang@wisc.edu).

Communicated by R. Venkataraman, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIT.2023.3239521>.

Digital Object Identifier 10.1109/TIT.2023.3239521

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. Fig. 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) [10], which extends the usual matrix stochastic block model [11] to tensors. The matrix analysis tools, however, are suboptimal for higher-order clustering. Developing tensor tools for solving block models has received increased interest recently [10], [12], [13].

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 nonrealistic. 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 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 [14] without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity [9]. The dTBM combines a higher-order checkerboard structure with degree parameter  $\theta = (\theta(1), \dots, \theta(p))^T$  to allow heterogeneity among  $p$  nodes. Fig. 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.

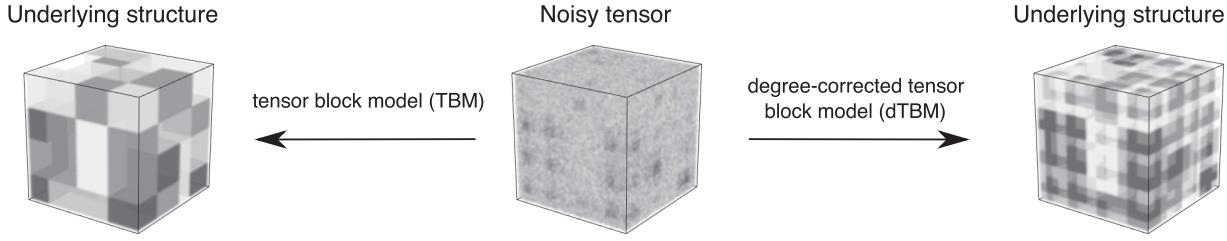


Fig. 1. Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

### 82 A. Our Contributions

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

- 86 • We develop a general dTBM and establish the identifiability  
87 for the uniqueness of clustering using the notion of angle separability.
- 88 • We present the phase transition of clustering performance  
89 with respect to three different statistical and computational behaviors. We characterize, for the first time,  
90 the critical signal-to-noise (SNR) thresholds in dTBMs,  
91 revealing the intrinsic distinctions among (vector) one-  
92 dimensional clustering, (matrix) biclustering, and (ten-  
93 sor) higher-order clustering. Specific SNR thresholds and  
94 algorithm behaviors are depicted in Fig. 2.
- 95 • We provide an angle-based algorithm that achieves exact  
96 clustering *in polynomial time* under mild conditions. Sim-  
97 ulation and data studies demonstrate that our algorithm  
98 outperforms existing higher-order clustering algorithms.

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

### 103 B. Related Work

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

- 107 • *Block model for clustering.* The block model such as  
108 stochastic block model (SBM) and degree-corrected SBM  
109 has been widely used for matrix clustering problems.  
110 The theoretical properties and algorithm performance for  
111 matrix block models have been well-studied [15]; see the  
112 review paper [11] and the references therein. However,  
113 The tensor counterparts are relatively less understood.
- 114 • *Tensor block model.* The (nondegree) tensor block model  
115 (TBM) is a higher-order extension of SBM, and its  
116 statistical-computational properties are investigated in  
117 recent literatures [7], [10], [13]. Some works [16] study  
118 the TBM with sparse observations, while, others [10],  
119 [13] and our work focus on the dense regime. Extending  
120 results from nondegree to degree-corrected model  
121 is highly challenging. Our dTBM parameter space is  
122 equipped with angle-based similarity and nuisance degree  
123 parameters. The extra complexity makes the Cartesian  
124 coordinates based analysis [13] nonapplicable to our  
125 setting. Towards this goal, we have developed a new polar

126 coordinates based analysis to control the model complex-  
127 ity. We have also developed a new angle-based iteration  
128 algorithm to achieve optimal clustering rates *without the*  
129 *need of estimating nuisance degree parameters.*

- 130 • *Degree-corrected block model.* The hypergraph  
131 degree-corrected block model (hDCBM) and its  
132 variant have been proposed in the literature [9],  
133 [17]. For this popular model, however, the optimal  
134 statistical-computational rates remain an open problem.  
135 Our main contribution is to provide a sharp statistical  
136 and computational critical phase transition in dTBM  
137 literature. In addition, our algorithm results in a faster  
138 exponential error rate, in contrast to the polynomial  
139 rate in [9]. The original hDCBM [9] is designed for  
140 binary observations only, and we extend the model to  
141 both continuous and binary observations. We believe  
142 our results are novel and helpful to the community. See  
143 Fig. 2 for overview of our results.

- 144 • *Global-to-local algorithm strategy.* Our methods gen-  
145 eralize the recent global-to-local strategy for matrix  
146 learning [15], [18], [19] to tensors [13], [16], [20].  
147 Despite the conceptual similarity, we address several  
148 fundamental challenges associated with this nonconvex,  
149 noncontinuous problem. We show the insufficiency of  
150 the conventional tensor HOSVD [21], and we develop  
151 a weighted higher-order initialization that relaxes the  
152 singular-value gap separation condition. Furthermore,  
153 our local iteration leverages the angle-based clustering  
154 in order to avoid explicit estimation of degree heteroge-  
155 neity. Our bounds reveal the interesting interplay between  
156 the computational and statistical errors. We show that  
157 our final estimate *provably* achieves the exact clustering  
158 within only polynomial-time complexity.

### 159 C. Notation

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

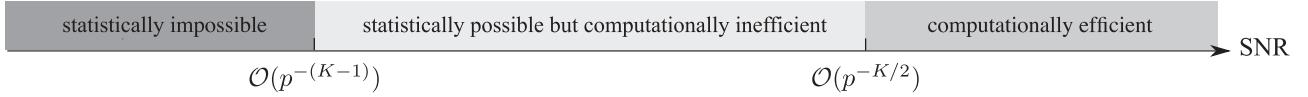


Fig. 2. SNR thresholds for statistical and computational limits in order- $K$  dTBM with dimension  $(p, \dots, p)$  and  $K \geq 2$ . The SNR gap between statistical possibility and computational efficiency exists only for tensors with  $K \geq 3$ .

TABLE I  
COMPARISON BETWEEN PREVIOUS METHODS WITH OUR METHOD

|                                       | Gao et al. (2018)[15] | Ahn et al. (2018)[16]   | Han et al. (2022)[13] | Ghoshdastidar et al. (2019)[7] | Ke et al. (2019)[9] | Ours             |
|---------------------------------------|-----------------------|-------------------------|-----------------------|--------------------------------|---------------------|------------------|
| Allow tensors of arbitrary order      | ×                     | ✓                       | ✓                     | ✓                              | ✓                   | ✓                |
| Allow degree heterogeneity            | ✓                     | ✗                       | ✗                     | ✓                              | ✓                   | ✓                |
| Singular-value gap-free clustering    | ✓                     | ✓                       | ✓                     | ✗                              | ✗                   | ✓                |
| Misclustering rate (for order $K^1$ ) | -                     | $p^{-(K-1)\alpha^{-1}}$ | $\exp(-p^{K/2})$      | $p^{-1}$                       | $p^{-2}$            | $\exp(-p^{K/2})$ |
| Consider sparse observation           | ✗                     | ✓                       | ✗                     | ✗                              | ✗                   | ✗                |

<sup>1</sup>We list the result for order- $K$  tensors with  $K \geq 3$  and general number of communities  $r = \mathcal{O}(1)$ .

<sup>2</sup>The parameter  $\alpha = f(p) > 0$  denotes the sparsity level which is some function of dimension  $p$ .

171 *a.* For two vector  $\mathbf{a}, \mathbf{b}$  of the same dimension, we denote the  
172 angle between  $\mathbf{a}, \mathbf{b}$  by

$$173 \cos(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|},$$

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

177 Let  $\mathcal{Y} \in \mathbb{R}^{p_1 \times \dots \times p_K}$  be an order- $K$   $(p_1, \dots, p_K)$ -  
178 dimensional tensor. We use  $\mathcal{Y}(i_1, \dots, i_K)$  to denote the  
179  $(i_1, \dots, i_K)$ -th entry of  $\mathcal{Y}$ . The multilinear multiplication of a  
180 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  
181 an order- $K$   $(p_1, \dots, p_K)$ -dimensional tensor  $\mathcal{X}$ , denoted

$$182 \mathcal{X} = \mathcal{S} \times_1 \mathbf{M}_1 \times \dots \times_K \mathbf{M}_K,$$

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

$$184 \mathcal{X}(i_1, \dots, i_K) \\ 185 = \sum_{(j_1, \dots, j_K)} \mathcal{S}(j_1, \dots, j_K) \mathbf{M}_1(i_1, j_1) \dots \mathbf{M}_K(i_K, j_K).$$

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

#### 200 D. Organization

201 The rest of this paper is organized as follows. Section II  
202 introduces the degree-corrected tensor block model (dTBM)  
203 with three motivating examples and presents the identifiability  
204 of dTBM under the angle gap condition. We show the phase  
205 transition and the existence of statistical-computational gaps  
206 for the higher-order dTBM in Section III. In Section IV,  
207 we provide a polynomial-time two-stage algorithm with

208 misclustering rate guarantees. Extension to Bernoulli models  
209 is also presented. In Section V, we compare our work with  
210 nondegree tensor block models. Numerical studies including  
211 the simulation, comparison with other methods, and two real  
212 dataset analyses are in Sections VI-VII. The main technical  
213 ideas we develop for addressing main theorems are provided  
214 in Section VIII. Detailed proofs and extra theoretical results  
215 are provided in Appendix.

## II. MODEL FORMULATION AND MOTIVATIONS

### A. Degree-Corrected Tensor Block Model

216 Suppose that we have an order- $K$  data tensor  $\mathcal{Y} \in \mathbb{R}^{p \times \dots \times p}$ .  
217 Assume that there exist  $r \geq 1$  disjoint communities among the  
218  $p$  nodes. We represent the community assignment by a function  
219  $z: [p] \mapsto [r]$ , where  $z(i) = a$  for  $i$ -th node that belongs to  
220 the  $a$ -th community. Then,  $z^{-1}(a) = \{i \in [p]: z(i) = a\}$   
221 denotes the set of nodes that belong to the  $a$ -th community,  
222 and  $|z^{-1}(a)|$  denotes the number of nodes in the  $a$ -th community.  
223 Let  $\boldsymbol{\theta} = (\theta(1), \dots, \theta(p))^T$  denote the degree heterogeneity for  
224  $p$  nodes. We consider the order- $K$  dTBM [7], [9],  
225

$$226 \mathcal{Y}(i_1, \dots, i_K) = \mathcal{S}(z(i_1), \dots, z(i_K)) \prod_{k=1}^K \theta_{i_k} + \mathcal{E}(i_1, \dots, i_K),$$

227 where  $\mathcal{S} \in \mathbb{R}^{r \times \dots \times r}$  is an order- $K$  tensor collecting the block  
228 means among communities, and  $\mathcal{E} \in \mathbb{R}^{p \times \dots \times p}$  is a noise tensor  
229 consisting of independent zero-mean sub-Gaussian entries  
230 with variance bounded by  $\sigma^2$ . The unknown parameters are  $z$ ,  
231  $S$ , and  $\boldsymbol{\theta}$ . The dTBM can be equivalently written in a compact  
232 form of tensor-matrix product:  
233

$$234 \mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \dots \times_K \boldsymbol{\Theta} \mathbf{M}, \quad (1)$$

235 where  $\boldsymbol{\Theta} = \text{diag}(\theta(1), \dots, \theta(p)) \in \mathbb{R}^{p \times p}$  is a diagonal matrix,  
236  $\mathbf{M} \in \{0, 1\}^{p \times r}$  is the membership matrix associated with  
237 community assignment  $z$  such that  $\mathbf{M}(i, j) = 1\{z(i) = j\}$ .  
238 By definition, each row of  $\mathbf{M}$  has one copy of 1’s and  
239 0’s elsewhere. Note that the discrete nature of  $\mathbf{M}$  renders  
240 our model (1) more challenging than Tucker decomposition.  
241 We call a tensor  $\mathcal{Y}$  an  $r$ -block tensor with degree  $\boldsymbol{\theta}$  if  $\mathcal{Y}$  admits  
242 dTBM (1) and let  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  denote the mean tensor. The goal  
243 of clustering is to estimate  $z$  from a single noisy tensor  $\mathcal{Y}$ .  
244 We are particularly interested in the high-dimensional regime  
245 where  $p$  grows whereas  $r = \mathcal{O}(1)$ .  
246

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  $(\Theta, M, z)$ ; the noise tensor  $\mathcal{E}$  and the data tensor  $\mathcal{Y}$  are still possibly asymmetric. In general, we allow asymmetric mean tensors with  $\{(\Theta_k, M_k, z_k)\}_{k=1}^K$ , one for each mode. The extension can be found in Appendix B.

### B. Motivating Examples

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

1) *Tensor Block Model*: Consider the model (1). Let  $\theta(i) = 1$  for all  $i \in [p]$ . The model (1) reduces to the tensor block model, which is widely used in previous clustering algorithms [10], [12], [13]. The theoretical results in TBM serve as benchmarks for dTBM.

2) *Community Detection in Hypergraphs*: The hypergraph network is a powerful tool to represent the complex entity relations with higher-order interactions [9]. 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 that there exist  $r$  disjoint communities among  $p$  nodes, and the connection probabilities depend on the community assignments and node-specific parameters. Then, the equation (1) models  $\mathbb{E}\mathcal{Y}$  with unknown degree heterogeneity  $\theta$  and sub-Gaussianity parameter  $\sigma^2 = 1/4$ .

3) *Multilayer Weighted Network*: Multilayer weighted network data consists of multiple networks over the same set of nodes. One representative example is the brain connectome data [22]. The multilayer 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 that there exist  $r$  disjoint communities among  $p$  nodes and  $r_l$  disjoint communities among the  $L$  layers. The multilayer network community detection is modeled by the general asymmetric dTBM model (1)

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

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

4) *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 multitissue gene expression dataset, which records different gene expressions 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 that 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 (1)

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \Theta_1 M_1 \times_2 \Theta_2 M_2 \times_3 \Theta_3 M_3,$$

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

*Remark 1 (Comparison With Nondegree Models)*: Our dTBM uses fewer block parameters than TBM. In particular, every nondegree  $r_1$ -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 (1). For ease of notation, we focus on symmetric tensors; the extension to asymmetric tensors are similar. We use  $\mathcal{P}$  to denote the following parameter space for  $(z, \mathcal{S}, \theta)$ ,

$$\begin{aligned} \mathcal{P} = & \left\{ (z, \mathcal{S}, \theta) : \theta \in \mathbb{R}_+^p, \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, \right. \\ & \left. c_3 \leq \|\text{Mat}(\mathcal{S})_{:a}\| \leq c_4, \|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r] \right\} \end{aligned} \quad (2)$$

where  $c_i > 0$ 's are universal constants. We briefly describe the rationale of the constraints in (2). First, the entrywise positivity constraint on  $\theta \in \mathbb{R}_+^p$  is imposed to avoid sign ambiguity between entries in  $\theta_{z^{-1}(a)}$  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 (1); see Example 1 below. Second, recall that the quantity  $|z^{-1}(a)|$  denotes the number of nodes in the  $a$ -th community. The constants  $c_1, c_2$  in the  $|z^{-1}(a)|$  bounds assume the roughly balanced size across  $r$  communities. Third, the constant  $c_3$  requires that all slides in  $\mathcal{S}$  have nondegenerate norm. Particularly, the lower bound  $c_3$  excludes the purely zero slide to avoid trivial nonidentifiability of model (1); see Example 2 below. The upper bound  $c_4$  is a technical constraint to avoid the slides with diverging norm as dimension grows. Lastly, the  $\ell_1$  normalization  $\|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$  is imposed to avoid the scalar ambiguity between  $\theta_{z^{-1}(a)}$  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.

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

$$357 \quad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta} \mathbf{M} \times_2 \boldsymbol{\Theta} \mathbf{M} \times_3 \boldsymbol{\Theta} \mathbf{M},$$

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

$$361 \quad \mathcal{X}_{::1} = \mathcal{X}_{::2} = -\mathcal{X}_{::3} = -\mathcal{X}_{::4} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

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

$$365 \quad \mathcal{X} = \mathcal{S}' \times_1 \boldsymbol{\Theta}' \mathbf{M}' \times_2 \boldsymbol{\Theta}' \mathbf{M}' \times_3 \boldsymbol{\Theta}' \mathbf{M}',$$

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

$$370 \quad \mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{M}' = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

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

377 *Example 2 (Nonidentifiability With Purely Zero Core Slice):*  
 378 Consider an order-2 dTBM with core tensor  $\mathcal{S} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$   
 379 degree matrices  $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \text{diag}(1, 1, 1, 1)$ , and mean tensor

$$380 \quad \mathcal{X} = \boldsymbol{\Theta}_1 \mathbf{M} \mathcal{S} \mathbf{M}^T \boldsymbol{\Theta}_2, \quad \text{with } \mathbf{M} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

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

383 We now provide the identifiability conditions for our model  
 384 before estimation procedures. When  $r = 1$ , the decomposi-  
 385 tion (1) is always unique (up to cluster label permutation) in  $\mathcal{P}$ ,  
 386 because dTBM is equivalent to the rank-1 tensor family under  
 387 this case. When  $r \geq 2$ , the Tucker rank of signal tensor  $\mathbb{E}\mathcal{Y}$   
 388 in (1) is bounded by, but not necessarily equal to, the number  
 389 of blocks  $r$  [10]. Therefore, one can not apply the classical  
 390 identifiability conditions for low-rank tensors to dTBM. Here,  
 391 we introduce a key separation condition on the core tensor.

392 *Assumption 1 (Angle Gap):* Let  $\mathbf{S} = \text{Mat}(\mathcal{S})$ . Assume that  
 393 the minimal gap between normalized rows of  $\mathbf{S}$  is bounded

away from zero; i.e.,

$$394 \quad \Delta_{\min} := \min_{a \neq b \in [r]} \left\| \frac{\mathbf{S}_{a:}}{\|\mathbf{S}_{a:}\|} - \frac{\mathbf{S}_{b:}}{\|\mathbf{S}_{b:}\|} \right\| > 0, \quad \text{for } r \geq 2. \quad (3)$$

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

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

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

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

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

#### A. Assumptions

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

$$420 \quad \text{SNR} := \Delta_{\min}^2 / \sigma^2 = p^\gamma, \quad (4)$$

422 with varying  $\gamma \in \mathbb{R}$  that quantifies different regimes of  
 423 interest. We call  $\gamma$  the *signal exponent*. Intuitively, a larger  
 424 SNR, or equivalently a larger  $\gamma$ , benefits the clustering in the  
 425 presence of noise. With quantification (4), we consider the  
 426 following parameter space,

$$427 \quad \mathcal{P}(\gamma) = \mathcal{P} \cap \{\mathcal{S} \text{ satisfies SNR condition (4) with } \gamma\}. \quad (5)$$

428 The 1-block dTBM does not belong to the space  $\mathcal{P}(\gamma)$  when  
 429  $\gamma < 0$ , due to the convention in Assumption 1. Our goal is to  
 430 characterize the clustering accuracy with respect to  $\gamma$  under  
 431 the space  $\mathcal{P}(\gamma)$ .

432 In our algorithmic development, we often refer to the  
 433 regime of balanced degree heterogeneity. We call the degree  
 434  $\boldsymbol{\theta}$  *balanced* if

$$435 \quad \min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|. \quad (6)$$

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

440 *Lemma 1 (Angle Gaps in  $\mathcal{X}$  and  $\mathcal{S}$ ):* Consider the dTBM  
 441 model (1) under the parameter space  $\mathcal{P}$  in (2) with  $r \geq 2$ .  
 442 Suppose  $\boldsymbol{\theta}$  is balanced satisfying (6) and  $\min_{i \in [p]} \theta(i) \geq c$

TABLE II  
PARAMETER SPACE COMPARISON BETWEEN PREVIOUS WORK WITH OUR ASSUMPTION

| Assumptions in parameter space | Gao et al. (2018)[15]         | Han et al. (2022)[13] | Ke et al. (2019)[9] | 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 |

from some constant  $c > 0$ . Then, as  $p \rightarrow \infty$ , 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 (6). 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

$$\mathbf{S} = \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix}, \quad (7)$$

and  $\boldsymbol{\theta}$  such that  $\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2 = p^m \|\boldsymbol{\theta}_{z^{-1}(2)}\|^2$ , where  $m \in [-1, 1]$  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

$$\Delta_{\mathbf{X}}^2 := \min_{i,j \in [p], z(i) \neq z(j)} \left\| \frac{\mathbf{X}_{i:}}{\|\mathbf{X}_{i:}\|} - \frac{\mathbf{X}_{j:}}{\|\mathbf{X}_{j:}\|} \right\|, \quad (8)$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . Take  $a = p^{-1/4}$  in the model setup (7). We have

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

Based on the Theorem 2 in Section III, the dTBM is impossible to solve when  $\Delta_{\mathbf{X}}^2 \lesssim p^{-1}$  even though  $\Delta_{\min}^2 \asymp p^{-1/2}$ ; 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 (6) 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 [15] that restricts the highest degree in the sublinear 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 [9], 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 degree balance across communities. In [15], 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 (2). Define the misclustering error by

$$\ell(\hat{z}, z) = \frac{1}{p} \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq \pi \circ z(i)\}, \quad (492)$$

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 infimum 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 (2) 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  $(\hat{z}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}, \hat{\boldsymbol{\theta}}_{\text{MLE}})$ , over the estimation space  $\mathcal{P}$ , where

$$(\hat{z}_{\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}, \hat{\boldsymbol{\theta}}_{\text{MLE}}) = \arg \min_{(z, \mathcal{S}, \boldsymbol{\theta}) \in \mathcal{P}} \|\mathcal{Y} - \mathcal{X}(z, \mathcal{S}, \boldsymbol{\theta})\|_F^2. \quad (9)$$

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

**• Impossibility.** Assume  $p \rightarrow \infty$  and  $2 \leq r \lesssim p^{1/3}$ . Let  $\mathcal{P}_{\mathcal{S}}(\gamma) := \{\mathcal{S} : c_3 \leq \|\text{Mat}(\mathcal{S})_{:a}\| \leq c_4, a \in [r]\} \cap \{\mathcal{S} : \Delta_{\min}^2 = p^\gamma\}$  denote the space for valid  $\mathcal{S}$  satisfying SNR condition (4), and  $\mathcal{P}_{z, \boldsymbol{\theta}} := \{\boldsymbol{\theta} \in \mathbb{R}_+^p, \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, \|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|, a \in [r]\}$  denote the space for valid  $(z, \boldsymbol{\theta})$ , where  $c_1, c_2, c_3, c_4$  are the constants in parameter space (2). If the signal exponent satisfies  $\gamma < -(K-1)$ , then, for any true core tensor  $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}(\gamma)$ , no estimator  $\hat{z}_{\text{stat}}$  achieves exact recovery in expectation;

526 that is, when  $\gamma < -(K - 1)$ , we have

$$\liminf_{p \rightarrow \infty} \inf_{\mathcal{S} \in \mathcal{P}_S(\gamma)} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \theta) \in \mathcal{P}_{z, \theta}} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1. \quad (10)$$

528 Further, we define the parameter space  $\mathcal{P}'(\gamma') := \mathcal{P} \cap$   
529  $\{\Delta_X^2 = p^{\gamma'}\}$ , where  $\Delta_X^2$  is the mean tensor minimal gap  
530 in (8). When  $\gamma' < -(K - 1)$ , we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \mathcal{S}, \theta) \in \mathcal{P}'(\gamma')} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1.$$

- 532 • **MLE achievability.** Suppose that the signal exponent  
533 satisfies  $\gamma > -(K - 1) + c_0$  for an arbitrary constant  
534  $c_0 > 0$ . Furthermore, assume that  $\theta$  is balanced and  
535  $\min_{i \in [p]} \theta(i) \geq c$  from some constant  $c > 0$ . Then, when  
536  $p \rightarrow \infty$ , for fixed  $r \geq 1$ , the MLE in (9) achieves exact  
537 recovery in high probability; that is,

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

539 with probability going to 1.

540 The proofs for the two parts in Theorem 2 are in the  
541 Appendices B-D and B-G, respectively. The first part of The-  
542 orem 2 demonstrates impossibility of exact recovery whenever  
543 the core tensor  $\mathcal{S}$  satisfies SNR condition (4) with exponent  
544  $\gamma < -(K - 1)$ . The proof is information-theoretical, and  
545 therefore the results apply to all statistical estimators, including  
546 but not limited to MLE and trace maximization [6]. The  
547 minimax bound (10) indicates the worst case impossibility for  
548 a particular core tensor  $\mathcal{S}$  with signal exponent  $\gamma < -(K - 1)$ ;  
549 i.e., under the assumptions of Theorem 2, when  $\gamma < -(K - 1)$ ,  
550 we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \sup_{(z, \mathcal{S}, \theta) \in \mathcal{P}(\gamma)} \mathbb{E} [p\ell(\hat{z}_{\text{stat}}, z)] \geq 1.$$

552 Such worst case impossibility is studied in related works [13],  
553 [15] while our lower bound (10) provides a stronger impossi-  
554 bility statement for arbitrary core tensors with weak signals.  
555 The second part of Theorem 2 shows the exact recovery of  
556 MLE when  $\gamma > -(K - 1) + c_0$  for an arbitrary constant  
557  $c_0 > 0$ . Combining the impossibility and achievability results,  
558 we conclude that the boundary  $\gamma_{\text{stat}} := -(K - 1)$  is the critical  
559 value for statistical performance of dTBM with respect to our  
560 SNR.

### 561 C. Computational Critical Value

562 The computational critical value means the minimal SNR  
563 required for exact recovery with *polynomial-time* computa-  
564 tional cost. An important ingredient to establish the computa-  
565 tional limits is the *hypergraphic planted clique (HPC) conjec-*  
566 *ture* [23], [24]. The HPC conjecture indicates the impossibility  
567 of fully recovering the planted cliques with polynomial-time  
568 algorithm when the clique size is less than the number of ver-  
569 tices in the hypergraph. The formal statement of HPC detection  
570 conjecture is provided in Definition 1 and Conjecture 1 as  
571 follows.

572 *Definition 1 (Hypergraphic Planted Clique (HPC) Detec-*  
573 *tion):* Consider an order- $K$  hypergraph  $H = (V, E)$  where

574  $V = [p]$  collects vertices and  $E$  collects all the order- $K$   
575 edges. Let  $\mathcal{H}_k(p, 1/2)$  denote the Erdős-Rényi  $K$ -hypergraph  
576 where the edge  $(i_1, \dots, i_K)$  belongs to  $E$  with probability  
577  $1/2$ . Further, we let  $\mathcal{H}_K(p, 1/2, \kappa)$  denote the hyphpergraph  
578 with planted cliques of size  $\kappa$ . Specifically, we generate a  
579 hypergraph from  $\mathcal{H}_k(p, 1/2)$ , pick  $\kappa$  vertices uniformly from  
580  $[p]$ , denoted  $K$ , and then connect all the hyperedges with  
581 vertices in  $K$ . Note that the clique size  $\kappa$  can be a function of  
582  $p$ , denoted  $\kappa_p$ . The order- $K$  HPC detection aims to identify  
583 whether there exists a planted clique hidden in an Erdős-  
584 Rényi  $K$ -hypergraph. The HPC detection is formulated as the  
585 following hypothesis testing problem

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

587 *Conjecture 1 (HPC Conjecture):* Consider the HPC detec-  
588 tion problem in Definition 1 with  $K \geq 2$ . Suppose the  
589 sequence  $\{\kappa_p\}$  such that  $\limsup_{p \rightarrow \infty} \log \kappa_p / \log \sqrt{p} \leq (1 - \tau)$   
590 for any  $\tau > 0$ . Then, for every sequence of polynomial-time  
591 test  $\{\varphi_p\} : H \mapsto \{0, 1\}$  we have

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

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

596 *Theorem 3 (Computational Critical Value):* Consider gen-  
597 eral Gaussian dTBMs under the parameter space  $\mathcal{P}$  with  
598  $K \geq 2$ . Then, we have the following computational phase  
599 transition.

- 600 • **Impossibility.** Assume HPC conjecture holds and  $r \geq$   
601 2. If the signal exponent satisfies  $\gamma < -K/2$ , then,  
602 no *polynomial-time estimator*  $\hat{z}_{\text{comp}}$  achieves exact recov-  
603 ery in expectation as  $p \rightarrow \infty$ ; that is, when  $\gamma < -K/2$ ,  
604 we have

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

- 605 • **Polynomial-time algorithm achievability.** Suppose that  
606 we have fixed  $r \geq 1$ , and the signal exponent satisfies  
607  $\gamma > -K/2 + c_0$  for an arbitrary constant  $c_0 > 0$ .  
608 Furthermore, assume that the degree  $\theta$  is balanced, lower  
609 bounded in that  $\min_{i \in [p]} \theta_i \geq c$  for some constant  $c > 0$ ,  
610 and satisfies the locally linear stability in Definition 2 in  
611 the neighborhood  $\mathcal{N}(z, \varepsilon)$  for all  $\varepsilon \leq E_0$  and some  $E_0 \gtrsim$   
612  $\log^{-1} p$ . Then, as  $p \rightarrow \infty$ , there exists a polynomial-time  
613 algorithm  $\hat{z}_{\text{poly}}$  that achieves exact recovery in high prob-  
614 ability; that is,

$$\ell(\hat{z}_{\text{poly}}, z) \lesssim \text{SNR}^{-1} \exp\left(-\frac{p^{K-1} \text{SNR}}{r^{K-1}}\right) \rightarrow 0,$$

617 with probability going to 1.

618 The proofs for the two parts in Theorem 3 are in the  
619 Appendices B-E and B-G, respectively. The first part of  
620 Theorem 3 indicates the impossibility of exact recovery by  
621 polynomial-time algorithms when  $\gamma < -K/2$ , and the sec-  
622 ond part shows the existence of such algorithm when  $\gamma >$   
623  $-K/2 + c_0$  for an arbitrary constant  $c_0 > 0$  under extra  
624 technical assumptions. In Section IV, we will present an

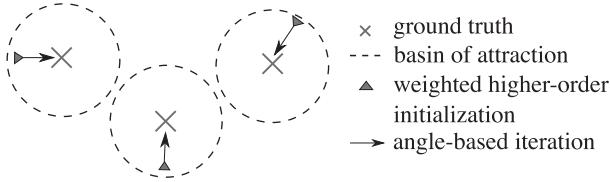


Fig. 3. Illustration of our global-to-local algorithm.

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. Fig. 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

$$Y = \Theta M S + E,$$

where  $Y \in \mathbb{R}^{p \times d}$  collects  $n$  data points in  $\mathbb{R}^d$ ,  $S \in \mathbb{R}^{r \times d}$  collects the  $d$ -dimensional centroids for  $r$  clusters, and  $\Theta \in \mathbb{R}^{p \times p}$ ,  $M \in \{0, 1\}^{p \times r}$ ,  $E \in \mathbb{R}^{p \times d}$  have the same meaning as in dTBM. [25] implies that polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Therefore, we conclude that the statistical-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 Fig. 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 dTBMs with symmetric mean tensor and independent sub-Gaussian noises such as Gaussian and uniform observations. The extensions for Bernoulli observations and other practical issues are in Sections IV-C and IV-D.

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

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

where the superscript  $\cdot^s$  denotes the normalized vector; i.e.,  $\mathbf{a}^s := \mathbf{a}/\|\mathbf{a}\|$  if  $\mathbf{a} \neq 0$  and  $\mathbf{a}^s = 0$  if  $\mathbf{a} = 0$  for any vector  $\mathbf{a}$ . 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 (Relationship Between Misclustering Error and Loss):* Consider the dTBM under the parameter space  $\mathcal{P}$ . Suppose  $\min_{i \in [p]} \theta(i) > c$  for some constant  $c > 0$ . 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 (1), for all  $i \in [p]$ , we have

$$\theta(i)^{-1} \mathbf{X}_{i:} = [\text{Mat}(\mathcal{S} \times_2 \Theta \mathbf{M} \times_3 \Theta \mathbf{M})]_{z(i)} : . \quad (688)$$

This implies that, all node  $i$  belonging to the  $a$ -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 the denoising step and the 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 [21] 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.,

$$\hat{\mathbf{U}}_1 = \text{SVD}_r(\text{Mat}_1(\mathcal{Y} \times_2 \mathbf{U}_{\text{pre},2} \mathbf{U}_{\text{pre},2}^T \times_3 \mathbf{U}_{\text{pre},3} \mathbf{U}_{\text{pre},3}^T)) . \quad (702)$$

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

$$\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\mathbf{U}}_1 \hat{\mathbf{U}}_1^T \times_2 \hat{\mathbf{U}}_2 \hat{\mathbf{U}}_2^T \times_3 \hat{\mathbf{U}}_3 \hat{\mathbf{U}}_3^T . \quad (705)$$

The double projection improves usual matrix spectral methods in order to alleviate the noise effects for  $K \geq 3$  [13].

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 [9], we apply the clustering on the unfolded tensor  $\hat{\mathbf{X}}$  rather than on the factors  $\hat{\mathbf{U}}_k$ . This strategy relaxes the singular-value gap condition [13], [15]. We assign

718 degenerate rows with purely zero entries to an arbitrarily  
 719 random cluster; these nodes are negligible in high-dimensions  
 720 because of the lower bound on  $\|\text{Mat}(\mathcal{S})_{a:}\|$  in (2). The final  
 721 result gives the initial cluster assignment  $z^{(0)}$ . Full procedures  
 722 for clustering are provided in Sub-algorithm 1.

723 We now establish the misclustering error rate of initialization.  
 724

725 *Theorem 4 (Error for Weighted Higher-Order Initialization):*  
 726 Consider the general sub-Gaussian dTBM with fixed  $r \geq 1$ ,  
 727  $K \geq 2$ , i.i.d. noise under the parameter space  $\mathcal{P}$ , and  
 728 Assumption 1. Assume  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  
 729  $c > 0$ . Let  $\Delta_{\mathbf{X}}$  denote the minimal gap in mean tensor defined  
 730 in (8), and let  $z_k^{(0)}$  denote the output of Sub-algorithm 1.  
 731 With probability going to 1, as  $p \rightarrow \infty$ , we have

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

732 Further, assume that  $\boldsymbol{\theta}$  is balanced as (6). We have

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

733 with probability going to 1 as  $p \rightarrow \infty$ .

734 *Remark 5 (Comparison to Previous Results):* For fixed  
 735 SNR, our initialization error rate with  $K = 2$  agrees with  
 736 the initialization error rate  $\mathcal{O}(p^{-1})$  in matrix models [15].  
 737 Furthermore, in the special case of nondegree TBMs with  
 738  $\boldsymbol{\theta} = \mathbf{1}_p$ , we achieve the same initial misclustering error  
 739  $\mathcal{O}(p^{-K/2})$  as in nondegree models [13]. Theorem 4 implies  
 740 the advantage of our algorithm in achieving both accuracy  
 741 and model flexibility.

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

747 *Remark 7 (Singular-Value Gap-Free Clustering):* Note  
 748 that our clustering directly applies to the estimated mean  
 749 tensor  $\hat{\mathbf{X}}$  rather than the leading tensor factors  $\hat{\mathbf{U}}_k$ .  
 750 Applying clustering to the tensor factors suffers from the  
 751 nonidentifiability issue due to the infinitely many orthogonal  
 752 rotations when the number of blocks  $r \geq 3$  in the absence  
 753 of singular-value gaps. Such ambiguity causes the trouble  
 754 for effective clustering [26]. In contrast, our initialization  
 755 algorithm applies the clustering to the overall mean tensor  $\hat{\mathbf{X}}$ .  
 756 This strategy avoids the nonidentifiability issue regardless of  
 757 the number of blocks and singular-value gaps.

## 760 B. Angle-Based Iteration

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

770 algorithm estimates the assignment  $z$  by minimizing the angle  
 771 between data vectors and centroids; i.e.,  
 772

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

773 The classical Euclidean-distance based clustering [13] fails  
 774 to recover  $z$  in the presence of degree heterogeneity, even  
 775 under noiseless case. In contrast, the proposed angle-based  
 776  $k$ -means algorithm achieves accurate recovery without the  
 777 explicit estimation of  $\boldsymbol{\theta}$ .

778 Our Sub-algorithm 2 shares the same spirit as in the angle-  
 779 based  $k$ -means. We still take the order-3 tensor for illustration.  
 780 Specifically, Sub-algorithm 2 updates estimated core tensor  
 781 and cluster assignment in each iteration. We use superscript  
 782  $(t)$  to denote the estimate from the  $t$ -th iteration, where  $t =$   
 783  $1, 2, \dots$ . For core tensor, we consider the following update  
 784 strategy

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

785 Intuitively,  $\mathcal{S}^{(t)}$  becomes closer to the true core  $\mathcal{S}$  as  $z_k^{(t)}$  is  
 786 more precise. For cluster assignment, we first aggregate the  
 787 slices of  $\mathcal{Y}$  and obtain the reduced tensor  $\mathcal{Y}_1^d \in \mathbb{R}^{p \times r \times r}$  on  
 788 the first mode with given  $z_k^{(t)}$ , where

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

789 Similarly, we also obtain  $\mathcal{Y}_2^d, \mathcal{Y}_3^d$ . We use  $\mathbf{Y}_k^d$  and  $\mathbf{S}_k^{(t)}$  to  
 790 denote the  $\text{Mat}_k(\mathcal{Y}^d)$  and  $\text{Mat}_k(\mathcal{S}^{(t)})$ . The rows  $\mathbf{Y}_{k,i}^d$  and  
 791  $\mathbf{S}_{k,a}^{(t)}$  correspond to the  $\mathbf{x}_i$  and  $\mathbf{s}_a$  in the one-dimensional  
 792 clustering (13). Then, we obtain the updated assignment by

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

793 provided that  $\mathbf{S}_{k,a}^{(t)}$  is a nonzero vector. Otherwise, if  $\mathbf{S}_{k,a}^{(t)}$  is  
 794 a zero vector, then we make the convention to assign  $z_k^{(t+1)}(i)$   
 795 randomly in  $[r]$ . Full procedures for our angle-based iteration  
 796 are described in Sub-algorithm 2.

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

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

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

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

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

819 Roughly speaking, the vector  $\mathbf{p}(\bar{z})$  represents the raw cluster  
 820 sizes, and  $\mathbf{p}_{\theta}(\bar{z})$  represents the relative cluster sizes weighted  
 821 by degrees. The local stability holds trivially for  $\varepsilon = 0$  based  
 822 on the construction of parameter space (2). The condition (14)  
 823 controls the impact of node degree to the  $\mathbf{p}_{\theta}(\cdot)$  with respect  
 824 to the misclustering rate  $\varepsilon$  and angle gap. Intuitively, the  
 825 condition (14) controls the skewness of degree so that the  
 826 angle between raw cluster size and degree-weighted cluster  
 827

**Algorithm 1** 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 matrices  $\mathbf{U}_{\text{pre},k} = \text{SVD}_r(\text{Mat}_k(\mathcal{Y}))$ ,  $k \in [K]$  and the  $(K-1)$ -mode projections

$$\mathcal{X}_{\text{pre},k} = \mathcal{Y} \times_1 \mathbf{U}_{\text{pre},1} \mathbf{U}_{\text{pre},1}^T \times_2 \dots \times_{k-1} \mathbf{U}_{\text{pre},k-1} \mathbf{U}_{\text{pre},k-1}^T \times_{k+1} \mathbf{U}_{\text{pre},k+1} \mathbf{U}_{\text{pre},k+1}^T \times_{k+2} \dots \times_K \mathbf{U}_{\text{pre},K} \mathbf{U}_{\text{pre},K}^T.$$

2: Compute factor matrices  $\hat{\mathbf{U}}_k = \text{SVD}_r(\text{Mat}_k(\mathcal{X}_{\text{pre},k}))$ ,  $k \in [K]$  and the 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  $\{\hat{x}_j\}_{j \in [r]}$  using weighted  $k$ -means, such that

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

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 the reduced tensor  $\mathcal{Y}_k^d \in \mathbb{R}^{r \times \dots \times r \times p \times r \times \dots \times r}$  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  $J_0 = \{i \in [p] : \|\mathbf{Y}_{i:\}\| = 0\}$ . Set  $z_k^{(t+1)}(i)$  randomly in  $[r]$  for  $i \in J_0$ .

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

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

14: **end for**

15: **end for**

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

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 fixed  $r \geq 1$ ,  $K \geq 2$ , independent noise under the parameter space  $\mathcal{P}$ , and Assumption 1. Assume that the locally linear stability of degree holds in the neighborhood  $\mathcal{N}(z, \varepsilon)$  for all  $\varepsilon \leq E_0$  and some  $E_0 \gtrsim \log^{-1} p$ . 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 the  $k$ -th mode. Suppose  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$ , the SNR  $\geq \tilde{C} p^{-(K-1)} \log p$  for some sufficiently large positive 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 as  $p \rightarrow \infty$ , 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 all parameter assumptions and the results in Theorems 4 and 5, with probability going to 1 as  $p \rightarrow \infty$ , our estimate  $z_k^{(T)}$  achieves exact recovery within

846 polynomial iterations; more precisely,

$$847 \quad z_k^{(T)} = \pi_k \circ z, \quad \text{for all } T \gtrsim \log_{1/\rho} p \text{ and } k \in [K].$$

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

849 Therefore, our combined algorithm is *computationally efficient*  
850 as long as  $\text{SNR} \gtrsim p^{-K/2} \log p$ . Note that, ignoring  
851 the logarithmic term, the minimal SNR requirement,  $p^{-K/2}$ ,  
852 coincides with the computational critical value in Theorem 3.  
853 Therefore, our algorithm is optimal regarding the signal  
854 requirement and lies in the sharpest *computationally efficient*  
855 regime in Fig. 2.

### 856 C. Extension to Bernoulli Observations

857 Bernoulli or network observations are common in multiple  
858 fields. Our iteration Theorem 5 holds for Bernoulli models,  
859 but our initialization Theorem 4 does not. Moreover, our  
860 current dTBM is insufficient to address sparsity with decaying  
861 mean tensor. Here, we provide extra discussions for Bernoulli  
862 initialization and strategies under sparse settings.

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

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

$$872 \quad \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \quad (16)$$

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

877 One possible remedy is to apply singular value decom-  
878 position to the *square unfolding* [27],  $\text{Mat}_{sq}(\cdot)$ , of Bernoulli  
879 tensor  $\mathcal{Y} \in \{0, 1\}^{p_1 \times \dots \times p_K}$ . Specifically, the square  
880 matricization  $\text{Mat}_{sq}(\mathcal{Y}) \in \{0, 1\}^{p^{\lfloor K/2 \rfloor} \times p^{\lceil K/2 \rceil}}$  has entries  
881  $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2) = \mathcal{Y}(i_1, \dots, i_K)$ , where

$$882 \quad j_1 = i_1 + p_1(i_2 - 1) + \dots + p_{\lfloor K/2 \rfloor - 1}(i_{\lfloor K/2 \rfloor} - 1), \\ 883 \quad j_2 = i_{\lceil K/2 \rceil} + p_{\lceil K/2 \rceil}(i_{\lceil K/2 \rceil + 1} - 1) + \dots \\ 884 \quad + p_{\lceil K/2 \rceil} \cdot p_{K-1}(i_K - 1).$$

885 The matrix  $\text{Mat}_{sq}(\mathcal{Y})$  is asymmetric. We interpret  
886  $\text{Mat}_{sq}(\mathcal{Y})$  as the adjacency matrix for a bipartite net-  
887 work with connections between two groups of nodes.  
888 The two groups of nodes in the bipartite network have  
889  $p_1 \cdots p_{\lfloor K/2 \rfloor}$  and  $p_{\lceil K/2 \rceil} \cdots p_K$  nodes, respectively. The  
890 entry  $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2)$  refers to the presence of con-  
891 nection between the nodes indexed by combinations  
892  $(i_1, \dots, i_{\lfloor K/2 \rfloor})$  and  $(i_{\lceil K/2 \rceil}, \dots, i_K)$ . We summarize the  
893 procedure in Algorithm 2.

894 *Proposition 1 (Error for Bernoulli Initialization):*

895 Consider the Bernoulli dTBM in the parameter space  $\mathcal{P}$   
896 with fixed  $r \geq 1, K \geq 2$ . Assume that Assumption 1  
897 holds,  $\boldsymbol{\theta}$  is balanced, and  $\min_{i \in [p]} \theta(i) \geq c$  for some

---

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

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

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

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

---

constant  $c > 0$ . Let  $z_k^{(0)}$  denote the output of Algorithm 2.  
With probability going to 1 as  $p \rightarrow \infty$ , we have

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

*Remark 8 (Comparison with Gaussian model):* 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 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.

*Remark 9 (Comparison With Previous Methods):*

Previous work [9] develops a spectral clustering method for Bernoulli dTBM. [9] adopts a different signal notion based on the singular gap in the core tensor, denoted as  $\Delta_{\text{singular}}$ . By [9, Theorem 1], the spectral method achieves exact recovery with  $\Delta_{\text{singular}} \gtrsim p^{-1/2}$ . However, we are not able to infer the exact recovery of spectral method by our angle-base SNR condition. Consider an order-2 dTBM with  $p > 2, \sigma^2 = 1, \boldsymbol{\theta} = \mathbf{1}_p$ , equal size assignment  $|z^{-1}(a)| = p/r$  for all  $a \in [r]$ , and core matrix equal to the 2-dimensional identity matrix  $\mathbf{S} = \mathbf{I}_2$ . The singular gap under this setting is  $\Delta_{\text{singular}} = \min\{\lambda_1 - \lambda_2, \lambda_2\} = 0$ , where  $\lambda_1 \geq \lambda_2$  are singular values of  $\mathbf{S}$ . In contrast, our angle gap  $\Delta_{\text{min}}^2 = 2$  satisfies the SNR condition in Theorem 5. Then, our algorithm achieves the exact recovery, but the spectral method in [9] fails.

Hence, for fair comparison, we compare the best performance of our algorithm and [9] under the strongest signal setting of each model. Since both methods contain an iteration procedure, we set the iteration number to infinity to avoid the computational error. Considering the largest angle-based SNR  $\asymp 1$  in Theorem 5, our Bernoulli clustering achieves exponential error rate of order  $\exp(-p^{(K-1)})$ ; considering the largest singular

gap  $\Delta_{\text{singular}} \asymp 1$  in Theorem 1 of [9], the spectral clustering has a polynomial error rate of order  $p^{-2}$ . Our algorithm still shows a better theoretical accuracy than the competitive work for Bernoulli observations.

- *Extension to sparse binary dTBMs.* The sparsity is often a popular feature in hypergraphs [9], [16], [28]. Specifically, the sparse binary dTBM assumes that, the entries of  $\mathcal{Y}$  follow independent Bernoulli distributions with the mean

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

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 [28]. Inspired by [28], we adopt the diagonal-deleted HOSVD (D-HOSVD) [9] 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 directions. 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 A.

#### D. Practical Issues

1) *Computational Complexity:* Our two-stage algorithm has a computational cost polynomial in tensor dimension  $p$ . Specifically, the complexity of Sub-algorithm 1 is  $\mathcal{O}(Kp^{K+1} + Kp^K)$ , 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  $\mathcal{Y}_k^d$ , and  $pr^K$  comes from the normalization of  $\mathcal{Y}_k^d$ , the calculation of  $\mathcal{S}^{(t)}$ , and the cluster assignment update in Step 13.

2) *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 (1),

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

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

$$\hat{\theta}(r) = (\hat{\theta}(1, r), \dots, \hat{\theta}(p, r))^T, \quad 1000$$

with  $\hat{\theta}(i, r) = \|\hat{\mathcal{Y}}^d(r)_{i:}\| / \sum_{j:\hat{z}(j,r)=\hat{z}(i,r)} \|\hat{\mathcal{Y}}^d(r)_{j:}\|$ ,  $\hat{\mathcal{Y}}^d(r) = \text{Mat}(\hat{\mathcal{Y}}^d(r))$ ,  $\hat{\mathcal{Y}}^d(r)(i, a_2, \dots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \dots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}$ , 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 NONDEGREE TENSOR BLOCK MODEL

We discuss the connections and differences between dTBM and TBM [13] 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 [13] 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 [13] 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 (4) and the Euclidean-based SNR in [13] as  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$ , respectively:

$$\Delta_{\text{ang}}^2 = 2(1 - \max_{a \neq b \in [r]} \cos(\mathbf{S}_{a:}, \mathbf{S}_{b:})), \quad 1020$$

$$\Delta_{\text{Euc}}^2 = \min_{a \neq b \in [r]} \|\mathbf{S}_{a:} - \mathbf{S}_{b:}\|^2. \quad 1021$$

By Lemma 4 in the Appendix B, we have

$$\Delta_{\text{ang}}^2 \max_{a \in [r]} \|\mathbf{S}_{a:}\|^2 \leq \Delta_{\text{Euc}}^2. \quad 1023$$

The above inequality indicates that 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

$$\mathbf{S} = \begin{pmatrix} p^{(\gamma+1)/2} + 2 & 2 \\ 2 & 4 \end{pmatrix}, \quad \text{with } \gamma \leq -1. \quad 1035$$

The core matrix  $\mathbf{S}$  lies in the parameter spaces of TBM and our dTBM. Here, the constraint  $\gamma \leq -1$  is added to ensure the bounded condition of  $\mathbf{S}$  in our parameter

space in (2). The angle-based and Euclidean-based signal levels of  $\mathcal{S}$  are

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

We conclude that TBM with  $\mathcal{S}$  achieves exact recovery with a polynomial-time algorithm; see [13, Theorem 4]. By contrast, the dTBM with the same  $\mathcal{S}$  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 [13] and compare with ours.

*Statistical critical value:*

Ours:  $\Delta_{\text{ang}}^2 \lesssim p^{-(K-1)} \Rightarrow$  statistically impossible;

$\Delta_{\text{ang}}^2 \gtrsim p^{-(K-1)} \Rightarrow$  MLE achieves exact recovery;

Han's:  $\Delta_{\text{Euc}}^2 \lesssim p^{-(K-1)} \Rightarrow$  statistically impossible;

$\Delta_{\text{Euc}}^2 \gtrsim p^{-(K-1)} \Rightarrow$  MLE achieves exact recovery.

*Computational critical value:*

Ours:  $\Delta_{\text{ang}}^2 \lesssim p^{-K/2} \Rightarrow$  computationally impossible;

$\Delta_{\text{ang}}^2 \gtrsim p^{-K/2} \Rightarrow$  computationally efficient;

Han's:  $\Delta_{\text{Euc}}^2 \lesssim p^{-K/2} \Rightarrow$  computationally impossible;

$\Delta_{\text{Euc}}^2 \gtrsim p^{-K/2} \Rightarrow$  computationally efficient.

The above comparison reveals four major differences.

First, none of our results in Section III are corollaries of [13]. Both models show the similar conclusion but under different conditions. While the TBM impossibility [13] provides a necessary condition for our dTBM impossibility, we find that such a condition is often loose. There exists a regime of  $\mathcal{S}$  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 [13] 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 dTBM-specific techniques to handle the extra degree heterogeneity. In our Theorem 2, we construct a special nontrivial degree heterogeneity to establish the lower bound for arbitrary core tensor with small angle gap, while, TBM [13] considers the constructions without degree parameter. In our Theorem 3, we construct a rank-2 tensor to relate HPC conjecture to  $\Delta_{\text{ang}}^2$ , while TBM [13] constructs a rank-1 tensor to relate HPC conjecture to  $\Delta_{\text{Euc}}^2$ . The asymptotic nonequivalence between

$\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$  renders our proof technically more involved.

Last, we discuss the statistical impossibility statements. Our Theorem 2 implies the statistical impossibility whenever the core tensor  $\mathcal{S}$  leads to an angle-based SNR below the critical value, while, Theorem 6 in [13] implies the worst case statistical impossibility for a particular core tensor  $\mathcal{S}$  with Euclidean-based SNR below the statistical limit. Hence, our Theorem 2 shows a stronger statistical impossibility for dTBM than that presented in TBM [13, Theorem 6]. However, inspecting the proof of [13], the proof of Theorem 6 indeed implies a stronger TBM impossibility statement for arbitrary core tensor; i.e., when  $\gamma < -(K - 1)$

$$\liminf_{p \rightarrow \infty} \inf_{\mathcal{S} \in \mathcal{P}_{\mathcal{S}, \text{TBM}} \cap \{\Delta_{\text{Euc}}^2 = p^\gamma\}} \inf_{\hat{z}_{\text{stats}}} \sup_{z \in \mathcal{P}_{z, \text{TBM}}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z)] \geq 1,$$

where  $\mathcal{P}_{\mathcal{S}, \text{TBM}}$  and  $\mathcal{P}_{z, \text{TBM}}$  refer to the space for core tensor  $\mathcal{S}$  and assignment  $z$  under TBM, respectively. Again, in terms of the strong statistical impossibility, both models show the similar conclusion but under different conditions. Since two impossibilities consider different core tensor regimes with nonequivalent  $\Delta_{\text{ang}}^2$  and  $\Delta_{\text{Euc}}^2$ , we emphasize that different proof techniques are required to obtain these similar conclusions. See our proof sketch in Section VIII-A, Appendices B-D and B-E for detail technical differences.

- *Algorithms.* Both [13] 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 [29], [30]. 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). The CER between  $(\hat{z}, z)$  is equivalent to misclustering error  $\ell(\hat{z}, z)$  up to constant multiplications [31], and a lower CER indicates a better performance.

We generate order-3 tensors with *assortative* [15] core tensors to control SNR; i.e., we set  $\mathcal{S}_{aaa} = s_1$  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 (4). Hence, we easily adjust SNR via varying  $\alpha$ . 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

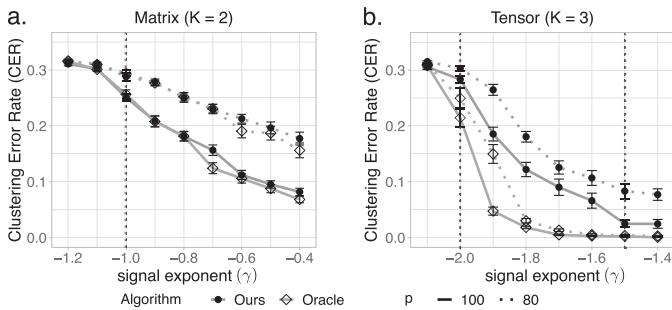


Fig. 4. SNR phase transitions for clustering in dTBM with  $p = \{80, 100\}$ ,  $r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

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  $|X_i| \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ ,  $i \in [p]$  and normalize  $\theta$  to satisfy (2). 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. Fig. 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, Fig. 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  $\gamma_{\text{comp}} = -1.5$ . Fig. 4 confirms the existence of the statistical-computational gap in our Theorems 2 and 3.

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  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$ . Fig. 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.

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 ( $p = 50, \sigma^2 = 1$ ), and the accuracy immediately increases with larger dimension  $p = 80$ . The improvement follows from the fact that a larger dimension  $p$  indicates

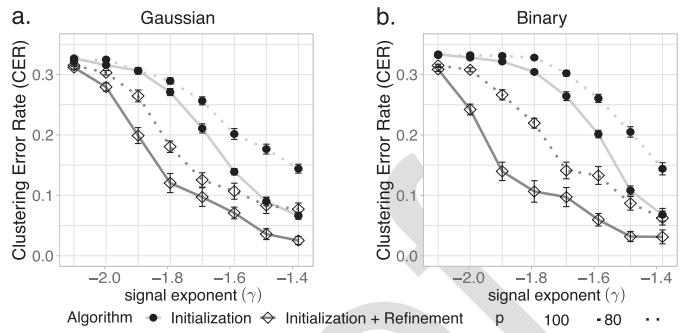


Fig. 5. CER versus signal exponent ( $\gamma$ ) for initialization only and for combined algorithm. We set  $p = \{80, 100\}$ ,  $r = 5$ ,  $\gamma \in [-2.1, -1.4]$  under (a) Gaussian models and (b) Bernoulli models.

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** [13]: High-order clustering algorithm developed for nondegree tensor block models;
- **SCORE** [9]: 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 [9]. The methods **SCORE** and **HOSVD+** are designed for degree models, whereas **HOSVD** and **HLloyd** are designed for nondegree 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  $\gamma \in [-1.5, -1.1]$ . Fig. 6 shows that our method **dTBM** outperforms all other algorithms. The suboptimality of **SCORE** and **HOSVD+** indicates the necessity of local iterations on the clustering. Furthermore, Fig. 6 shows the inadequacy of nondegree algorithms in the presence of mild degree heterogeneity. The experiment demonstrates the benefits of addressing heterogeneity in higher-order clustering tasks.

The only exception in Fig. 6 is the slightly better performance of **HLloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLloyd** 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

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

| Settings                           | $p = 50, \sigma^2 = 0.25$ |           | $p = 50, \sigma^2 = 1$ |           | $p = 80, \sigma^2 = 0.25$ |      | $p = 80, \sigma^2 = 1$ |           |
|------------------------------------|---------------------------|-----------|------------------------|-----------|---------------------------|------|------------------------|-----------|
|                                    | 2                         | 4         | 2                      | 4         | 2                         | 4    | 2                      | 4         |
| True cluster number $r$            |                           |           |                        |           |                           |      |                        |           |
| 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) |

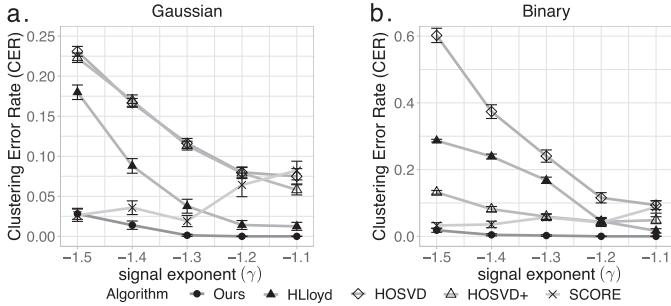


Fig. 6. CER versus signal exponent (denoted  $\gamma$ ) for different methods. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under (a) Gaussian and (b) Bernoulli models.

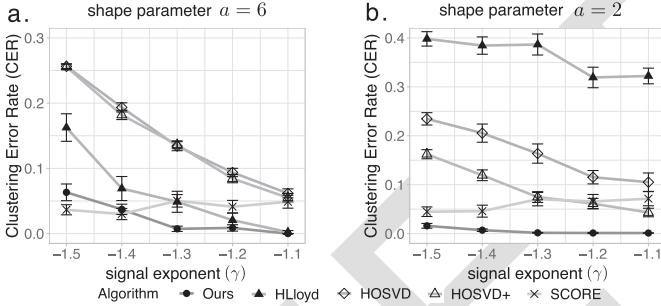


Fig. 7. CER comparison versus signal exponent (denoted  $\gamma$ ) under (a) low (shape parameter  $a = 6$ ) (b) high (shape parameter  $a = 2$ ) degree heterogeneity. We set  $p = 100, r = 5, \gamma \in [-1.5, -1.1]$  under Gaussian model.

1229  $a$  is called *shape parameter*. We vary  $a \in \{2, 6\}$  and choose  $b$   
1230 such that  $\mathbb{E}X = a(a-1)^{-1}b = 1$  for  $X$  following  $\text{Pareto}(a, b)$ .  
1231 Note that a smaller  $a$  leads to a larger variance in  $\theta$  and hence a  
1232 larger degree heterogeneity. We consider the Gaussian model  
1233 under low ( $a = 6$ ) and high ( $a = 2$ ) degree heterogeneity.  
1234 Fig. 7 shows that the errors for nondegree algorithms (**HLloyd**,  
1235 **HOSVD**) increase with degree heterogeneity. In addition, the  
1236 advantage of **HLloyd** over **HOSVD+** disappears with higher  
1237 degree heterogeneity.

1238 The last experiment investigates the effects of degree hetero-  
1239 geneity to clustering performance. We fix the signal exponent  
1240  $\gamma = -1.2$  and vary the extent of degree heterogeneity.  
1241 In this experiment, we generate  $\theta$  from Pareto distribution  
1242 prior to normalization. We vary the shape parameter  $a \in$   
1243  $[3, 6]$  in the Pareto distribution to investigate a range of  
1244 degree heterogeneities. Fig. 8 demonstrates the stability of  
1245 degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over  
1246 the entire range of degree heterogeneity under considera-  
1247 tion. In contrast, nondegree algorithms (**HLloyd**, **HOSVD**)  
1248 show poor performance with large heterogeneity, especially in  
1249 Bernoulli cases. This experiment, again, highlights the benefit  
1250 of addressing degree heterogeneity in higher-order clustering.

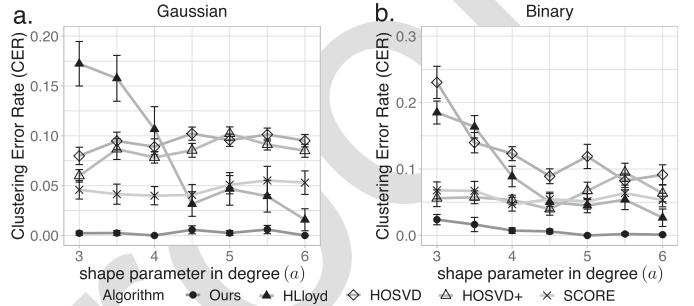


Fig. 8. CER versus shape parameter in degree ( $a \in [3, 6]$ ) for different methods. We set  $p = 100, r = 5, \gamma = -1.2$  under (a) Gaussian and (b) Bernoulli models.

## 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 [32]. We preprocess the original dataset following [33] 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 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 [34]. Because of the symmetry in the data, the estimated brain node clustering results are the same on the first and second modes. Fig. 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.

Fig. 10 illustrates the estimated core tensor  $\hat{\mathcal{S}}$  with estimated clustering, and Fig. 11 visualizes the average brain connections and the connection enrichment in contrast to average networks in each group. In general, we find that the

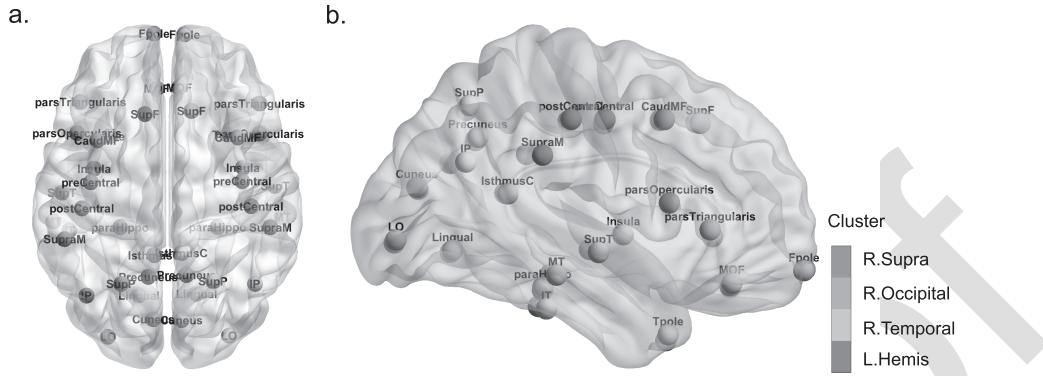


Fig. 9. Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

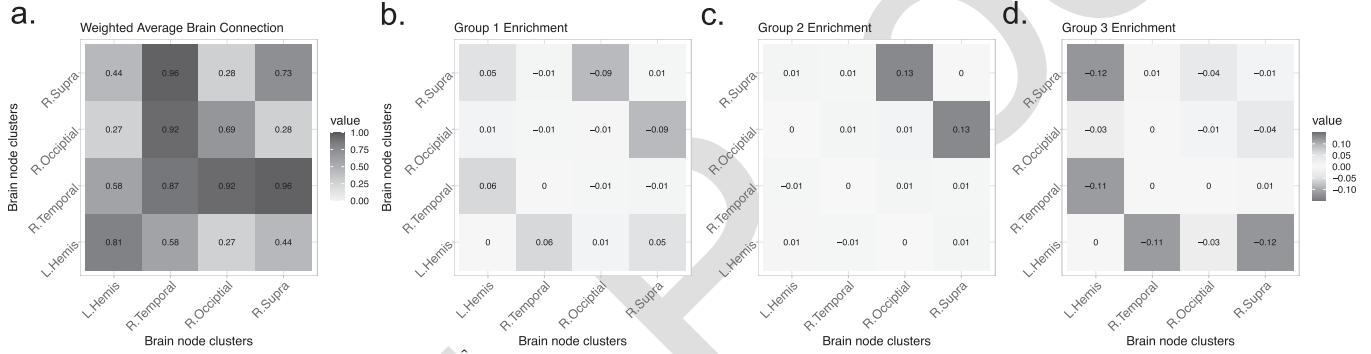


Fig. 10. Mode 3 slices of estimated core tensor  $\hat{S}$ . (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the difference between each slice of  $\hat{S}$  and the averaged slice.

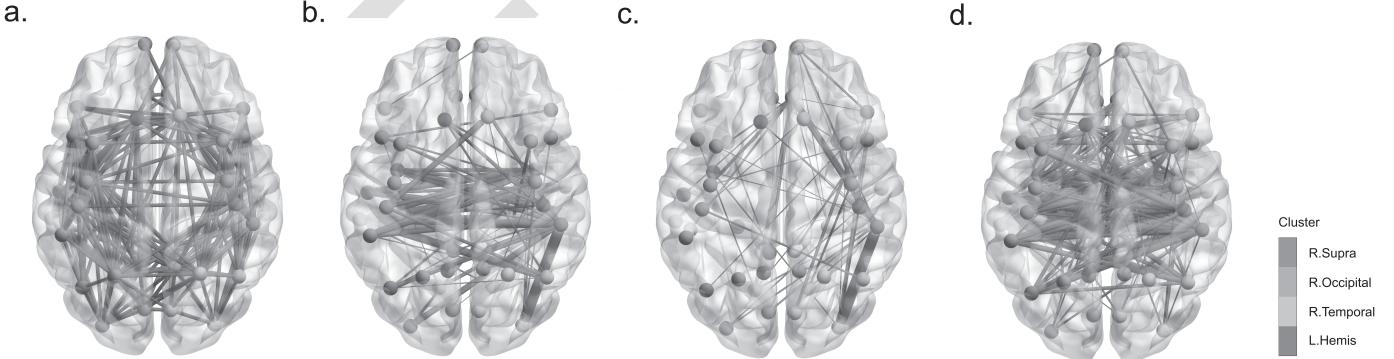


Fig. 11. Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain network enrichments in Groups 1-3. Red edges represent the positive enrichment and blue edges represent the negative enrichment.

inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Fig. 10a). 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 (Fig. 10b), while group 4 with only 36% females exhibits a reduction (Fig. 10d). This result agrees with previous findings in [34]. The enrichment on the back-front connection is also recognized in group 3 (Fig. 10c). The interpretive patterns in our results demonstrate the usefulness of our clustering methods 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 [35]. 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

1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309

TABLE IV  
CLUSTERING ERRORS (MEASURED BY CER) FOR VARIOUS METHODS IN  
THE ANALYSIS OF PERU LEGISLATION DATASET

| Method | dTBM  | HOSVD | HOSVD+ | HLloyd | SCORE |
|--------|-------|-------|--------|--------|-------|
| CER    | 0.116 | 0.22  | 0.213  | 0.149  | 0.199 |

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.

### VIII. PROOF SKETCHES

In this section, we provide the proof sketches for the main Theorem 2 (Impossibility), Theorem 3 (Impossibility), and Theorems 4-5. Detail proofs and extra theoretical results are provided in Appendix B.

#### A. Proof Sketch of Theorem 2 (Impossibility) and Theorem 3 (Impossibility)

The proofs of impossibility in Theorems 2 and 3 share the same proof idea with [13, Theorems 6 and 7] and [15, Theorem 2]. In both proofs of statistical and computational impossibilities, the key idea is to construct a particular set of parameters to lower bound the minimax rate. Specifically, for statistical impossibility in Theorem 2, we construct a particular  $(z_{\text{stats}}^*, \theta_{\text{stats}}^*) \in \mathcal{P}_{z,\theta}$  such that for all  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$

$$\begin{aligned} & \inf_{\hat{z}_{\text{stats}}} \sup_{(z,\theta) \in \mathcal{P}_{z,\theta}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z)] \\ & \geq \inf_{\hat{z}_{\text{stats}}} \mathbb{E}[p\ell(\hat{z}_{\text{stats}}, z_{\text{stats}}^*) | (z_{\text{stats}}^*, \mathcal{S}^*, \theta_{\text{stats}}^*)] \geq 1; \end{aligned} \quad (20)$$

for computational impossibility in Theorem 3, we construct a particular  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*) \in \mathcal{P}(\gamma)$  such that

$$\begin{aligned} & \inf_{\hat{z}_{\text{comp}}} \sup_{(z,\mathcal{S},\theta) \in \mathcal{P}(\gamma)} \mathbb{E}[p\ell(\hat{z}_{\text{comp}}, z)] \\ & \geq \inf_{\hat{z}_{\text{comp}}} \mathbb{E}[p\ell(\hat{z}_{\text{comp}}, z_{\text{comp}}^*) | (z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)] \geq 1. \end{aligned}$$

The constructions of  $(z_{\text{stats}}^*, \theta_{\text{stats}}^*)$  and  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)$  are the most critical steps. With good constructions, the lower bound “ $\geq 1$ ” can be verified by classical statistical conclusions (e.g. Neyman-Pearson Lemma) or prior work (e.g. HPC Conjecture).

A notable detail in the proof of statistical impossibility is the arbitrariness of  $\mathcal{S}^*$ . The first infimum over  $\mathcal{P}_S(\gamma)$  in the minimax rate (10) requires that the lower bound (20) holds for any  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$ . The arbitrary choice of  $\mathcal{S}^*$  brings extra difficulties in the parameter construction, and consequently a nontrivial  $\theta_{\text{stats}}^* \neq 1$  is chosen to address the arbitrariness. Previous TBM construction in the proof of [13, Theorem 6] with  $\theta_{\text{stats}}^* = 1$  is no longer applicable in our case. Meanwhile, our construction  $(z_{\text{comp}}^*, \mathcal{S}_{\text{comp}}^*, \theta_{\text{comp}}^*)$  leads to a rank-2 mean tensor to relate the HPC Conjecture while TBM [13, Theorem

7] constructs a rank-1 mean tensor. Hence, we emphasize that dTBM-specific techniques are required to obtain our impossibility results, though the proof idea is common for minimax lower bound analysis.

#### B. Proof Sketch of Theorem 4

The proof of Theorem 4 is inspired by the proof idea of [15, 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

$$\begin{aligned} \ell(z^{(0)}, z) & \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ & \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K} \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ & \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \end{aligned} \quad (21)$$

where  $\mathcal{X} = \mathbb{E}\mathcal{Y}$  is the true mean. The first inequality in (21) holds with the assumption  $\min_{i \in [p]} \theta(i) \geq c > 0$  in Theorem 4. The second inequality relies on the key Lemma 1, which indicates

$$\min_{z(i) \neq z(j)} \|[\mathbf{X}_{i:}]^s - [\mathbf{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \quad (22)$$

where  $\mathbf{X} = \text{Mat}(\mathcal{X})$ . The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (22) (or the proof of Lemma 1), in which the proof of [15] is no longer applicable due to different angle gap assumption in our dTBM. To address the angle gap notion, we develop the extra padding technique in Lemma 5 and balance assumption (6). Last, we finish the proof of Theorem 4 by showing the third inequality of (21) using [13, Proposition 1].

#### C. Proof Sketch of Theorem 5

The proof of Theorem 5 is inspired by the proof idea of [13, 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)

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

We show that  $L^{(t)}$  provides an upper bound for the misclustering error of interest via the inequality  $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2}$  in Lemma 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

1397  $t$ -th iteration quantities  $\mathbf{W}^{(t)}, \mathbf{V}^{(t)}$  corresponding to  $\mathbf{M}^{(t)}$  (or  
 1398 equivalently  $z^{(t)}$ ). To evaluate  $L^{(t+1)}$ , we prove the bound

$$\begin{aligned} 1399 & \mathbb{1} \left\{ z^{(t+1)}(i) = b \right\} \\ 1400 & = \mathbb{1} \left\{ \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right\} \\ 1401 & \leq A_{ib} + B_{ib}, \end{aligned} \quad (23)$$

1402 where  $\mathbf{Y} = \text{Mat}(\mathcal{Y})$ ,  $\mathbf{S} = \text{Mat}(\mathcal{S})$ ,  $\mathbf{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$  and

$$\begin{aligned} 1403 & A_{ib} = \mathbb{1} \left\{ \left\langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \lesssim -\|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \right\}, \\ 1404 & B_{ib} = \mathbb{1} \left\{ \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2 \lesssim F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\}. \end{aligned}$$

1405 The terms  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$  are controlled by  $z^{(t)}, \mathcal{S}^{(t)}$ ; see the  
 1406 detailed definitions in (68), (69), (70). Note that the event  $A_{ib}$   
 1407 only involves the oracle estimator independent of  $t$ , while all  
 1408 the terms related to the  $t$ -th iteration are in  $B_{ib}$ . Thus, the  
 1409 inequality (23) decomposes the misclustering loss in the  $(t +$   
 1410 1)-th iteration into the oracle loss and the loss in  $t$ -th iteration.  
 1411 This decomposition leads to the separation of statistical error  
 1412 and computational error in the final upper bound of Theorem 5.  
 1413 Specifically, we prove the contraction inequality

$$\begin{aligned} 1414 & L^{(t+1)} \leq M\xi + \rho L^{(t)}, \\ 1415 & \text{with } \xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \end{aligned} \quad (24)$$

1416 where  $M$  is a positive constant,  $\rho \in (0, 1)$  is the contraction  
 1417 parameter, and we call  $\xi$  the oracle loss. Controlling the  
 1418 probability of event  $B_{ib}$  and obtaining the  $\rho L^{(t)}$  term in the  
 1419 right hand side of (24) are the most challenging parts in  
 1420 the proof of Theorem 5. Note that the true and estimated  
 1421 core tensors are involved via their normalized rows such  
 1422 as  $\mathbf{S}_{a:}^s, \tilde{\mathbf{S}}_{a:}^s, [\mathbf{S}_{a:}^{(t)}]^s$ . The Cartesian coordinate based analysis  
 1423 in [13] is no longer applicable in our case. Instead, we use  
 1424 the polar-coordinate based analysis and the geometry property  
 1425 of trigonometric functions to derive the high probability upper  
 1426 bounds for  $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ .

1427 Further, by sub-Gaussian concentration, we prove the high  
 1428 probability upper bound for oracle loss

$$\xi \lesssim \text{SNR}^{-1} \exp \left( -\frac{p^{K-1} \text{SNR}}{r^{K-1}} \right). \quad (25)$$

1430 Combining the decomposition (24) and the oracle bound (25),  
 1431 we finish the proof of Theorem 5.

1432 The proof of MLE error shares the similar idea as Theorems 4-5. We first show a weaker polynomial rate for MLE  
 1433 and then improve the rate from polynomial to exponential  
 1434 through the iterations. The only difference is that the MLE  
 1435 remains the same over iterations due to its global optimality.  
 1436 See Appendix B-G for the detailed proof.

## APPENDIX A

### ADDITIONAL NUMERICAL EXPERIMENTS

#### A. Bernoulli Phase Transition

1441 The first additional experiment verifies the  
 1442 statistical-computational gap in Section III under the Bernoulli  
 1443 model. Consider the Bernoulli model with  $p = \{80, 100\}$ ,

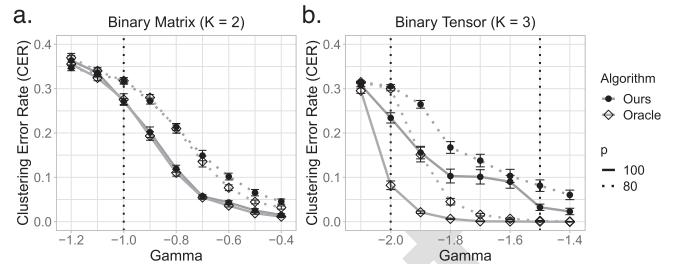


Fig. 12. SNR phase transitions for Bernoulli dTBM with  $p = \{80, 100\}$ ,  $r = 5$  under (a) matrix case with  $\gamma \in [-1.2, -0.4]$  and (b) tensor case with  $\gamma \in [-2.1, -1.4]$ .

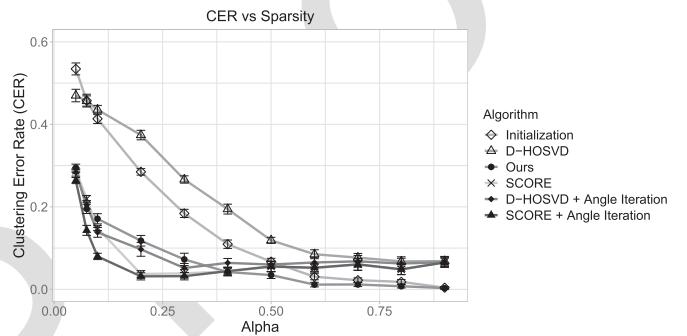


Fig. 13. CER comparison versus sparsity parameter  $\alpha_p$  in  $[0.05, 0.9]$ . We set  $p = 100, r = 5$  and  $\gamma = -1.2$  under sparse binary dTBM.

1444  $r = 5$ . We vary  $\gamma$  in  $[-1.2, -0.4]$  and  $[-2.1, -1.4]$  for  
 1445 matrix ( $K = 2$ ) and tensor ( $K = 3$ ) clustering, respectively.  
 1446 We approximate MLE using an oracle estimator, i.e., the  
 1447 output of Sub-algorithm 2 initialized from the true assignment.  
 1448 Fig. 12 shows a similar pattern as Fig. 4. The algorithm and  
 1449 oracle estimators have no gap in the matrix case, while an  
 1450 error gap emerges between the critical values  $\gamma_{\text{stat}} = -2$  and  
 1451  $\gamma_{\text{comp}} = -1.5$  in the tensor case. Fig. 4 suggests the  
 1452 statistical-computational gap in Bernoulli models.

#### B. Sparsity

1453 The second additional experiment evaluates the algorithm  
 1454 performances under the sparse binary dTBM (18). We fix the  
 1455 signal exponent  $\gamma = -1.2$  and vary the sparsity parameter  
 1456  $\alpha_p \in [0.05, 0.9]$ . A smaller  $\alpha_p$  leads to a higher probability  
 1457 of zero entries in the observation. In addition to the three  
 1458 algorithms mentioned in Section VI-B (denoted **Initialization**,  
 1459 **dTBM**, and **SCORE**), we consider other three algorithms based  
 1460 on the discussion in Section IV-C:

- **D-HOSVD**, the diagonal-deleted HOSVD in [9];
- **D-HOSVD + Angle**, the combined algorithm of our  
 1463 angle-based iteration with initialization from **D-HOSVD**;
- **SCORE + Angle**, the combined algorithms of our  
 1465 angle-based iteration with initialization from **SCORE**.

1466 Fig. 13 shows a slightly larger error in **dTBM** than that in  
 1467 **SCORE**, **D-HOSVD + Angle**, and **SCORE + Angle** under the  
 1468 sparse setting with  $\alpha_p < 0.3$ . The small gap between **dTBM**  
 1469 and other sparse-specific methods implies the robustness of our  
 1470 algorithm. In addition, comparing **SCORE** versus **SCORE + Angle**  
 1471 (or **D-HOSVD** versus **D-HOSVD + Angle**) indicates the  
 1472 benefit of our angle iterations under the sparse dTBM. In the  
 1473

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 B PROOFS

We provide the proofs for all the theorems in our main paper. In each subsection, 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.

#### 1) Preliminaries:

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

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

• For a vector  $\mathbf{a}$ , let  $\mathbf{a}^s := \mathbf{a}/\|\mathbf{a}\|$  denote the normalized vector. We make the convention that  $\mathbf{a}^s = \mathbf{0}$  if  $\mathbf{a} = \mathbf{0}$ .

• For a matrix  $\mathbf{A} \in \mathbb{R}^{n \times m}$ , let  $\mathbf{A}^{\otimes K} := \mathbf{A} \otimes \cdots \otimes \mathbf{A} \in \mathbb{R}^{n^K \times m^K}$  denote the Kronecker product of  $K$  copies of matrices  $\mathbf{A}$ .

• For a matrix  $\mathbf{A}$ , let  $\|\mathbf{A}\|_\sigma$  denote the spectral norm of matrix  $\mathbf{A}$ , which is equal to the maximal singular value of  $\mathbf{A}$ ; let  $\lambda_k(\mathbf{A})$  denote the  $k$ -th largest singular value of  $\mathbf{A}$ ; let  $\|\mathbf{A}\|_F$  denote the Frobenius norm of matrix  $\mathbf{A}$ .

2) 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

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where } \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \mathbf{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \mathbf{M}_K, \quad (26)$$

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  $\mathbf{M}_k \in \{0, 1\}^{p_k \times r_k}$  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  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ ,

$$\begin{aligned} \mathcal{P}(\{r_k\}) &= \left\{ (\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) : \boldsymbol{\theta}_k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} |z_k^{-1}(a)| \leq \frac{c_2 p_k}{r_k}, \right. \\ &\quad c_3 \leq \|\mathbf{S}_{k,a,:}\| \leq c_4, \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|_1 = |z_k^{-1}(a)|, \\ &\quad \left. \text{for all } a \in [r_k], k \in [K] \right\}. \end{aligned} \quad (27)$$

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

$$\min_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\| = (1 + o(1)) \max_{a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{-1}(a)}\|. \quad (28)$$

We also consider the generalized Assumption 1 on angle gap.

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

$$\Delta_{\min} := \min_{k \in [K]} \min_{a \neq b \in [r_k]} \|\mathbf{S}_{k,a,:}^s - \mathbf{S}_{k,b,:}^s\| > 0. \quad (29)$$

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

$$\mathcal{P}(\gamma) = \mathcal{P}(\{r_k\}) \cap \{\mathcal{S} \text{ satisfies } \text{SNR} = p^\gamma \text{ and } p_k \asymp p, k \in [K]\}. \quad (30)$$

### B. Proof of Theorem 1

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

$$\begin{aligned} \mathcal{X} &= \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \mathbf{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \mathbf{M}'_K \\ &= \mathcal{S}' \times_1 \boldsymbol{\Theta}'_1 \mathbf{M}'_1 \times_2 \cdots \times_K \boldsymbol{\Theta}'_K \mathbf{M}'_K, \end{aligned} \quad (31)$$

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 (26) violates Assumption 2. Note that  $\Delta_{\min}^2 = 1$  when there exists  $k \in [K]$  such that  $r_k = 1$ . Hence, we consider the case that  $r_k \geq 2$  for all  $k \in [K]$ . Without loss of generality, we assume  $\|\mathbf{S}_{1,1,:}^s - \mathbf{S}_{1,2,:}^s\| = 0$ .

By constraints in parameter space (27), 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

$$\mathcal{S} = \mathcal{S}_0 \times_1 \mathbf{C} \mathbf{R}, \quad (32)$$

where  $\mathbf{C} = \text{diag}(1, c, 1, \dots, 1) \in \mathbb{R}^{r_1 \times r_1}$  and

$$\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{1}_{r_1-2} \end{pmatrix} \in \mathbb{R}^{r_1 \times (r_1-1)}. \quad (33)$$

Let  $\mathbf{D} = \text{diag}(1 + c, 1, \dots, 1) \in \mathbb{R}^{r_1-1 \times r_1-1}$ . Consider the parameterization  $\mathbf{M}'_1 = \mathbf{M}_1 \mathbf{R}$ ,  $\mathcal{S}' = \mathcal{S}_0 \times_1 \mathbf{D}$ , and

$$\theta'_1(i) = \begin{cases} \frac{1}{1+c} \theta_1(i) & i \in z_1^{-1}(1), \\ \frac{c}{1+c} \theta_1(i) & i \in z_1^{-1}(2), \\ \theta_1(i) & \text{otherwise,} \end{cases} \quad (34)$$

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 (26) 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  $k \in [K]$ . When  $r_k = 1$ , all possible  $\mathbf{M}_k$ 's are equal to the vector  $\mathbf{1}_{p_k}$ , and the uniqueness holds trivially. Hence, we consider the case that  $r_k \geq 2$ . Without loss of generality, we consider  $k = 1$  with  $r_1 \geq 2$  and show the uniqueness of 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  $\|(\mathbf{S}')_{1,z'_1(i)}^s - (\mathbf{S}')_{1,z'_1(j)}^s\| = 0$  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  $\|(\mathbf{S}')_{1,z'_1(i)}^s - (\mathbf{S}')_{1,z'_1(j)}^s\| \neq 0$  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

$$\frac{\mathbf{X}_{1,z_1(i)}^s}{\mathbf{X}_{1,z_1(j)}^s} = \frac{\mathbf{X}'_{1,z_1(i)}^s}{\mathbf{X}'_{1,z_1(j)}^s}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta'_1(j)}{\theta'_1(i)}. \quad (29)$$

Let  $\theta'_1(j) = c\theta_1(j)$  for some positive constant  $c$ . By equation (29), 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 (28), we have

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

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 (26) with  $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$  and  $r_k \geq 2, k \in [K]$ . Then, for any  $k \in [K]$  and index pair  $(i, j) \in [p_k]^2$ , we have

$$\begin{aligned} \|\mathbf{S}_{k,z_k(i)}^s - \mathbf{S}_{k,z_k(j)}^s\| &= 0 \quad \text{if and only if} \\ \|\mathbf{X}_{k,z_k(i)}^s - \mathbf{X}_{k,z_k(j)}^s\| &= 0. \end{aligned}$$

*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

$$\mathbf{X}_{j:} = \theta_1(j) \mathbf{S}_{z_1(j):} [\boldsymbol{\Theta}_2 \mathbf{M}_2 \otimes \cdots \otimes \boldsymbol{\Theta}_K \mathbf{M}_K]^T. \quad (1620)$$

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

$$\|\mathbf{a}^s - \mathbf{b}^s\| = \|(c_1 \mathbf{a})^s - (c_2 \mathbf{b})^s\|. \quad (1623)$$

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

$$\|\mathbf{S}_{z_1(i):}^s - \mathbf{S}_{z_1(j):}^s\| = 0 \quad \text{if and only if} \quad (1626)$$

$$\left\| \left[ \mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T \right]^s - \left[ \mathbf{S}_{z_1(j):} \tilde{\mathbf{M}}^T \right]^s \right\| = 0. \quad (1627)$$

$$(\Leftarrow) \text{ Suppose } \left\| \left[ \mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T \right]^s - \left[ \mathbf{S}_{z_1(j):} \tilde{\mathbf{M}}^T \right]^s \right\| = 0. \quad (1628)$$

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

$$\mathbf{S}_{z_1(i):} = \mathbf{S}_{z_1(i):} \tilde{\mathbf{M}}^T \left[ \tilde{\mathbf{M}} \left( \tilde{\mathbf{M}}^T \tilde{\mathbf{M}} \right)^{-1} \right], \quad (1631)$$

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

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

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

$$\begin{aligned} \mathbf{w}_{a_2 \dots a_K} &= [\theta_{z^{-1}(a_2)}^T \otimes \cdots \otimes \theta_{z^{-1}(a_K)}^T] \in \mathbb{R}^{|z^{-1}(a_2)| \times \cdots \times |z^{-1}(a_K)|}, \end{aligned} \quad (1646)$$

for all  $a_k \in [r], k = 2, \dots, K$ , where  $\otimes$  denotes the Kronecker product. Therefore, we have  $\mathbf{X}_{i:} = \theta(i) \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(i)})$  where  $\mathbf{w} = \{\mathbf{w}_{a_2, \dots, a_K}\}_{a_k \in [r], k \in [K]/\{1\}}$ . Specifically, we have  $\|\mathbf{w}_{a_2, \dots, a_K}\|^2 = \prod_{k=2}^K \|\theta_{z^{-1}(a_k)}\|^2$ , and by the balanced assumption (6) we have

$$\max_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2 = (1 + o(1)) \min_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2. \quad (30)$$

Consider the inner product of  $\mathbf{X}_{i:}$  and  $\mathbf{X}_{j:}$  for  $z(i) \neq z(j)$ . By the definition of weighted padding operator (56) and the balanced assumption (30), we have

$$\begin{aligned} &\langle \mathbf{X}_{i:}, \mathbf{X}_{j:} \rangle \\ &= \theta(i)\theta(j) \langle \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(i)}), \text{Pad}_{\mathbf{w}}(\mathbf{S}_{z(j)}) \rangle \\ &= \theta(i)\theta(j) \min_{(a_2, \dots, a_K)} \|\mathbf{w}_{a_2, \dots, a_K}\|^2 \langle \mathbf{S}_{z(i)}, \mathbf{S}_{z(j)} \rangle (1 + o(1)). \end{aligned} \quad (1658)$$

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

Then, we have

$$\begin{aligned} \cos(\mathbf{S}_{z_1(i):}, \mathbf{S}_{z_1(j):}) &= \frac{\langle \mathbf{S}_{z_1(i):}, \mathbf{S}_{z_1(j):} \rangle}{\|\mathbf{S}_{z_1(i):}\| \|\mathbf{S}_{z_1(j):}\|} \\ &= (1 + o(1)) \frac{\langle \mathbf{X}_{i:}, \mathbf{X}_{j:} \rangle}{\|\mathbf{X}_{i:}\| \|\mathbf{X}_{j:}\|} \\ &= (1 + o(1)) \cos(\mathbf{X}_{i:}, \mathbf{X}_{j:}), \end{aligned}$$

where the second inequality follows by the balance assumption on  $\theta$ .

Further, notice that  $\|\mathbf{v}_1^s - \mathbf{v}_2^s\|^2 = 2(1 - \cos(\mathbf{v}_1, \mathbf{v}_2))$ . For all  $i, j$  such that  $z(i) \neq z(j)$ , when  $p \rightarrow \infty$ , we have

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

Combining the inequalities (12) and (12) in the proof of Theorem 2 in [15], we have

$$\begin{aligned} \inf_{\hat{z}_1} \mathbb{E} [\ell(\hat{z}_1, z_1^*) | (z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*)] &\geq \\ \frac{C}{r^3 |T_1^c|} \sum_{i \in T_1^c} \inf_{\hat{z}_1(i)} \{ \mathbb{P}[\hat{z}_1(i) = 1 | z_1^*(i) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ &\quad + \mathbb{P}[\hat{z}_1(i) = 2 | z_1^*(i) = 1, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \}, \end{aligned} \quad (31)$$

where  $C$  is some positive constant,  $\hat{z}_1$  on the left hand side denote the generic assignment functions in  $\mathcal{P}(\gamma)$ , and the infimum on the right hand side is taken over the generic assignment function family of  $\hat{z}_1(i)$  for all nodes  $i \in T_1^c$ . Here, the factor  $r^3 = r \cdot r^2$  in (31) 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.

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

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

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):* Consider the general asymmetric dTBM (26) in the special case that  $p_k = p$  and  $r_k = r$  for all  $k \in [K]$  with  $K \geq 2$ ,  $2 \leq r \lesssim p^{1/3}$  as  $p \rightarrow \infty$ . For simplicity, we show the minimax rate for the estimation on the first mode  $\hat{z}_1$ ; the proof for other modes are essentially the same.

To prove the minimax rate (10), it suffices to take an arbitrary  $\mathcal{S}^* \in \mathcal{P}_S(\gamma)$  with  $\gamma < -(K-1)$  and construct  $(z_k^*, \boldsymbol{\theta}_k^*)$  such that

$$\inf_{\hat{z}_1} \mathbb{E} [p\ell(\hat{z}_1, z_1^*) | (z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*)] \geq 1.$$

We first define a subset of indices  $T_k \subset [p_k]$ ,  $k \in [K]$  in order to avoid the complication of label permutation. Based on [13, Proof of Theorem 6], we consider the restricted family of  $\hat{z}_k$ 's for which the following three conditions are satisfied:

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

for all  $k \in [K]$ . Now, we consider the construction:

- (i)  $\{z_k^*\}$  satisfies properties (a)-(c) with misclassification sets  $T_k^c$  for all  $k \in [K]$ ;
- (ii)  $\{\boldsymbol{\theta}_k^*\}$  such that  $\boldsymbol{\theta}_k^*(i) \leq \sigma r^{(K-1)/2} p^{-(K-1)/2}$  for all  $i \in T_k^c$ ,  $k \in [K]$  and  $\max_{k \in [K], a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{*, -1}(a)}\|_2^2 \asymp p/r$ .

Next, we need to find the lower bound of the rightmost side in (31). We consider the hypothesis test based on model (26). First, we reparameterize the model under the construction (i)-(ii).

$$\mathbf{x}_a^* = [\text{Mat}_1(\mathcal{S}^* \times_2 \boldsymbol{\Theta}_2^* \mathbf{M}_2^* \times_3 \cdots \times_K \boldsymbol{\Theta}_K^* \mathbf{M}_K^*)]_{a:}, \quad (1717)$$

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 (31) for  $i = 1$ . The analysis for other  $i \in T_1^c$  are similar. For notational simplicity, we suppress the subscript  $i$  and write  $\mathbf{y}, \boldsymbol{\theta}^*, z$  in place of  $\mathbf{y}_1, \boldsymbol{\theta}_1^*(1)$  and  $z_1(1)$ , respectively. The equivalent vector problem for assessing the summand in (31) is

$$\mathbf{y} = \boldsymbol{\theta}^* \mathbf{x}_z^* + \mathbf{e}, \quad (32)$$

where  $z \in \{1, 2\}$  is an unknown parameter,  $\boldsymbol{\theta}^* \in \mathbb{R}_+$  is the given heterogeneity degree,  $\mathbf{x}_1^*, \mathbf{x}_2^* \in \mathbb{R}^{p^{K-1}}$  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 (32):

$$H_0 : z = 1, \mathbf{y} = \boldsymbol{\theta}^* \mathbf{x}_1^* + \mathbf{e} \leftrightarrow H_1 : z = 2, \mathbf{y} = \boldsymbol{\theta}^* \mathbf{x}_2^* + \mathbf{e}, \quad (33)$$

The hypothesis testing (33) is a simple versus simple testing, since the assignment  $z$  is the only unknown parameter in the test. By Neyman-Pearson lemma, the likelihood ratio test is optimal with minimal Type I + II error. Under Gaussian model, the likelihood ratio test of (33) is equivalent to the least square estimator  $\hat{z}_{LS} = \arg \min_{a=\{1,2\}} \|\mathbf{y} - \boldsymbol{\theta}^* \mathbf{x}_a^*\|_F^2$ .

Let  $\mathbf{S} = \text{Mat}_1(\mathcal{S})$ . Note that

$$\begin{aligned} &\|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F \\ &\leq \boldsymbol{\theta}^* \|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \prod_{k=2}^K \lambda_{\max}(\boldsymbol{\Theta}_k^* \mathbf{M}_k^*) \\ &\leq \boldsymbol{\theta}^* \|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \max_{k \in [K]/\{1\}, a \in [r]} \|\boldsymbol{\theta}_{k,z_k^{*, -1}(a)}\|_2^{K-1} \\ &\leq \sigma r^{(K-1)/2} p^{-(K-1)/2} 2 c_4 p^{(K-1)/2} r^{-(K-1)/2} \\ &\leq 2 c_4 \sigma, \end{aligned}$$

where  $\lambda_{\max}(\cdot)$  denotes the maximal singular value, the second inequality follows from Lemma 6, and the third inequality

follows from property (ii) and the boundedness constraint in  $\mathcal{P}_S(\gamma)$  such that  $\|\mathbf{S}_{1:}^* - \mathbf{S}_{2:}^*\|_F \leq \|\mathbf{S}_{1:}^*\|_F + \|\mathbf{S}_{2:}^*\|_F \leq 2c_4$ .  
Hence, we have

$$\begin{aligned} & \inf_{\hat{z}_1(1)} \{\mathbb{P}[\hat{z}_1(1) = 1 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ & \quad + \mathbb{P}[\hat{z}_1(1) = 2 | z_1^*(1) = 1, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*]\} \\ &= 2\mathbb{P}[\hat{z}_{LS} = 1 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ &= 2\mathbb{P}[\|\mathbf{y} - \boldsymbol{\theta}^* \mathbf{x}_1^*\|_F^2 \leq \|\mathbf{y} - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F^2 | z_1^*(1) = 2, z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \\ &= 2\mathbb{P}[2\langle \mathbf{e}, \boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^* \rangle \geq \|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F^2] \\ &= 2\mathbb{P}[N(0, 1) \geq \boldsymbol{\theta}^* \|\mathbf{x}_1^* - \mathbf{x}_2^*\|_F / (2\sigma)] \\ &\geq 2\mathbb{P}[N(0, 1) \geq c_4] \geq c, \end{aligned} \quad (34)$$

where the first equation holds by symmetry, the third equation holds by rearrangement, the fourth equation holds from the fact that  $\langle \mathbf{e}, \boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^* \rangle \sim N(0, \sigma \|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F)$ , and  $c$  is some positive constant in the last inequality.

Plugging the inequality (34) into the inequality (31) for all  $i \in T_1^c$ , then, we have

$$\liminf_{p \rightarrow \infty} \inf_{\hat{z}_1} \mathbb{E}[\rho(\hat{z}_1, z_1^*) | z_k^*, \boldsymbol{\theta}_k^*, \mathcal{S}^*] \geq \liminf_{p \rightarrow \infty} \frac{Ccp}{r^3} \geq Cc,$$

where the last inequality follows by the condition  $r = o(p^{1/3})$ . By the discrete nature of the misclustering error, we obtain our conclusion

$$\liminf_{p \rightarrow \infty} \inf_{\mathcal{S}^* \in \mathcal{P}_S(\gamma)} \inf_{\hat{z}_{\text{stat}}} \sup_{(z^*, \boldsymbol{\theta}^*) \in \mathcal{P}_{z, \boldsymbol{\theta}}} \mathbb{E}[\rho(\hat{z}_{\text{stat}}, z)] \geq 1.$$

Last, with constructed  $z_k^*, \boldsymbol{\theta}_k^*$  satisfying properties (i) and (ii) and  $\gamma' < -(K-1)$ , we construct a core tensor  $\mathcal{S}^*$  such that  $\Delta_{\mathbf{X}^*}^2 \leq p^{-(K-1)}$ . Based on the property (ii) and the boundedness constraint of  $\mathcal{S}^*$  in  $\mathcal{P}$ , we still have  $\|\boldsymbol{\theta}^* \mathbf{x}_1^* - \boldsymbol{\theta}^* \mathbf{x}_2^*\|_F \leq 2c_4\sigma$ . Hence, we obtain the desired result

$$\begin{aligned} & \liminf_{p \rightarrow \infty} \inf_{\hat{z}_1} \sup_{(z, \mathcal{S}, \boldsymbol{\theta}) \in \mathcal{P}'(\gamma')} \mathbb{E}[\rho(\hat{z}_1, z_1)] \\ & \geq \liminf_{p \rightarrow \infty} \inf_{\hat{z}_{\text{stat}}} \mathbb{E}[\rho(\hat{z}_1, z_1^*) | z_k^*, \mathcal{S}^*, \boldsymbol{\theta}_k^*] \geq 1. \end{aligned}$$

□

### 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  $K \geq 2$  and  $r \geq 2$ . 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

$$\mathcal{S} = \mathcal{S}(\gamma) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}^{\otimes K}. \quad (35)$$

Then, we consider the signal tensor family

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

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 (5). 1789
- (ii) For every  $\gamma \in \mathbb{R}$ ,  $\{\mathcal{X} - \mathbf{1}: \mathcal{X} \in \mathcal{P}_{\text{shifted}}(\gamma)\} \subset \mathcal{P}_{\text{nondegree}}(\gamma)$ , 1790 where  $\mathcal{P}_{\text{nondegree}}(\gamma)$  denotes the subfamily of rank-one tensor 1791 block model constructed in the proof of [13, Theorem 7]. 1792

The verification of the above two properties is provided in the end 1794 of this proof. 1795

Now, following the proof of [13, Theorem 7], when  $\gamma < -K/2$ , 1796 every polynomial-time algorithm estimator  $(\hat{\mathbf{M}}_k)_{k \in [K]}$  obeys 1797

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{nondegree}}(\gamma)} \mathbb{P}(\exists k \in [K], \hat{\mathbf{M}}_k \neq \mathbf{M}_k) \geq 1/2, \quad (36) \quad 1798$$

under the HPC Conjecture 1. The inequality (36) implies 1799

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}_{\text{nondegree}}(\gamma)} \max_{k \in [K]} \mathbb{E}[\rho(\mathbf{z}_k, \hat{\mathbf{z}}_k)] \geq 1. \quad 1800$$

Based on properties (i)-(ii), we conclude that 1801

$$\liminf_{p \rightarrow \infty} \sup_{\mathcal{X} \in \mathcal{P}(\gamma)} \max_{k \in [K]} \mathbb{E}[\rho(\mathbf{z}_k, \hat{\mathbf{z}}_k)] \geq 1. \quad 1802$$

We complete the proof by verifying the properties (i)-(ii). For (i), 1803 we verify that the angle gap for the core tensor  $\mathcal{S}$  in (35) is on the 1804 order of  $\sigma p^{-\gamma/2}$ . Specifically, write  $\mathbf{1} = (1, 1, 1)$  and  $\mathbf{e} = (1, -1, 0)$ . 1805 We have 1806

$$\text{Mat}(\mathcal{S}) = \begin{bmatrix} \text{Vec}(\mathbf{1}^{\otimes K-1}) + \sigma p^{-\gamma/2} \text{Vec}(\mathbf{e}^{\otimes(K-1)}) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) - \sigma p^{-\gamma/2} \text{Vec}(\mathbf{e}^{\otimes(K-1)}) \\ \text{Vec}(\mathbf{1}^{\otimes K-1}) \end{bmatrix}. \quad 1807$$

Based on the orthogonality  $\langle \mathbf{1}, \mathbf{e} \rangle = 0$ , the minimal angle gap among 1808 rows of  $\text{Mat}(\mathcal{S})$  is 1809

$$\begin{aligned} \Delta_{\min}^2(\mathcal{S}) &\asymp \tan^2(\text{Mat}(\mathcal{S})_{1:}, \text{Mat}(\mathcal{S})_{3:}) \\ &= \left( \frac{\|\mathbf{e}\|_2}{\|\mathbf{1}\|_2} \right)^{2(K-1)} \sigma^2 d^{-\gamma} \\ &\asymp \sigma^2 d^{-\gamma}. \end{aligned} \quad 1810 \quad 1811 \quad 1812$$

Therefore, we have shown that  $\mathcal{P}_{\text{shifted}}(\gamma) = \mathcal{P}(\gamma)$ . Finally, the 1813 property (ii) follows directly by comparing the definition of  $\mathcal{S}$  in (35) 1814 with that in the proof of [13, Theorem 7]. 1815 □

### F. Proof of Theorem 4 and Proposition 1

*Proof of Theorem 4:* We prove Theorem 4 under the dTBM (1) 1817 with symmetric mean tensor, parameters  $(z, \mathcal{S}, \boldsymbol{\theta})$ , fixed  $r \geq 1, K \geq 1818 2$ , and i.i.d. noise. For the case  $r = 1$ , we have  $L(z^{(0)}, z) = 1819 0, \ell(z^{(0)}, z) = 0$  trivially. Hence, we focus on the proof of the first 1820 mode clustering  $z_1^{(0)}$  with  $r \geq 2$ ; the proofs for the other modes can 1821 be extended similarly. We drop the subscript  $k$  in the matricizations 1822  $\mathbf{M}_k, \mathbf{X}_k, \mathbf{S}_k$  and in the estimate  $z_1^{(0)}$ . We firstly show the proof with 1823 balanced  $\boldsymbol{\theta}$ . 1824

We firstly show the upper bound for misclustering error  $\ell(z^{(0)}, z)$ . 1825 First, by Lemma 1, there exists a positive constant such that 1826  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\min}$ . By the balance assumption 1827 on  $\boldsymbol{\theta}$  and Lemma 8, we have 1828

$$\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_I} \theta(i)^2 + 4 \sum_{i \in S} \theta(i)^2, \quad (37) \quad 1829$$

where 1830

$$S_0 = \{i: \|\hat{\mathbf{X}}_{i:}\| = 0\}, S = \{i \in S_0^c: \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\| \geq c_0 \Delta_{\min}/2\}. \quad 1831$$

1832 On one hand, note that for any set  $P \in [p]$ ,

$$\begin{aligned} 1833 \quad \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 &= \sum_{i \in P} \|\theta(i) \mathbf{S}_{z(i)} : (\Theta \mathbf{M})^{T, \otimes(K-1)}\|^2 \\ 1834 \quad &\geq \sum_{i \in P} \theta(i)^2 \min_{a \in [r]} \|\mathbf{S}_{a:}\|^2 \lambda_r^{2(K-1)}(\Theta \mathbf{M}) \\ 1835 \quad &\gtrsim \sum_{i \in P} \theta(i)^2 p^{K-1} r^{-(K-1)}, \end{aligned}$$

1836 where the last inequality follows Lemma 6, the assumption that  
1837  $\min_{i \in [p]} \theta(i) \geq c$ , and the constraint  $\min_{a \in [r]} \|\mathbf{S}_{a:}\| \geq c_3$  in  
1838 the parameter space (2). Thus, we have

$$1839 \quad \sum_{i \in P} \theta(i)^2 \lesssim \sum_{i \in P} \|\mathbf{X}_{i:}\|^2 p^{-(K-1)} r^{K-1}. \quad (38)$$

1840 On the other hand, note that

$$\begin{aligned} 1841 \quad &\sum_{i \in S} \|\mathbf{X}_{i:}\|^2 \\ 1842 \quad &\leq 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 + 2 \sum_{i \in S} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \\ 1843 \quad &\leq \frac{8}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1844 \quad &\quad (40) \end{aligned}$$

$$\begin{aligned} 1845 \quad &\leq \frac{16}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \left[ \|\hat{\mathbf{x}}_{z^{(0)}(i)} - \hat{\mathbf{X}}_{i:}^s\|^2 + \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 \right] \\ 1846 \quad &\quad + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1847 \quad &\leq \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} \sum_{i \in S} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_{i:}^s - \mathbf{X}_{i:}^s\|^2 + 2 \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (42) \end{aligned}$$

$$1848 \quad \leq \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \quad (43)$$

$$1849 \quad \lesssim \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (44)$$

1850 where inequalities (39) and (41) follow from the triangle  
1851 inequality, (40) follows from the definition of  $S$ , (42) follows  
1852 from the update rule of  $k$ -means in Step 6 of Sub-algorithm 1,  
1853 (43) follows from Lemma 4, and the last inequality (44)  
1854 follows from Lemma 7. Also, note that

$$\begin{aligned} 1855 \quad \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 &= \sum_{i \in S_0} \|\hat{\mathbf{X}}_{i:} - \mathbf{X}_{i:}\|^2 \\ 1856 \quad &\leq \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1857 \quad &\lesssim (p^{K/2} r + pr^2 + r^K) \sigma^2, \quad (45) \end{aligned}$$

1858 where the equation follows from the definition of  $S_0$ . There-  
1859 fore, combining the inequalities (37), (38), (44), and (45),  
1860 we have

$$\begin{aligned} 1861 \quad &\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ 1862 \quad &\lesssim \left( \sum_{i \in S} \|\mathbf{X}_{i:}\|^2 + \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \right) p^{-(K-1)} r^{K-1} \\ 1863 \quad &\lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}} (p^{K/2} r + pr^2 + r^K). \quad (46) \end{aligned}$$

With the assumption that  $\min_{i \in [p]} \theta(i) \geq c$ , we finally obtain  
1864 the result  
1865

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \quad (46)$$

where the last inequality follows from the definition  $\text{SNR} = \Delta_{\min}^2 / \sigma^2$ .  
1867  
1868

Without the balanced  $\theta$ , we have  
1869  $\min_{z(i) \neq z(j)} \|\mathbf{X}_{i:}^s - \mathbf{X}_{j:}^s\| \geq c_0 \Delta_{\mathbf{X}}$ . Replacing the definition  
1870 of  $S$  with  $\Delta_{\mathbf{X}}$ , we obtain the desired result.  
1871

Next, we Show the Bound for  $L(z^{(0)}, z)$ . Note that  $\mathbf{X}_{i:}^s$  have  
1872 only  $r$  different values. We let  $\mathbf{X}_a^s = \mathbf{X}_{i:}^s$  for all  $i$  such that  
1873  $z(i) = a, a \in [r]$ .  
1874

Notice that

$$\|\mathbf{X}_{i:}\|^2 \gtrsim p^{K-1} r^{-(K-1)} \quad (47)$$

and

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

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

$$\begin{aligned} 1879 \quad &\sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1880 \quad &\lesssim \sum_{i \in [p]} \left( \|\mathbf{X}_{i:}\|^2 - \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 \right) \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1881 \quad &\lesssim \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1882 \quad &\lesssim \eta \sum_{i \in [p]} \|\hat{\mathbf{X}}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \mathbf{X}_{i:}^s\|^2 \\ 1883 \quad &\lesssim \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \\ 1884 \quad &\lesssim p^{K/2} r + pr^2 + r^K. \quad (47) \end{aligned}$$

Hence, we have

$$\begin{aligned} 1887 \quad \sum_{i \in [p]} \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1888 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} \sum_{i \in [p]} \|\mathbf{X}_{i:}\|^2 \|\hat{\mathbf{X}}_i^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1889 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} (p^{K/2} r + pr^2 + r^K), \quad (48) \end{aligned}$$

where the first inequality follows from the assumption  
1890  $\min_{i \in [p]} \theta(i) \geq c > 0$ , the second inequality follows from  
1891 the inequality (38), and the last inequality comes from the  
1892 inequality (47).  
1893

Next, we consider the following quantity,

$$\begin{aligned} 1896 \quad &\sum_{i \in [p]} \theta(i) \|\mathbf{X}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1897 \quad &\lesssim \sum_{i \in [p]} \theta(i)^2 \|\mathbf{X}_{i:}^s - \hat{\mathbf{X}}_{i:}^s\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1898 \quad &\lesssim \sum_{i \in [p]} \frac{\theta(i)^2}{\|\mathbf{X}_{i:}\|^2} \|\mathbf{X}_{i:} - \hat{\mathbf{X}}_{i:}\|^2 + \sum_{i \in [p]} \theta(i)^2 \|\hat{\mathbf{X}}_{i:}^s - \hat{\mathbf{x}}_{z^{(0)}(i)}\|^2 \\ 1899 \quad &\lesssim \frac{r^{K-1}}{p^{K-1}} (p^{K/2} r + pr^2 + r^K), \quad (49) \end{aligned}$$

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 (48). In addition, with Theorem 4 and the condition  $\text{SNR} \gtrsim p^{-K/2} \log p$ , 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

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

where the last inequality follows from the inequality (48).

Finally, we obtain

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

where the first inequality follows from Lemma 1, the third inequality follows from inequalities (49) and (50), and the last inequality follows from the assumption that  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$ .  $\square$

*Proof of Proposition 1:* Algorithm 2 shares the same algorithm strategy as Sub-algorithm 1 but with a different estimation of the mean tensor,  $\hat{\mathcal{X}}'$ . Hence, the proof of Proposition 1 follows the same proof idea with the proof of Theorem 4. Replacing the estimation  $\hat{\mathcal{X}}$  by  $\hat{\mathcal{X}}'$  in the proof of Theorem 4, we have

$$\begin{aligned} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \\ \lesssim \left( \sum_{i \in S} \|\mathbf{X}_{i:}\|^2 + \sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \right) p^{-(K-1)} r^{K-1}. \end{aligned} \quad (51)$$

By inequalities (43) and (45), we have

$$\sum_{i \in S} \|\mathbf{X}_{i:}\|^2 \leq \left( \frac{16(1+\eta)}{c_0^2 \Delta_{\min}^2} + 2 \right) \|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2, \quad (52)$$

$$\sum_{i \in S_0} \|\mathbf{X}_{i:}\|^2 \leq \|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2. \quad (53)$$

Hence, it suffices to find the upper bound of the estimation error  $\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2$  to complete our proof. Note that the matricization  $\text{Mat}_{sq}(\mathcal{X}) \in \mathbb{R}^{p^{\lceil K/2 \rceil} \times p^{\lceil K/2 \rceil}}$  has  $\text{rank}(\text{Mat}_{sq}(\mathcal{X})) \leq r^{\lceil K/2 \rceil}$ , and Bernoulli random variables follow the sub-Gaussian distribution with bounded variance  $\sigma^2 = 1/4$ . Apply Lemma 9 to  $\mathbf{Y} = \text{Mat}_{sq}(\mathcal{Y})$ ,  $\mathbf{X} = \text{Mat}_{sq}(\mathcal{X})$ , and  $\hat{\mathbf{X}} = \text{Mat}_{sq}(\hat{\mathcal{X}}')$ . Then, with probability tending to 1 as  $p \rightarrow \infty$ , we have

$$\|\hat{\mathcal{X}}' - \mathcal{X}\|_F^2 = \|\text{Mat}_{sq}(\hat{\mathcal{X}}') - \text{Mat}_{sq}(\mathcal{X})\|_F^2 \lesssim p^{\lceil K/2 \rceil}. \quad (54)$$

Combining the estimation error (54) with inequalities (52), (53), and (51), we obtain

$$\min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}} p^{\lceil K/2 \rceil}. \quad (55)$$

Replace the inequality (46) in the proof of Theorem 4 by inequality (55). With the the same procedures to obtain  $\ell(\hat{z}^{(0)}, z)$  and  $L(\hat{z}^{(0)}, z)$  for Theorem 4, we finish the proof of Proposition 1.  $\square$

#### 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

$$\sin(\mathbf{v}_1, \mathbf{v}_2) \leq \|\mathbf{v}_1^s - \mathbf{v}_2^s\| \leq \frac{2 \|\mathbf{v}_1 - \mathbf{v}_2\|}{\max(\|\mathbf{v}_1\|, \|\mathbf{v}_2\|)}. \quad (56)$$

*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

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = \sqrt{2(1 - \cos \alpha)} = 2 \sin \frac{\alpha}{2} \geq \sin \alpha, \quad (57)$$

where the equations follow 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

$$\text{Pad}_{\mathbf{w}}(\mathbf{a}) = [a_1 \circ \mathbf{w}_1, \dots, a_d \circ \mathbf{w}_d]^T, \quad (58)$$

where  $a_i \circ \mathbf{w}_i = [a_i w_{i1}, \dots, a_i w_{ip_i}]^T$ , for all  $i \in [d]$ . 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

$$\min_{i \in [d]} \|w_i\|^2 \|\mathbf{a}\|^2 \leq \|\text{Pad}_{\mathbf{w}}(\mathbf{a})\|^2 \leq \max_{i \in [d]} \|w_i\|^2 \|\mathbf{a}\|^2. \quad (59)$$

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

$$\text{Pad}_{\mathbf{w}}^{-1}(\text{Pad}_{\mathbf{w}}(\mathbf{a})) = \mathbf{a}. \quad (60)$$

**Lemma 5 (Angle for Weighted Padding Vectors):** Suppose that we have two nonzero vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ . Given the weight collection  $\mathbf{w}$ , we have

$$\begin{aligned} \frac{\min_{i \in [d]} \|\mathbf{w}_i\|}{\max_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}) &\stackrel{*}{\leq} \sin(\text{Pad}_{\mathbf{w}}(\mathbf{a}), \text{Pad}_{\mathbf{w}}(\mathbf{b})) \\ &\stackrel{**}{\leq} \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}). \quad (58) \end{aligned}$$

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

First, we prove the inequality  $**$  in (58). 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 (57) to vectors  $\mathbf{b}$  and  $\mathbf{a}^\perp$ .

Next, we prove inequality  $*$  in (58). With the decomposition of  $\text{Pad}_{\mathbf{w}}(\mathbf{b})$  and the inverse weighted padding operator, we have

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

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 (57) 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 (2) and assumption that  $\min_{i \in [p]} \theta(i) \geq c$  for some constant  $c > 0$ , the singular values of  $\Theta M$  are bounded as

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

**Proof of Lemma 6:** Note that

$$(\Theta M)^T \Theta M = D,$$

with  $D = \text{diag}(D_1, \dots, D_r)$  where  $D_a = \|\theta_{z^{-1}(a)}\|^2, a \in [r]$ . By the definition of singular values, we have

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

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

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

where the last inequality follows from the constraint in parameter space (2). Finally, notice that

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

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

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

where  $C, c$  are some positive constants.

**Proof of Lemma 7:** See [13, 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  $a \in [r]$  with  $r \geq 2$ . 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 (6) such that  $\frac{\max_{a \in [r]} \|\theta_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\theta_{z^{-1}(a)}\|^2} = 1 + o(1)$ . Consider an arbitrary estimate  $\hat{z}$  with  $\hat{\mathbf{x}}_i = \hat{\mathbf{v}}_{\hat{z}(i)}$  for all  $i \in S$ . Then, if

$$\min_{a \neq b \in [r]} \|\mathbf{v}_a - \mathbf{v}_b\| \geq 2c, \quad (59)$$

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 4 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

$$R_1 = \{u \in [r] : C_u = \emptyset\},$$

$$R_2 = \{u \in [r] : C_u \neq \emptyset, \text{ for all } i, j \in C_u, \hat{z}(i) = \hat{z}(j)\},$$

$$R_3 = \{u \in [r] : C_u \neq \emptyset, \text{ there exist } i, j \in C_u, \hat{z}(i) \neq \hat{z}(j)\}.$$

2056 Note that  $\cup_{u \in [r]} C_u = S_0^c / S^c$  and  $C_u \cap C_v = \emptyset$  for any  $u \neq v$ .  
 2057 Suppose there exist  $i \in C_u$  and  $j \in C_v$  with  $u \neq v \in [r]$  and  
 2058  $\hat{z}(i) = \hat{z}(j)$ . Then we have

2059  $\|\mathbf{v}_{z(i)} - \mathbf{v}_{z(j)}\| \leq \|\mathbf{v}_{z(i)} - \hat{\mathbf{x}}_i\| + \|\mathbf{v}_{z(j)} - \hat{\mathbf{x}}_j\| < 2c,$

2060 which contradicts to the assumption (59). Hence, the estimates  
 2061  $\hat{z}(i) \neq \hat{z}(j)$  for the nodes  $i \in C_u$  and  $j \in C_v$  with  $u \neq v$ .  
 2062 By the definition of  $R_2$ , the nodes in  $\cup_{u \in R_2} C_u$  have the same  
 2063 assignment with  $z$  and  $\hat{z}$ . Then, we have

2064  $\min_{\pi \in \Pi} \sum_{i: \hat{z}(i) \neq \pi(z(i))} \theta(i)^2 \leq \sum_{i \in S_0} \theta(i)^2 + \sum_{i \in S} \theta(i)^2 + \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2.$

2065 We only need to bound  $\sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2$  to finish the proof.  
 2066 Note that every  $C_u$  with  $u \in R_3$  contains at least two  
 2067 nodes assigned to different clusters by  $\hat{z}$ . Then, we have  
 2068  $|R_2| + 2|R_3| \leq r$ . Since  $|R_1| + |R_2| + |R_3| = r$ , we have  
 2069  $|R_3| \leq |R_1|$ . Hence, we obtain

2070 
$$\begin{aligned} \sum_{i \in \cup_{u \in R_3} C_u} \theta(i)^2 &\leq |R_3| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ 2071 &\leq |R_1| \max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2 \\ 2072 &\leq \frac{\max_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2}{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{-1}(a)}\|^2} \sum_{i \in \cup_{u \in R_1} z^{-1}(u)} \theta(i)^2 \\ 2073 &\leq 2 \sum_{i \in S} \theta(i)^2, \end{aligned}$$

2074 where the last inequality holds by the balanced assumption on  
 2075  $\boldsymbol{\theta}$  when  $p$  is large enough, and the fact that  $\cup_{u \in R_1} z^{-1}(u) \subset S$ .  
 2076  $\square$

2077 *Lemma 9 (Low-Rank Matrix Estimation):* Let  $\mathbf{Y} = \mathbf{X} +$   
 2078  $\mathbf{E} \in \mathbb{R}^{m \times n}$ , where  $n > m$  and  $\mathbf{E}$  contains independent mean-  
 2079 zero sub-Gaussian entries with bounded variance  $\sigma^2$ . Suppose  
 2080  $\text{rank}(\mathbf{X}) = r$ . Consider the least square estimator

2081 
$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}' \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{X}') \leq r} \|\mathbf{X}' - \mathbf{Y}\|_F^2.$$

2082 There exist positive constants  $C_1, C_2$  such that

2083 
$$\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 \leq C_1 \sigma^2 nr,$$

2084 with probability at least  $1 - \exp(-C_2 nr)$ .

2085 *Proof of Lemma 9:* Note that  $\|\hat{\mathbf{X}} - \mathbf{Y}\|_F^2 \leq \|\mathbf{X} - \mathbf{Y}\|_F^2$  by  
 2086 the definition of least square estimator.

2087 We have

2088 
$$\begin{aligned} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 \\ 2089 &\leq 2 \langle \hat{\mathbf{X}} - \mathbf{X}, \mathbf{Y} - \mathbf{X} \rangle \\ 2090 &\leq 2 \|\hat{\mathbf{X}} - \mathbf{X}\|_F \sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{T}, \mathbf{Y} - \mathbf{X} \rangle \quad (60) \end{aligned}$$

2091 with probability at least  $1 - \exp(-C_2 nr)$ , where the second  
 2092 inequality follows by re-arrangement.

2093 Consider the SVD for matrix  $\mathbf{T} = \mathbf{U} \Sigma \mathbf{V}^T$  with orthogonal  
 2094 matrices  $\mathbf{U} \in \mathbb{R}^{m \times 2r}, \mathbf{V} \in \mathbb{R}^{n \times 2r}$  and diagonal matrix  $\Sigma \in$

$\mathbb{R}^{2r \times 2r}$ . We have

2095 
$$\begin{aligned} &\sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{T}, \mathbf{Y} - \mathbf{X} \rangle \\ 2096 &= \sup_{\mathbf{T} \in \mathbb{R}^{m \times n}, \text{rank}(\mathbf{T}) \leq 2r, \|\mathbf{T}\|_F=1} \langle \mathbf{U} \Sigma, \mathbf{EV} \rangle \\ 2097 &= \sup_{\mathbf{v} \in \mathbb{R}^{2nr}} \mathbf{v}^T \mathbf{e} \leq C \sigma \sqrt{nr}, \end{aligned} \quad (61) \quad 2098$$

2099 with probability  $1 - \exp(-C_2 nr)$ , where  $C, C_2$  are two  
 2100 positive constants, the vectorization  $\mathbf{e} = \text{Vec}(\mathbf{EV}) \in \mathbb{R}^{2nr}$   
 2101 has independent mean-zero sub-Gaussian entries with bounded  
 2102 variance  $\sigma^2$  due to the orthogonality of  $\mathbf{V}$ , and the last  
 2103 inequality follows from [36, Theorem 1.19].  
 2104 Combining inequalities (60) and (61), we obtain the desired  
 2105 conclusion.  $\square$

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

2106 *Proof of Theorem 2 (Achievability) and Theorem 5:* The  
 2107 proofs of Theorem 2 (Achievability) and Theorem 5 share the  
 2108 same idea. We prove the contraction step by step. In each  
 2109 step, we show the specific procedures for the algorithm loss  
 2110 and address the MLE loss by stating the difference.

2111 We consider dTBM (1) with symmetric mean tensor, param-  
 2112 eters  $(z, \mathcal{S}, \boldsymbol{\theta})$ , fixed  $r \geq 1, K \geq 2$ , and i.i.d. noise. Let  
 2113  $(\hat{z}, \hat{\mathcal{S}}, \hat{\boldsymbol{\theta}})$  denote the MLE in (9), and  $(z_k^{(0)}, \mathcal{S}^{(0)}, \boldsymbol{\theta}_k^{(0)})$  denote  
 2114 parameters related to the initialization. For the case  $r = 1$ ,  
 2115  $\ell(z_k^{(t)}, z) = 0$  trivially for all  $t \geq 0, k \in [k]$ . Hence, we focus  
 2116 on the proof of the first mode clustering  $z_1^{(t+1)}$  with  $r \geq 2$ ; the  
 2117 extension for other modes can be obtained similarly. We drop  
 2118 the subscript  $k$  in the matricizations  $\boldsymbol{\Theta}, \mathbf{M}_k, \mathbf{S}_k, \mathbf{X}_k$  and in  
 2119 estimates  $z_k^{(0)}, z_k^{(t+1)}, z_k^{(t)}$  for ease of the notation. Without  
 2120 loss of generality, we assume that the variance  $\sigma = 1$ , and that  
 2121 the identity permutation minimizes the initial misclustering  
 2122 error; i.e.,  $\pi^{(0)} = \arg \min_{\pi \in \Pi} \sum_{i \in [p]} \mathbb{1} \{z^{(0)}(i) \neq \pi \circ z(i)\}$   
 2123 and  $\pi^{(0)}(a) = a$  for all  $a \in [r]$ , and so for  $\hat{z}$ .

2124 *Step 1 (Notation and Conditions):* We first introduce addi-  
 2125 tional notations and the necessary conditions used in the proof.  
 2126 We will verify that the conditions hold in our context under  
 2127 high probability in the last step of the proof.

#### Notation:

2128 1) Projection. We use  $\mathbf{I}_d$  to denote the identity matrix of  
 2129 dimension  $d$ . For a vector  $\mathbf{v} \in \mathbb{R}^d$ , let  $\text{Proj}(\mathbf{v}) \in \mathbb{R}^{d \times d}$  denote  
 2130 the projection matrix to  $\mathbf{v}$ . Then,  $\mathbf{I}_d - \text{Proj}(\mathbf{v})$  is the projection  
 2131 matrix to the orthogonal complement  $\mathbf{v}^\perp$ .

2132 2) We define normalized membership matrices

2133 
$$\mathbf{W} = \mathbf{M} \left( \text{diag}(\mathbf{1}_p^T \mathbf{M}) \right)^{-1}, \quad \mathbf{W}^{(t)} = \mathbf{M}^{(t)} \left( \text{diag}(\mathbf{1}_p^T \mathbf{M}^{(t)}) \right)^{-1},$$

2134 weighted normalized membership matrices

2135 
$$\mathbf{P} = \boldsymbol{\Theta} \mathbf{M} (\text{diag}(\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2, \dots, \|\boldsymbol{\theta}_{z^{-1}(r)}\|^2))^{-1},$$

2136 
$$\hat{\mathbf{P}} = \hat{\boldsymbol{\Theta}} \hat{\mathbf{M}} (\text{diag}(\|\hat{\boldsymbol{\theta}}_{z^{-1}(1)}\|^2, \dots, \|\hat{\boldsymbol{\theta}}_{z^{-1}(r)}\|^2))^{-1},$$

2137 and the dual normalized and dual weighted normalized mem-  
 2138 bership matrices

2139 
$$\mathbf{V} = \mathbf{W}^{\otimes(K-1)}, \quad \mathbf{V}^{(t)} = \left( \mathbf{W}^{(t)} \right)^{\otimes(K-1)},$$

2140 
$$\hat{\mathbf{Q}} = \mathbf{P}^{\otimes K-1}, \quad \hat{\mathbf{Q}} = \hat{\mathbf{P}}^{\otimes K-1}.$$

2141 Also, let  $\mathbf{B} = (\boldsymbol{\Theta} \mathbf{M})^{\otimes(K-1)}, \hat{\mathbf{B}} = (\hat{\boldsymbol{\Theta}} \hat{\mathbf{M}})^{\otimes(K-1)}$ . By the  
 2142 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.,

$$\begin{aligned}\mathcal{S}^{(t)} &= \mathcal{Y} \times_1 (\mathbf{W}^{(t)})^T \times_2 \cdots \times_K (\mathbf{W}^{(t)})^T, \\ \tilde{\mathcal{S}} &= \mathcal{Y} \times_1 \mathbf{W}^T \times_2 \cdots \times_K \mathbf{W}^T, \\ \hat{\mathcal{S}} &= \mathcal{Y} \times_1 \hat{\mathbf{P}}^T \times_2 \cdots \times_K \hat{\mathbf{P}}^T, \\ \bar{\mathcal{S}} &= \mathcal{Y} \times_1 \mathbf{P}^T \times_2 \cdots \times_K \mathbf{P}^T.\end{aligned}$$

4) We define the matricizations of tensors

$$\begin{aligned}\mathbf{S} &= \text{Mat}(\mathcal{S}), \quad \mathbf{Y} = \text{Mat}(\mathcal{Y}), \quad \mathbf{X} = \text{Mat}(\mathcal{X}), \quad \mathbf{E} = \text{Mat}(\mathcal{E}), \\ \mathbf{S}^{(t)} &= \text{Mat}(\mathcal{S}^{(t)}), \quad \hat{\mathcal{S}} = \text{Mat}(\hat{\mathcal{S}}), \quad \tilde{\mathcal{S}} = \text{Mat}(\tilde{\mathcal{S}}), \quad \bar{\mathcal{S}} = \text{Mat}(\bar{\mathcal{S}}).\end{aligned}$$

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 (2), 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

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

We also define the loss for oracle and weighted oracle estimators

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

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

Then we introduce the necessary conditions in Condition 1.

*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

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

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

$$\|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_b]_s\|^2 \leq \|[\mathbf{Y}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{S}_{z(i)}]_s\|^2. \quad (67)$$

Note that the event (67) 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 (67) yields the decomposition

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

where

$$\begin{aligned}F_{ib}^{(t)} &= 2 \langle \mathbf{E}_{i:} \mathbf{V}^{(t)}, ([\tilde{\mathcal{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s) - ([\tilde{\mathcal{S}}_b]_s - [\mathbf{S}_b]_s) \rangle \\ &\quad + 2 \langle \mathbf{E}_{i:} (\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathcal{S}}_{z(i)}]_s - [\tilde{\mathcal{S}}_b]_s \rangle,\end{aligned}$$

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

$$\begin{aligned}H_{ib}^{(t)} &= \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]_s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]_s \\ &\quad - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 + \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2.\end{aligned}$$

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

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

Note that

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

where the first inequality follows from the property of eigenvalues; the last inequality follows from Lemma 6, Lemma 10, 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  $\|\mathbf{X}_{i:} \mathbf{V}^{(t)}\|$  (72) into the inequality (71) gives

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

where

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

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

Taking the weighted summation of (73) 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

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} A_{ib} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2. \quad (74)$$

Similarly to  $\xi$  in (74), 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  $\hat{\theta}$  satisfies  $\hat{\theta}(i) = \langle \mathbf{Y}_{i:}, \hat{\mathbf{A}}_{\hat{z}(i):} \rangle / \|\hat{\mathbf{A}}_{\hat{z}(i):}\|_F^2$  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

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

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

$$\|\mathbf{EV}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}} \left( p^{1/2} + r^{(K-1)/2} \right), \quad \|\mathbf{EV}\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}}, \quad \|\mathbf{W}_a^T \mathbf{EV}\| \lesssim \frac{r^K}{p^{K/2}}, \text{ for all } a \in [r], \quad (62)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_\sigma \lesssim \left( \sqrt{r^{K-1}} + K\sqrt{pr} \right), \quad (63)$$

$$\sup_{\mathbf{U}_k \in \mathbb{O}_{p,r}, k=2,\dots,K} \|\mathbf{E}(\mathbf{U}_2 \otimes \cdots \otimes \mathbf{U}_K)\|_F \lesssim \left( \sqrt{pr^{K-1}} + K\sqrt{pr} \right), \quad (64)$$

$$\xi \leq \exp \left( -M \frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad \xi' \lesssim \exp \left( -\frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}} \right), \quad (65)$$

$$L^{(t)} \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \quad \text{for } t = 0, 1, \dots, T, \quad L(\hat{z}) \leq \frac{\bar{C}}{\tilde{C}} \frac{\Delta_{\min}^2}{r \log p}, \quad (66)$$

where  $M$  is a positive universal constant in inequality (84),  $\bar{C}, \tilde{C}$  are positive universal constants in the proof of Theorem 4 and assumption  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$ , respectively. Further, inequality (62) 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 (66) holds.

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

$$A_{ib}' = \mathbb{1} \left\{ \hat{z}(i) = b, \langle \mathbf{E}_{i:}, [\tilde{\mathbf{A}}_{z(i):}]^s - [\tilde{\mathbf{A}}_{b:}]^s \rangle \right. \quad (2232)$$

$$\left. \leq -\frac{m'}{4} \sqrt{\frac{p^{K-1}}{r^{K-1}}} \|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2 \right\}, \quad (2233)$$

$$B_{ib}' = \mathbb{1} \left\{ \hat{z}(i) = b, -\frac{1}{2} \|[\mathbf{A}_{z(i):}]^s - [\mathbf{A}_{b:}]^s\|_F^2 \right. \quad (2234)$$

$$\left. \leq \sqrt{\frac{r^{K-1}}{(m')^2 p^{K-1}}} \hat{F}_{ib} + \hat{G}_{ib} + \hat{H}_{ib} \right\} \quad (2235)$$

with terms

$$\hat{F}_{ib} = 2 \left\langle \mathbf{E}_{i:}, ([\tilde{\mathbf{A}}_{z(i):}]^s - [\hat{\mathbf{A}}_{a:}]^s) - ([\tilde{\mathbf{A}}_{b:}]^s - [\hat{\mathbf{A}}_{b:}]^s) \right\rangle, \quad (2237)$$

$$\hat{G}_{ib} = \left( \|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{z(i):}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \right) \quad (2238)$$

$$- \left( \|\mathbf{X}_{i:}^s - [\hat{\mathbf{A}}_{b:}]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \right), \quad (2239)$$

$$\hat{H}_{ib} = \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:z(i)}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 - \|\mathbf{X}_{i:}^s - [\mathbf{P}_{:b}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\|_F^2 \quad (2240)$$

$$+ \|\mathbf{A}_{z(i):}^s - \mathbf{A}_{b:}^s\|_F^2. \quad (2241)$$

*Step 3 (Derivation of Contraction Inequality):* In this step we derive the upper bound of  $\zeta_{ib}$  and obtain the contraction inequality (24). We show the analysis in the following one-column box for a better presentation.

*Step 4 (Verification of Condition 1):* Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (62), (63), and (64) describe the property of the sub-Gaussian noise tensor  $\mathcal{E}$ , and the readers can find the proof directly in [13, Step 5, Proof of Theorem 2]. The initial condition (66) for MLE is satisfied by Lemma 13. Here, we include only the verification of inequalities (65) and (66) for algorithm estimators.

Now, we verify the oracle loss condition (65). Recall the definition of  $\xi$ ,

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \left\{ \langle \mathbf{E}_{i:} \mathbf{V}, [\tilde{\mathbf{S}}_{z(i):}]^s - [\tilde{\mathbf{S}}_{b:}]^s \rangle \right. \quad (2256)$$

$$\leq -\frac{\theta(i)m}{4} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \left\{ \cdot \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right\}$$

where

2267

$$P_1 = \mathbb{P} \left( \langle e_i, [S_{z(i)}]_s^s - [S_b]_s^s \rangle \leq -\frac{\theta(i)m}{8} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right), \quad 2268$$

$$P_2 = \mathbb{P} \left( \langle e_i, [\tilde{S}_{z(i)}]_s^s - [S_{z(i)}]_s^s \rangle \leq -\frac{\theta(i)m}{16} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right), \quad 2269$$

$$P_3 = \mathbb{P} \left( \langle e_i, [S_b]_s^s - [\tilde{S}_b]_s^s \rangle \leq -\frac{\theta(i)m}{16} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right). \quad 2270$$

2258 Let  $e_i = E_{i,:} V$  denote the aggregated noise vector for all  $i \in [p]$ , and  $e_i$ 's are independent zero-mean sub-Gaussian vector  
2259 in  $\mathbb{R}^{r^{K-1}}$ . The entries in  $e_i$  are independent zero-mean sub-  
2260 Gaussian variables with sub-Gaussian norm upper bounded by  
2261  $m_1 \sqrt{r^{K-1}/p^{K-1}}$  with some positive constant  $m_1$ . We have  
2262 the probability inequality  
2263

$$\begin{aligned} 2265 \mathbb{P} \left( \langle e_i, [\tilde{S}_{z(i)}]_s^s - [\tilde{S}_b]_s^s \rangle \leq -\frac{\theta(i)m}{4} \|[S_{z(i)}]_s^s - [S_b]_s^s\|^2 \right) \\ 2266 \leq P_1 + P_2 + P_3, \end{aligned}$$

Step 3: Choose the constant  $\tilde{C}$  in the condition  $\text{SNR} \geq \tilde{C} p^{-K/2} \log p$  that satisfies the condition of Lemma 11, inequalities (98), and (102). Note that

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

where the first inequality follows from the inequality (89) in Lemma 11, and the last inequality follows from the assumption that  $\min_{i \in [p]} \theta(i) \geq c > 0$ . Following [13, Step 4, Proof of Theorem 2] and Lemma 11, we have

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

for a positive universal constant  $C$  and

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

where the last inequality follows from the definition of  $L^{(t)}$  and the constraint of  $\theta$  in parameter space (2). For  $\tilde{C}$  also satisfies

$$\frac{C_0 \bar{C}}{cm^2 \tilde{C}^2} \leq \frac{1}{512}, \quad (75)$$

we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta_{ib}^{(t)} \leq \frac{1}{8} L^{(t+1)} + \frac{1}{4} L^{(t)}. \quad (76)$$

Plugging the inequality (76) into the decomposition (74), we obtain the contraction inequality

$$L^{(t+1)} \leq \frac{3}{2} \xi + \frac{1}{2} L^{(t)}, \quad (77)$$

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

Therefore, with  $\tilde{C}$  satisfying inequalities (75), (98) and (102), we obtain the conclusion in Theorem 5 via inequality (77) combining the inequality (65) in Condition 1 and Lemma 2.

We also have the contraction inequality for MLE.

Following the same derivation of (77) with the upper bound of  $\hat{F}_{ib}, \hat{G}_{ib}, \hat{H}_{ib}$  in Lemma 12, we also have

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

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

constant  $m_2$ . Then, by Chernoff bound, we have

$$P_1 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(j)}]^s - [\mathbf{S}_{b:}]^s\|^2\right). \quad (78)$$

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

$$P_2 \leq P_{21} + P_{22}, \quad (79)$$

where with some positive constant  $t > 0$ ,

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

For  $P_{21}$ , note that the term  $\mathbf{W}_{:z(i)}^T \mathbf{EV} = \frac{\sum_{j \neq i, j \in [p]} \mathbb{1}\{z(j)=z(i)\} e_j}{\sum_{j \in [p]} \mathbb{1}\{z(j)=z(i)\}}$  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

$$\begin{aligned} P_{21} &\leq \mathbb{P}\left(t \|\mathbf{S}_{z(i)}\| \leq \|\tilde{\mathbf{S}}_{z(i)} - \mathbf{S}_{z(i)}\|\right) \\ &\leq \mathbb{P}\left(c_3 t \leq \|\mathbf{W}_{:z(i)}^T \mathbf{EV}\|\right) \\ &\lesssim \exp\left(-\frac{p^K t^2}{r^K}\right), \end{aligned} \quad (80)$$

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 (2), 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

$$P_{22} \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1} t^2} \|[\mathbf{S}_{z(j)}]^s - [\mathbf{S}_{b:}]^s\|^4\right). \quad (81)$$

We take  $t = \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|$  in  $P_{21}$  and  $P_{22}$ , and plug the inequalities (80) and (81) into to the upper bound for  $P_2$  in (79). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2\right). \quad (82)$$

Combining the upper bounds (78) and (82) gives

$$\begin{aligned} \mathbb{P}\left(\left\langle e_i, [\tilde{\mathbf{S}}_{z(i)}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2\right) \\ \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2\right). \end{aligned} \quad (83)$$

Hence, we have

$$\begin{aligned} \mathbb{E}\xi &= \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{P}\left\{ \left\langle \mathbf{E}_i \mathbf{V}, [\tilde{\mathbf{S}}_{z(i)}]^s - [\tilde{\mathbf{S}}_{b:}]^s \right\rangle \right. \\ &\quad \left. \leq -\frac{\theta(i)m}{4} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \right\} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \\ &\lesssim \frac{1}{p} \sum_{i \in [p]} \theta(i) \max_{i \in [p], b \in [r]} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2 \\ &\quad \cdot \exp\left(\frac{p^{K-1}}{r^{K-1}} \|[\mathbf{S}_{z(i)}]^s - [\mathbf{S}_{b:}]^s\|^2\right) \\ &\leq \exp\left(-M \frac{p^{K-1}}{r^{K-1}} \Delta_{\min}^2\right), \end{aligned} \quad (84)$$

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 (83).

By Markov's inequality, we have

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

and thus the condition (65) 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

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

where the first inequality follows from the property in later inequality (105). We also have

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

Finally, we verify the bounded loss condition (66) 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

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

Therefore, the condition (66) holds for  $t = 0$ . Assume that the condition (66) also holds for all  $t \leq t_0$ . Then, by the decomposition (77), we have

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

where the second inequality follows from the condition (65) and the last inequality follows from the assumption that  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ . Thus, the condition (66) holds for  $t_0+1$ , and the condition (66) is proved by induction.  $\square$

### Useful Lemmas for the Proof of Theorem 5:

*Lemma 10 (Singular-Value Property of Membership Matrices):* Under the setup of Theorem 5, suppose that the condition (66) holds. Then, for all  $a \in [r]$ , we have  $|(\mathbf{z}^{(t)})^{-1}(a)| \asymp p/r$ . Moreover, we have

$$\lambda_r(\mathbf{M}) \asymp \|\mathbf{M}\|_\sigma \asymp \sqrt{p/r}, \quad \lambda_r(\mathbf{W}) \asymp \|\mathbf{W}\|_\sigma \asymp \sqrt{r/p}, \\ \lambda_r(\mathbf{P}) \asymp \|\mathbf{P}\|_\sigma \asymp \min_{a \in [r]} \|\theta_{z^{-1}(a)}\|^{-1} \lesssim \sqrt{r/p}. \quad (85)$$

The inequalities (85) also hold by replacing  $\mathbf{M}$  and  $\mathbf{W}$  to  $\mathbf{M}^{(t)}$  and  $\mathbf{W}^{(t)}$  respectively. Further, we have

$$\lambda_r(\mathbf{W}\mathbf{W}^T) \asymp \|\mathbf{W}\mathbf{W}^T\|_\sigma \asymp r/p, \quad (86)$$

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

*Proof of Lemma 10:* The proof for the inequality (85) for  $\mathbf{M}, \mathbf{W}$  can be found in [13, 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 (86), note that for all  $k \in [r]$ ,

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

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 11 (Upper Bound for  $F_{ib}^{(t)}, G_{ib}^{(t)}$  and  $H_{ib}^{(t)}$ ):* Under the Condition 1 and the setup of Theorem 5 with fixed  $r \geq 2$ , 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 (98) and (102). As  $p \rightarrow \infty$ , we have

$$\begin{aligned} \max_{i \in [p]} \max_{b \neq z(i)} \frac{(F_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \\ \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2} \|\mathbf{E}_{i:}\mathbf{V}\|^2 + \left(1 + \frac{rL^{(t)}}{\Delta_{\min}^2}\right) \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2, \quad (87) \end{aligned}$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(G_{ib}^{(t)})^2}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{512} (\Delta_{\min}^2 + L^{(t)}), \quad (88)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{|H_{ib}^{(t)}|}{\|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2} \leq \frac{1}{4}. \quad (89)$$

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

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(\hat{F}_{ib})^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \lesssim p^{K-1} \frac{rL(\hat{z})}{\Delta_{\min}^2} \quad (2381)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{(\hat{G}_{ib})^2}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \leq \frac{1}{512} (\Delta_{\min}^2 + L(\hat{z})), \quad (2382)$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{|\hat{H}_{ib}|}{\|[\mathbf{A}_{z(i)}]_s - [\mathbf{A}_b]_s\|^2} \leq \frac{1}{4}. \quad (2383)$$

*Proof of Lemma 11:* We prove the the first three inequalities in Lemma 11 separately.

1) Upper bound for  $F_{ib}^{(t)}$ , i.e., inequality (87). Recall the definition of  $F_{ib}^{(t)}$ ,

$$\begin{aligned} F_{ib}^{(t)} = 2 \left\langle \mathbf{E}_{i:}\mathbf{V}^{(t)}, \left([\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s\right) - \left([\tilde{\mathbf{S}}_b]_s - [\mathbf{S}_b]_s\right) \right\rangle \\ + 2 \left\langle \mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle. \end{aligned} \quad (2388)$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned} &(F_{ib}^{(t)})^2 \\ &\leq 8 \left( \left\langle \mathbf{E}_{i:}\mathbf{V}^{(t)}, \left([\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s\right) - \left([\tilde{\mathbf{S}}_b]_s - [\mathbf{S}_b]_s\right) \right\rangle \right)^2 \\ &\quad + 8 \left( \left\langle \mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)}), [\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s \right\rangle \right)^2 \\ &\leq 8 \left( \|\mathbf{E}_{i:}\mathbf{V}\|^2 + \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2 \right) \max_{a \in [r]} \|[\tilde{\mathbf{S}}_a]_s - [\mathbf{S}_a]_s\|^2 \\ &\quad + \|\mathbf{E}_{i:}(\mathbf{V} - \mathbf{V}^{(t)})\|^2 \|[\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\|. \end{aligned} \quad (90)$$

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

$$\begin{aligned} \|[\tilde{\mathbf{S}}_a]_s - [\mathbf{S}_a]_s\|^2 &= \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 \\ &\leq 2 \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]_s\|^2 \\ &\quad + 2 \|[\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}]_s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]_s\|^2 \\ &\lesssim \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\ &\lesssim rL^{(t)} + \frac{rr^{2K} + pr^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \\ &\lesssim rL^{(t)}, \end{aligned} \quad (91)$$

where the second inequality follows from the inequalities (108) and (109) in Lemma 12, the third inequality follows from the condition (66) 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

$$\begin{aligned} &\|[\tilde{\mathbf{S}}_{z(i)}]_s - [\tilde{\mathbf{S}}_b]_s\|^2 \\ &= \|[\tilde{\mathbf{S}}_{z(i)}]_s - [\mathbf{S}_{z(i)}]_s + [\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s + [\mathbf{S}_b]_s - [\tilde{\mathbf{S}}_b]_s\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 + \max_{a \in [r]} \|[\mathbf{S}_a]_s - [\tilde{\mathbf{S}}_a]_s\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2 + \max_{a \in [r]} \frac{1}{\|\mathbf{S}_a\|^2} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}\|^2 \\ &\lesssim \|[\mathbf{S}_{z(i)}]_s - [\mathbf{S}_b]_s\|^2, \end{aligned} \quad (92)$$

where the second inequality follows from Lemma 4, and the last inequality follows from the assumptions on  $\|\mathbf{S}_{a:}\|$  in the parameter space (2), the inequality (62) in Condition 1 and the assumption  $\Delta_{\min}^2 \gtrsim p^{-K/2} \log p$ .

Therefore, we finish the proof of inequality (87) by plugging the inequalities (91) and (92) into the upper bound (90).

2) Upper bound for  $G_{ib}^{(t)}$ , i.e., inequality (88). By definition of  $G_{ib}^{(t)}$ , we rearrange terms and obtain

$$\begin{aligned} G_{ib}^{(t)} &= \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &\quad - \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &= 2 \left\langle [\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s, \left( [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s \right) \right. \\ &\quad \left. - \left( [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s \right) \right\rangle \\ &= G_1 + G_2 - G_3, \end{aligned} \tag{93}$$

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

$$\begin{aligned} |G_1|^2 &\leq \left| \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \right. \\ &\quad \left. - \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{b:}^{(t)}]^s\|^2 \right|^2 \\ &\leq \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^4 \\ &\leq C^4 \frac{r^4}{\Delta_{\min}^4} (L^{(t)})^4 + \frac{r^2 r^{4K} + p^2 r^{2K+4}}{p^{2K}} \frac{(L^{(t)})^2}{\Delta_{\min}^4} \\ &\leq C^4 \frac{\bar{C}}{\tilde{C}^3} \left( \Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)} \right), \end{aligned} \tag{94}$$

where the third inequality follows from the inequality (110) in Lemma 12 and the last inequality follows from the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) 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

$$\begin{aligned} |G_2|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\quad \cdot \|[\mathbf{W}_{:z(i)}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{S}_{z(i):}^{(t)}]^s\|^2 \\ &\leq \frac{2}{\|\mathbf{W}_{z(i):}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2 \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \end{aligned}$$

$$\leq C' \frac{r^{2K-1} + K p r^{K+1}}{p^K} \tag{2447}$$

$$\cdot \left( \frac{r^2}{\Delta_{\min}^2} (L^{(t)})^2 + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) \tag{2448}$$

$$\leq \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)}, \tag{2449}$$

where  $C'$  is a positive universal constant, the second inequality follows from Lemma 4, the third inequality follows from the inequality (63) in Condition 1, the inequalities (110) and (129) in the proof of Lemma 12, and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) in Condition 1.

For  $G_3$ , note that by triangle inequality

$$\begin{aligned} &\|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2}, \end{aligned} \tag{96}$$

where the last inequality follows from the inequality (128) in the proof of Lemma 12 and  $C$  is a positive constant. Then we have

$$\begin{aligned} |G_3|^2 &\leq 2 \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq 2 \left( \|[\mathbf{X}_{i:} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \right. \\ &\quad \left. + \|[\mathbf{W}_{:b}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \right) \\ &\quad \cdot \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{Y} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}]^s\|^2 \\ &\leq C^2 \left( \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 + C \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \right) \\ &\quad \cdot \left( \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) + \frac{C'}{\tilde{C}^2} \Delta_{\min}^2 L^{(t)} \\ &\leq \frac{C^2 \bar{C}^2}{\tilde{C}} \|\mathbf{S}_{z(i):}^s - \mathbf{S}_{b:}^s\|^2 (\Delta_{\min}^2 + L^{(t)}) \\ &\quad + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} \left( \Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)} \right), \end{aligned} \tag{97}$$

where the third inequality follows from the same procedure to derive (94) and (95), and the last inequality follows from the assumption  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$  and inequality (66) in Condition 1.

Choose the  $\tilde{C}$  such that

$$3 \left( C^4 \frac{\bar{C}}{\tilde{C}^3} + \frac{C'}{\tilde{C}^2} + \frac{C^2 \bar{C}^2}{\tilde{C}} + \frac{C^3 C' \bar{C}^2}{\tilde{C}^2} \right) \leq \frac{1}{512}. \tag{98}$$

Then, we finish the proof of inequality (88) by plugging the inequalities (94), (95), and (97) into the upper bound (93).

3) Upper bound for  $H_{ib}^{(t)}$ , i.e., the inequality (89). By definition of  $H_{ib}$ , we rearrange terms and obtain

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

where

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

For  $H_1$ , we have

$$\begin{aligned} |H_1| &\leq \frac{4 \max_{a \in [r]} \|\mathbf{W}_{:a}^T \mathbf{E} \mathbf{V}^{(t)}\|^2}{\|\mathbf{W}_{:z(i)}^T \mathbf{X} \mathbf{V}^{(t)}\|^2} \\ &\leq \frac{r^{2K-1} + K p r^{K+1}}{p^K} \\ &\leq \tilde{C}^{-2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{b:}]^s\|^2, \end{aligned} \quad (99)$$

following the derivation of  $G_2$  in inequality (95) and the assumption that  $\Delta_{\min}^2 \geq \tilde{C} p^{-K/2} \log p$ .

For  $H_2$ , by the inequality (96), we have

$$\begin{aligned} |H_2| &\lesssim 2 \max_{a \in [r]} \|[\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s - [\mathbf{W}_{:a}^T \mathbf{X} \mathbf{V}^{(t)}]^s\|^2 \\ &\lesssim \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \\ &\leq C \frac{\bar{C}^2}{\tilde{C}^2} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \end{aligned} \quad (100)$$

where the last inequality follows from the condition (66) in Condition 1.

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

$$\begin{aligned} |H_3| &\lesssim \|[\mathbf{X}_{i:}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}^{(t)}]^s\| |H_1|^{1/2} \\ &\leq 2 \tilde{C}^{-1} \|[\mathbf{S}_{z(i):}]^s - [\mathbf{S}_{a:}]^s\|^2, \end{aligned} \quad (101)$$

following the inequalities (96) and (99).

Choose  $\tilde{C}$  such that

$$\tilde{C}^{-2} + C \frac{\bar{C}^2}{\tilde{C}^2} + \tilde{C}^{-1} \leq \frac{1}{4}. \quad (102)$$

Therefore, we finish the proof of inequality (89) combining inequalities (99), (100), and (101).

Next, we Show the Upper Bounds for  $\hat{F}_{ib}$ ,  $\hat{G}_{ib}$  and  $\hat{H}_{ib}$ . By Lemma 1, we have

$$\|\mathbf{S}_{a:}^s - \mathbf{S}_{b:}^s\| = (1 + o(1)) \|\mathbf{A}_{a:}^s - \mathbf{A}_{b:}^s\|. \quad (2517)$$

Also, notice that the matrix product of  $\mathbf{B}^T$  corresponds to the padding operation in Lemma 5, and the padding weights are balanced such that  $\|\mathbf{v} \mathbf{B}\| = (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{v}\|$  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

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = (1 + o(1)) \|[\mathbf{v}_1 \mathbf{B}^T]^s - [\mathbf{v}_2 \mathbf{B}^T]^s\|. \quad (103) \quad (2523)$$

The equation (103) also holds for  $\hat{\mathbf{B}}^T$ .

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

$$\begin{aligned} \|\mathbf{A}_{i:} \hat{\mathbf{Q}}\| &= \|\mathbf{S}_{z(i):} \mathbf{B}^T \hat{\mathbf{Q}}\| \\ &= \|\mathbf{S}_{z(i):} \hat{\mathbf{D}}^{\otimes(K-1)}\| \\ &= (1 + o(1)) \|\mathbf{S}_{z(i):}\| \\ &= (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{-(K-1)/2} \|\mathbf{A}_{i:}\|, \end{aligned} \quad (104) \quad (2526-2529)$$

where the third inequality follows from the singular property of MLE confusion matrix (135) and the last inequality follows from the fact that  $\mathbf{A}_i = \mathbf{S}_{z(i):} \mathbf{B}^T$  and Lemma 10. 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

$$\begin{aligned} &\|[\mathbf{X}_i \hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}]^s\| \\ &= \left\| \frac{\mathbf{A}_{z(i):} \hat{\mathbf{Q}}}{\|\mathbf{A}_{z(i):} \hat{\mathbf{Q}}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|} \right\| \\ &= (1 + o(1)) \left\| \frac{\mathbf{A}_{z(i):}}{\|\mathbf{A}_{z(i):}\|} - \frac{\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T}{\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\|} \right\| \\ &= (1 + o(1)) \|[\mathbf{X}_i]^s - [\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T]^s\| \end{aligned} \quad (105) \quad (2538-2541)$$

where the second equation follows from (104),  $\|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}} \hat{\mathbf{B}}^T\| = (1 + o(1)) \max_a \|\theta_{z^{-1}(a)}\|^{(K-1)/2} \|\mathbf{P}_{:a}^T \mathbf{Y} \hat{\mathbf{Q}}\|$ , 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

$$\begin{aligned} (\hat{F}_{ib})^2 &\leq \|\mathbf{E}_{i:}\|^2 \|[\bar{\mathbf{A}}_{a:}]^s - [\hat{\mathbf{A}}_{a:}]^s\|^2 \\ &\leq \|\mathbf{E}_{i:}\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T]^s - [\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| \right. \\ &\quad \quad \left. + \|[\bar{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s - [\hat{\mathbf{S}}_{a:} \hat{\mathbf{B}}^T]^s\| \right]^2 \\ &\lesssim \|\mathbf{E}_{i:}\|^2 \left[ \|[\bar{\mathbf{S}}_{a:} \mathbf{B}^T \hat{\mathbf{Q}}]^s - [\bar{\mathbf{S}}_{a:}]^s\| + \|[\bar{\mathbf{S}}_{a:}]^s - [\hat{\mathbf{S}}_{a:}]^s\| \right]^2. \end{aligned} \quad (2550-2553)$$

Following similar derivations in inequalities (91), (92), and the upper bound for  $J_1$  in the proof of Lemma 12, respectively,

we have

$$\|[\bar{S}_{a:}]^s - [\hat{S}_{a:}]^s\| \lesssim rL(\hat{z}), \quad \|[\bar{S}_{a:}]^s - [\bar{S}_{b:}]^s\| \lesssim \|S_{a:}^s - S_{b:}^s\|^2,$$

and

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

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 (105), we have

$$(1 + o(1))\hat{G}_{ib}$$

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

$$(1 + o(1))\hat{H}_{ib}$$

$$= \|[\mathbf{X}_{i:}\hat{\mathbf{Q}}]^s - [\mathbf{P}_{:a}^T\mathbf{Y}\hat{\mathbf{Q}}]^s\|_F^2 - \|[\mathbf{X}_{i:}\hat{\mathbf{Q}}]^s - [\mathbf{P}_{:b}^T\mathbf{Y}\hat{\mathbf{Q}}]^s\|_F^2 \\ + \|A_{a:}^s - A_{b:}^s\|_F^2.$$

We obtain the upper bounds following the proof for inequalities (88) and (89).  $\square$

**Lemma 12 (Relationship Between Misclustering Loss and Intermediate Parameters):** Under the Condition 1 and the setup of Theorem 5 with fixed  $r \geq 2$ , as  $p \rightarrow \infty$ , we have

$$\|\mathbf{V} - \mathbf{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}} \frac{r}{\Delta_{\min}^2} L^{(t)}}, \quad (106)$$

$$\|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}} \frac{r}{\Delta_{\min}^2} L^{(t)}}, \quad (107)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}]^s\| \\ \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} \right), \quad (108)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ \leq C \left( \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} + \frac{rL^{(t)}}{\Delta_{\min}} \right), \quad (109)$$

$$\max_{b \in [r]} \|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ \leq C \left( \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} \right), \quad (110)$$

for some positive universal constant  $C$ . In addition, the inequality (109) 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 12:* We follow and use several intermediate conclusions in [13, Proof of Lemma 5]. We prove each inequality separately.

1) Inequality (106). By [13, 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 (106) by applying Lemma 2 to the above inequality.

2) Inequality (107). By [13, Proof of Lemma 5], we have

$$\|\mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}} r \ell^{(t)}}. \quad (2595)$$

Also, we complete the proof of inequality (106) by applying Lemma 2 to the above inequality.

3) Inequality (108). We upper bound the desired quantity by triangle inequality,

$$\|[\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}]^s\| \leq I_1 + I_2 + I_3, \quad (2600)$$

where

$$I_1 = \left\| \frac{\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}}{\|\mathbf{W}_{:b}^T\mathbf{X}\mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \right\|, \quad (2602)$$

$$I_2 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \right) \mathbf{W}_{:b}^T\mathbf{Y}\mathbf{V} \right\|, \quad (2603)$$

$$I_3 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \right) \mathbf{W}_{:b}^{(t),T}\mathbf{Y}\mathbf{V} \right\|. \quad (2604)$$

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

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

$$I_1 \leq I_{11} + I_{12}, \quad (2607)$$

where

$$I_{11} = \left\| \frac{\mathbf{W}_{:b}^T\mathbf{X}\mathbf{V}}{\|\mathbf{W}_{:b}^T\mathbf{X}\mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \right\|, \quad (2609)$$

and

$$I_{12} = \left\| \frac{\mathbf{W}_{:b}^T\mathbf{E}\mathbf{V}}{\|\mathbf{W}_{:b}^T\mathbf{X}\mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T}\mathbf{E}\mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T}\mathbf{X}\mathbf{V}\|} \right\|. \quad (2611)$$

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

$$D_{ab} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = a, z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}}, \text{ for all } a, b \in [r]. \quad (2614)$$

By Lemma 10, we have  $\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} \gtrsim p/r$ . Then, we have

$$\sum_{a \neq b, a, b \in [r]} D_{ab} \lesssim \frac{r}{p} \sum_{i: z^{(t)}(i) \neq z(i)} \theta(i) \lesssim \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \frac{1}{\log p}, \quad (2617)$$

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

$$D_{bb} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z(i) = z^{(t)}(i) = b\}}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \quad (2619)$$

$$\geq \frac{c(\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\} - p\ell^{(t)})}{\sum_{i \in [p]} \mathbb{1}\{z^{(t)}(i) = b\}} \quad (2620)$$

$$\gtrsim 1 - \frac{1}{\log p}, \quad (2621)$$

under the inequality (66) in Condition 1. By the definition of  $\mathbf{W}, \mathbf{W}^{(t)}, \mathbf{V}$ , we have

$$\frac{\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} = [\mathbf{S}_{b:}]^s,$$

and

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

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 (2) and the bounds of  $\mathbf{D}$  in (111) and (112).

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

$$\begin{aligned} & \| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \| \\ &= \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \| (\mathbf{I}_d - \text{Proj}(\mathbf{S}_{b:})) \mathbf{S}_{a:} \| \\ &= \sum_{a \neq b, a \in [r]} D_{ab} \| \mathbf{S}_{a:} \sin(\mathbf{S}_{b:}, \mathbf{S}_{a:}) \| \\ &\leq \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\|, \end{aligned} \quad (113)$$

where the first inequality follows from the triangle inequality, and the last inequality follows from Lemma 4. Note that with bounds (111) and (112), when  $p$  is large enough, we have

$$\begin{aligned} \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| &= \|D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}\| \\ &\geq D_{bb} \|\mathbf{S}_{b:}\| - \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \\ &\geq C_1, \end{aligned} \quad (114)$$

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

$$\begin{aligned} I_{11} &\leq \sqrt{2} \sin \alpha \\ &= \frac{\| [D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:}] \sin \alpha \|}{\| D_{bb} \mathbf{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} \mathbf{S}_{a:} \|} \\ &\leq \frac{1}{C_1} \sum_{a \neq b, a \in [r]} D_{ab} \|\mathbf{S}_{a:}\| \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \\ &\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1} \{ z^{(t)}(i) = b \} \|\mathbf{S}_{b:}^s - \mathbf{S}_{a:}^s\| \\ &\leq \frac{r L^{(t)}}{\Delta_{\min}}, \end{aligned} \quad (115)$$

where the second inequality follows from (113) and (114), 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 (2).

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

$$\begin{aligned} I_{12} &\leq \frac{1}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| \\ &\quad + \frac{\| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \|}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\|. \end{aligned} \quad (116)$$

By [13, Proof of Lemma 5], we have

$$\begin{aligned} \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}. \quad (116) \\ \text{Notice that} \quad & \\ \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \| &\leq \|\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}\| \|\mathbf{X} \mathbf{V}\|_F \quad (116) \\ &\lesssim \frac{r^{3/2} L^{(t)}}{\sqrt{p} \Delta_{\min}^2} \|S\| \|\Theta M\|_\sigma \quad (116) \\ &\lesssim \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}}, \end{aligned} \quad (117)$$

where the second inequality follows from [13, Inequality (121), Proof of Lemma 5] and the last inequality follows from Lemma 6 and (66) 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 (114). Therefore, we have

$$\begin{aligned} I_{12} &\lesssim \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{E} \mathbf{V} \| \\ &\quad + \| (\mathbf{W}_{:b}^T - \mathbf{W}_{:b}^{(t),T}) \mathbf{X} \mathbf{V} \| \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\| \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}} + \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}} \sqrt{\frac{r^{2K}}{p^K}} \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}, \end{aligned} \quad (118)$$

where second inequality follows from the inequalities (116), (117), and (62) in Condition 1.

Hence, combining inequalities (115) and (118) yields

$$I_1 \lesssim \frac{r L^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}. \quad (119)$$

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 (114). By triangle inequality and (62) in Condition 1, we have

$$I_2 \leq \frac{\|\mathbf{W}_{:b}^T \mathbf{E} \mathbf{V}\|}{\|\mathbf{W}_{:b}^T \mathbf{X} \mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^T \mathbf{E} \mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}, \quad (120)$$

and

$$I_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}. \quad (121)$$

Therefore, combining the inequalities (119), (120), and (121), we finish the proof of inequality (108).

4) Inequality (109). Here we only show the proof of inequality (109) 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

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

where

$$J_1 = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \right\|,$$

$$J_2 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V} \right\|,$$

$$J_3 = \left\| \left( \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)}\|} - \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \right) \mathbf{W}_{:b}^{(t),T} \mathbf{Y} \mathbf{V}^{(t)} \right\|.$$

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

$$J_{11} = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \right\|$$

and

$$J_{12} = \left\| \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} - \frac{\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}^{(t)}}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \right\|.$$

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

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

for  $k = 1, \dots, K-1$ . Let  $\beta_k$  denote the angle between  $\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^k$  and  $\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{k+1}$ . With the same idea to prove  $I_{11}$  in inequality (115), 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  $\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^k$  and  $\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{k+1}$ . Before the specific derivation of the inner product, note that

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

where

$$\mathcal{T}_k = \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T$$

$$\times_{k+1} \mathbf{W}^{(t),T} \times_{k+2} \cdots \times_K \mathbf{W}^{(t),T}$$

$$\mathcal{T}_{k+1} = \mathcal{X} \times_1 \mathbf{W}_{:b}^{(t),T} \times_2 \mathbf{W}^T \times_3 \cdots \times_k \mathbf{W}^T$$

$$\times_{k+1} \mathbf{W}^T \times_{k+2} \cdots \times_K \mathbf{W}^{(t),T}.$$

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

$$\langle \mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^k, \mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{k+1} \rangle$$

$$= \langle \text{Mat}_{k+1}(\mathcal{T}_k), \text{Mat}_{k+1}(\mathcal{T}_{k+1}) \rangle$$

$$= \langle \mathbf{D}^T \mathbf{S} \mathbf{Z}^k, \mathbf{S} \mathbf{Z}^k \rangle$$

$$= \sum_{b \in [r]} \left( D_{bb} \|\mathbf{S}_{b:} \mathbf{Z}^k\|^2 + \sum_{a \neq b, a \in [r]} D_{ab} \langle \mathbf{S}_{a:} \mathbf{Z}^k, \mathbf{S}_{b:} \mathbf{Z}^k \rangle \right)$$

$$\gtrsim (1 - \log p^{-1}) \min_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2 - \log p^{-1} \max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\|^2,$$

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 (111) and (112).

Note that

$$\|\mathbf{D}\|_\sigma \leq \|\mathbf{D}\|_F$$

$$\leq \sqrt{\sum_{b \in [r]} D_{bb}^2 + (\sum_{a \neq b, a \in [r]} D_{ab})^2}$$

$$\lesssim \sqrt{r + \log^2 p^{-1}} \lesssim 1,$$

where the second inequality follows from inequality (111), 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

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

following the Lemma 6 and Lemma 10. Then, for all  $k \in [K]$ , we have

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

$$\leq \|\mathbf{Z}^k\|_\sigma \leq \|\mathbf{D}_{:b}\| \|\mathbf{D}\|_\sigma^{K-k-1} \lesssim 1.$$

Thus, we have bounds

$$\max_{a \in [r]} \|\mathbf{S}_{a:} \mathbf{Z}^k\| \leq \max_{a \in [r]} \|\mathbf{S}_{a:}\| \|\mathbf{Z}^k\|_\sigma \lesssim 1,$$

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

Hence, when  $p$  is large enough, the inner product (122) is positive, which implies  $\beta_k \in [0, \pi/2]$  and thus  $2 \sin \frac{\beta_k}{2} \leq \sqrt{2} \sin \beta_k$ .

2756 Next, we upper bound the trigonometric function  $\sin \beta_k$ .  
 2757 Note that

$$\begin{aligned} 2758 \quad \sin \beta_k &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}, \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}) \\ 2759 &\leq \sin \beta_{k1} + \sin \beta_{k2}, \end{aligned}$$

2760 where

$$\begin{aligned} 2761 \quad \sin \beta_{k1} &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}, \\ 2762 &\quad \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \tilde{\mathbf{D}} \otimes \mathbf{D}^{\otimes K-k-1}), \\ 2763 \quad \sin \beta_{k2} &= \sin(\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \tilde{\mathbf{D}} \otimes \mathbf{D}^{\otimes K-k-1}, \\ 2764 &\quad \mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}), \end{aligned}$$

2765 and  $\tilde{\mathbf{D}}$  is the normalized confusion matrix with entries  $\tilde{\mathbf{D}}_{ab} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z^{(t)}=b, z(i)=a\}}{\sum_{i \in [p]} \theta(i) \mathbb{1}\{z^{(t)}=b\}}$ .

2766 To bound  $\sin \beta_{k1}$ , recall Definition 2 that for any cluster  
 2767 assignment  $\bar{z}$  in the  $\varepsilon$ -neighborhood of true  $z$ ,

$$\begin{aligned} 2769 \quad \mathbf{p}(\bar{z}) &= (|\bar{z}^{-1}(1)|, \dots, |\bar{z}^{-1}(r)|)^T, \\ 2770 \quad \mathbf{p}_{\theta}(\bar{z}) &= (\|\theta_{\bar{z}^{-1}(1)}\|_1, \dots, \|\theta_{\bar{z}^{-1}(r)}\|_1)^T. \end{aligned}$$

2771 Note that we have  $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2} \leq \frac{\bar{C}}{C} r \log^{-1}(p)$  by Condition 1 and Lemma 2. Then, with the locally linear stability  
 2772 assumption, the  $\theta$  is  $\ell^{(t)}$ -locally linearly stable; i.e.,  
 2773

$$2774 \quad \sin(\mathbf{p}(z^{(t)}), \mathbf{p}_{\theta}(z^{(t)})) \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

2775 Note that  $\text{diag}(\mathbf{p}(z^{(t)}))\mathbf{D} = \text{diag}(\mathbf{p}_{\theta}(z^{(t)}))\tilde{\mathbf{D}}$ , and  
 2776  $\sin(\mathbf{a}, \mathbf{b}) = \min_{c \in \mathbb{R}} \frac{\|\mathbf{a}-c\mathbf{b}\|}{\|\mathbf{a}\|}$  for vectors  $\mathbf{a}, \mathbf{b}$  of same  
 2777 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,  
 2778 we have

$$\begin{aligned} 2779 \quad &\min_{c \in \mathbb{R}} \|\mathbf{D} - c\tilde{\mathbf{D}}\|_F \\ 2780 &\leq \|\mathbf{I}_r - c_0 \text{diag}(\mathbf{p}(z^{(t)})) \text{diag}^{-1}(\mathbf{p}_{\theta}(z^{(t)}))\|_F \|\mathbf{D}\|_F \\ 2781 &\lesssim \frac{\|\mathbf{p}(z^{(t)}) - c_0 \mathbf{p}_{\theta}(z^{(t)})\|}{\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1} \\ 2782 &= \frac{\|\mathbf{p}(z^{(t)})\|}{\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1} \sin(\mathbf{p}(z^{(t)}), \mathbf{p}_{\theta}(z^{(t)})) \\ 2783 &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned}$$

2784 where the last inequality follows from Lemma 10, the  
 2785 constraint  $\min_{i \in [p]} \theta(i) \geq c > 0$ ,  $\|\mathbf{p}(z^{(t)})\| \lesssim p$  and  
 2786  $\min_{a \in [r]} \|\theta_{z^{(t)}, -1(a)}\|_1 \gtrsim p$ .

2787 By the geometry property of trigonometric function,  
 2788 we have

$$\begin{aligned} 2789 \quad \sin \beta_{k1} &= \min_{c \in \mathbb{R}} \frac{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{D} - c\tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1}\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes \mathbf{D}^{\otimes K-k}\|} \\ 2790 &\leq \frac{\|\mathbf{D}_{:b}^T \mathbf{S}\| \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_{\sigma} \|\mathbf{D}\|_{\sigma}^{K-k-1}}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k}(\mathbf{D})} \\ 2791 &\lesssim \|\mathbf{D} - c_0 \tilde{\mathbf{D}}\|_F \\ 2792 &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{126}$$

2793 where the second inequality follows from the singular property  
 2794 of  $\mathbf{D}$  in (123), (124) and the constraint of  $\mathbf{S}$  in (2).

2795 To bound  $\sin \beta_{k2}$ , let  $\mathbf{C} = \text{diag}(\{\|\mathbf{S}_{a:}\|\}_{a \in [r]})$ . We have

$$\begin{aligned} 2796 \quad \sin \beta_{k2} &\lesssim \frac{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k-1} \otimes (\mathbf{I}_r - \tilde{\mathbf{D}}) \otimes \mathbf{D}^{\otimes K-k-1}\|}{\|\mathbf{D}_{:b}^T \mathbf{S} \mathbf{I}_r^{\otimes k} \otimes \mathbf{D}^{\otimes K-k-1}\|} \\ 2797 &\lesssim \frac{\|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{Z}^k\|_F}{\|\mathbf{D}_{:b}^T \mathbf{S}\| \lambda_r^{K-k-1}(\mathbf{D})} \\ 2798 &\lesssim \|(\mathbf{I}_r - \tilde{\mathbf{D}}^T) \mathbf{S} \mathbf{C}^{-1}\|_F \|\mathbf{C} \mathbf{Z}^k\|_{\sigma} \\ 2799 &\lesssim \frac{r}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]} \mathbb{1}\{z^{(t)}(i) = b\} \|\mathbf{S}_{b:}^s - \mathbf{S}_{z(i)}^s\| \\ 2800 &\lesssim \frac{L^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{127}$$

2801 where the third inequality follows from the singular property  
 2802 of  $\mathbf{D}$  and the boundedness of  $\mathbf{S}$ , and the fourth inequality  
 2803 follows from the definition of  $\tilde{\mathbf{D}}$ , boundedness of  $\mathbf{S}$ , the  
 2804 lower bound of  $\theta$ , and the singular property of  $\mathbf{Z}^k$  in inequality  
 2805 (125), and the last line follows from the definition of  $L^{(t)}$ .  
 2806

Combining (126) and (127) yields

$$\sin \beta_k \leq \sin \beta_{k1} + \sin \beta_{k2} \lesssim \frac{L^{(t)}}{\Delta_{\min}}.$$

Finally, by triangle inequality, we obtain

$$2808 \quad J_{11} \leq \sum_{k=1}^{K-1} J_{11}^k \lesssim \sum_{k=1}^{K-1} \sin \beta_k \lesssim (K-1) \frac{r L^{(t)}}{\Delta_{\min}}. \tag{128}$$

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

$$\begin{aligned} 2810 \quad J_{12} &\leq \frac{1}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| \\ 2811 &\quad + \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}\| \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\|} \|\mathbf{W}_{:b}^{(t),T} \mathbf{E} \mathbf{V}^{(t)}\|. \end{aligned} \tag{129}$$

Note that

$$\begin{aligned} 2814 \quad \|\mathbf{W}_{:b}^{(t),T} \mathbf{X} \mathbf{V}^{(t)}\| &= \|\mathbf{D}^T \mathbf{S} \mathbf{Z}^1\| \\ 2815 &\geq \lambda_r(\mathbf{D}) \|\mathbf{S}\| \lambda_{r^{K-2}}(\mathbf{Z}^1) \gtrsim 1, \\ 2816 \end{aligned} \tag{129}$$

where the inequality follows from the bounds (124) and (125).  
 2817

By [13, Proof of Lemma 5], we have

$$\begin{aligned} 2819 \quad &\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}(\mathbf{V} - \mathbf{V}^{(t)})\| \\ 2820 &\lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}} \frac{(K-1)\sqrt{L^{(t)}}}{\Delta_{\min}}. \end{aligned} \tag{130}$$

Notice that

$$\begin{aligned} 2822 \quad &\|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F \\ 2823 &\leq \|(\mathbf{I} - \mathbf{D}^T) \mathbf{S}(\mathbf{I}_r^{\otimes (k-1)} \otimes \mathbf{D}^{\otimes (K-k-1)})\|_F \\ 2824 &\leq \|(\mathbf{W}^T - \mathbf{W}^{(t),T}) \Theta \mathbf{M}\|_F \|\mathbf{S}\|_F \|\mathbf{D}\|_{\sigma}^{K-k-1} \\ 2825 &\lesssim \|\mathbf{W}^T - \mathbf{W}^{(t),T}\| \|\Theta \mathbf{M}\|_{\sigma} \\ 2826 &\lesssim \frac{\sqrt{r L^{(t)}}}{\Delta_{\min}}, \end{aligned} \tag{131}$$

where the first inequality follows from the tensor algebra in inequality (122), the second inequality follows from the fact that  $\mathbf{I} = \mathbf{W}^T \Theta \mathbf{M}$ , and the last inequality follows from [13, Proof of Lemma 5]. It follows from (131) and Lemma 10 that

$$\begin{aligned} \|\mathbf{W}_{:b}^{(t),T} \mathbf{X}(\mathbf{V} - \mathbf{V}^{(t)})\| &\leq \|\mathbf{W}_{:b}^{(t),T}\| \sum_{k=1}^{K-1} \|\mathbf{X}(\mathbf{V}^k - \mathbf{V}^{k+1})\|_F \\ &\lesssim \frac{\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}}. \end{aligned} \quad (132)$$

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 (114) and (129), respectively. We have

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

where the second inequality follows from inequalities (130), (132), and the inequality (62) 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 (114) and (129), respectively. By triangle inequality and inequality (62) in Condition 1, we have

$$J_2 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}\mathbf{V}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}, \quad (133)$$

and

$$J_3 \leq \frac{\|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}^{(t)}\|}{\|\mathbf{W}_{:b}^{(t),T} \mathbf{X}\mathbf{V}^{(t)}\|} \lesssim \|\mathbf{W}_{:b}^{(t),T} \mathbf{E}\mathbf{V}\| \lesssim \frac{r^K}{p^{K/2}}. \quad (134)$$

Therefore, combining the inequalities (128), (133), and (134), we finish the proof of inequality (109).

5) Inequality (110). By triangle inequality, we upper bound the desired quantity

$$\begin{aligned} &\|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ &\leq \|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}^{(t)}]^s - [\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}]^s\| \\ &\quad + \|[\mathbf{W}_{:b}^T \mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}]^s\| \\ &\quad + \|[\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}]^s - [\mathbf{W}_{:b}^{(t),T} \mathbf{Y}\mathbf{V}^{(t)}]^s\| \\ &\lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K} \frac{\sqrt{L^{(t)}}}{\Delta_{\min}}}, \end{aligned}$$

following the inequalities (108) and (109). Therefore, we finish the proof of inequality (110).

Next, we Show the Intermediate Inequalities Holds With  $\mathbf{P}, \mathbf{Q}$  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

$$\hat{D}_{ab} = \frac{\sum_{i \in [p]} \theta(i) \hat{\theta}(i) \mathbb{1}\{\hat{z}(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2}$$

$$= \frac{\sum_{i \in [p]} (1 + o(p^{K-2})) (\hat{\theta}(i))^2 \mathbb{1}\{\hat{z}(i) = a, \hat{z}(i) = b\}}{\|\hat{\theta}_{\hat{z}^{-1}(b)}\|^2}, \quad (135)$$

where the second equation follows from Lemma 13, and thus  $\sum_{a \in [r]} \hat{D}_{ab} = 1 + o(1)$ . By the derivation of (111), (112), (124), and (123), we have

$$\sum_{a \neq b \in [r]} \hat{D}_{ab} \lesssim \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\hat{\theta}(i))^2 \lesssim \frac{1}{\log p},$$

$$\hat{D}_{bb} \gtrsim 1 - \frac{1}{\log p}, \quad \lambda_{\min}(\hat{\mathbf{D}}) \asymp \|\hat{\mathbf{D}}\|_{\sigma} = (1 + o(1)).$$

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 12. The analogies of inequalities (106), (107), (108), (109), and (110) hold by using the MLE confusion matrix and the definition of  $L(\hat{z})$ .

Particularly, for the analogy of (109), the usage of MLE confusion matrix avoids the stability condition on  $\theta$ . Let  $\bar{\mathbf{D}}$  be the normalized version of  $\hat{\mathbf{D}}$ . The angle in inequality (126) decays to 0 at speed  $p^{-(K-2)} \lesssim \Delta_{\min}$  when  $K \geq 3$ , and the inequality (127) holds by the fact that

$$\|(\mathbf{I}_r - \bar{\mathbf{D}})\mathbf{S}\mathbf{C}^{-1}\|_F \lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{S}_{b:}^s - \mathbf{S}_{z(i):}^s\|$$

$$\lesssim \frac{r}{p} \sum_{i \in [p]} (\theta(i))^2 \sum_{b \in [r]} \|\mathbf{A}_{b:}^s - \mathbf{A}_{z(i):}^s\|.$$

□

**Lemma 13 (Polynomial Estimation Error of MLE):** Let  $(\hat{z}, \hat{\mathcal{S}}, \hat{\theta})$  denote the MLE in (9) with fixed  $K \geq 2$  and symmetric mean tensor, and  $\hat{\mathcal{X}}$  denote the mean tensor consisting of parameter  $(\hat{z}, \hat{\mathcal{S}}, \hat{\theta})$ . With high probability going to 1 as  $p \rightarrow \infty$ , we have

$$\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \lesssim \sigma^2 (r^K + Kpr),$$

with probability going to 1. When SNR  $\gtrsim p^{-(K-1)} \log p$ ,  $\theta$  is balanced, and  $\min_{i \in [p]} \theta(i) \geq c$  for some positive constant  $c$ , the MLE satisfies

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\theta(i))^2 \lesssim \frac{1}{r \log p},$$

$$\frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\hat{\theta}(i))^2 \lesssim \frac{1}{r \log p},$$

$$\text{and } L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p},$$

Further, we have

$$\theta(i)^2 = (1 + o(p^{-(K-2)})) \hat{\theta}(i)^2.$$

Proof of Lemma 13: Without loss of generality, we assume  $\sigma^2 = 1$  and identity mapping minimizes the misclustering error for MLE. For arbitrary two sets of parameters  $(z, \mathcal{S}, \boldsymbol{\theta}), (z', \mathcal{S}', \boldsymbol{\theta}') \in \mathcal{P}(\gamma)$  and corresponding mean tensors  $\mathcal{X}, \mathcal{X}'$ , we have

$$\begin{aligned} & \text{rank}(\text{Mat}_k(\mathcal{X}) - \text{Mat}_k(\mathcal{X}')) \\ & \leq \text{rank}(\text{Mat}_k(\mathcal{X})) + \text{rank}(\text{Mat}_k(\mathcal{X}')) \\ & \leq 2r, \quad k \in [K]. \end{aligned}$$

Hence, we have

$$\mathcal{X} - \mathcal{X}' \in \mathcal{Q}(2r, \dots, 2r), \quad (136)$$

where  $\mathcal{Q}(r, \dots, r) := \{\text{Tucker tensor with rank } (r, \dots, r)\}$ . Then, we obtain that

$$\begin{aligned} & \mathbb{P}(\|\mathcal{X} - \hat{\mathcal{X}}_{ML}\|_F \geq t) \\ & \leq 2\mathbb{P}\left(\sup_{\mathcal{X}, \mathcal{X}' \in \mathcal{Q}(r, \dots, r)} \left\langle \frac{\mathcal{X} - \mathcal{X}'}{\|\mathcal{X} - \mathcal{X}'\|_F}, \mathcal{E} \right\rangle \geq t\right) \\ & \leq 2\mathbb{P}\left(\sup_{\mathcal{T} \in \mathcal{Q}(2r, \dots, 2r) \cap \{\|\mathcal{T}\|_F=1\}} \langle \mathcal{T}, \mathcal{E} \rangle \geq t\right) \\ & \lesssim \exp(-Kpr), \end{aligned}$$

with the choice  $t \asymp \sigma\sqrt{(Kpr + r^K)}$ . Here the first inequality follows from [10, Lemma 1], the second inequality follows from (136), and the last inequality follows from [37, Lemma E5].

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

$$\begin{aligned} \frac{1}{p} \sum_{i \in [p]} \mathbb{1}\{\hat{z}(i) \neq z(i)\} (\theta(i))^2 & \lesssim \frac{r^{K-1}}{\Delta_{\min}^2 p^K} \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \\ & \lesssim \frac{r^{K-2}}{p^{K-1} \Delta_{\min}^2} \\ & \lesssim \frac{1}{r \log p}, \end{aligned}$$

and

$$L(\hat{z}) \lesssim \frac{\Delta_{\min}^2}{r \log p}.$$

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  $a \in [r]$

$$\begin{aligned} & (1 - O(1)) \frac{p^2}{r^2} \|\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}}\|_F^2 \\ & \leq \sum_{\hat{z}(i)=z(i)=a} (\theta(i) \mathbf{W}_{:a}^T \mathbf{X} - \hat{\theta}(i) \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}})^2 \\ & \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr, \end{aligned}$$

where the first inequality follows from the facts that  $\ell(\hat{z}, z) \lesssim \frac{1}{\log p}, |z^{-1}(a)| \asymp p/r$ ,

$$\begin{aligned} & |z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq |\hat{z}^{-1}(a)| \leq |z^{-1}(a)| + C \frac{p}{r} \ell(\hat{z}, z), \\ & |z^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{z(i)=\hat{z}(i)=a} \theta(i) \leq |z^{-1}(a)|, \end{aligned}$$

and

$$|\hat{z}^{-1}(a)| - C \frac{p}{r} \ell(\hat{z}, z) \leq \sum_{\hat{z}(i)=z(i)=a} \hat{\theta}(i) \leq |\hat{z}^{-1}(a)|.$$

Hence, for all  $i \in [p]$

$$\begin{aligned} & (\theta(i) - \hat{\theta}(i))^2 \|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - O(p) \\ & \leq \|(\theta(i) - \hat{\theta}(i)) \mathbf{W}_{:a}^T \mathbf{X}\|_F^2 - \|\hat{\theta}(i)(\mathbf{W}_{:a}^T \mathbf{X} - \hat{\mathbf{W}}_{:a}^T \hat{\mathbf{X}})\|_F^2 \\ & \leq \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leq pr, \end{aligned}$$

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  $a \in [r]$

$$\|\mathbf{W}_{:a}^T \mathbf{X}\|_F^2 \geq \|\mathbf{S}_{:a}\|_F^2 \lambda_{\min}^{2(K-1)} (\Theta \mathbf{M}) \gtrsim p^{K-1}.$$

The inequality indicates that  $\theta(i)^2 = (1 + o(p^{-(K-2)}))\hat{\theta}(i)^2$ .  $\square$

## ACKNOWLEDGMENT

The authors would like to thank Zheng Tracy Ke, Anru Zhang, Rungang Han, and Yuetian Luo for helpful discussions and for sharing software packages.

## REFERENCES

- [1] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, “Tensor decompositions for learning latent variable models,” *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, Aug. 2014.
- [2] L. Wang, D. Durante, R. E. Jung, and D. B. Dunson, “Bayesian network-response regression,” *Bioinformatics*, vol. 33, no. 12, pp. 1859–1866, 2017.
- [3] P. Koniusz and A. Cherian, “Sparse coding for third-order supersymmetric tensor descriptors with application to texture recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5395–5403.
- [4] M. Wang, J. Fischer, and Y. S. Song, “Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition,” *Ann. Appl. Statist.*, vol. 13, no. 2, pp. 1103–1127, Jun. 2019.
- [5] V. Hore et al., “Tensor decomposition for multiple-tissue gene expression experiments,” *Nature Genet.*, vol. 48, no. 9, p. 1094, 2016.
- [6] D. Ghoshdastidar and A. Dukkipati, “Uniform hypergraph partitioning: Provable tensor methods and sampling techniques,” *J. Mach. Learn. Res.*, vol. 18, no. 50, pp. 1–41, 2017.
- [7] D. Ghoshdastidar and A. Dukkipati, “Consistency of spectral hypergraph partitioning under planted partition model,” *Ann. Statist.*, vol. 45, no. 1, pp. 289–315, 2017.
- [8] K. Ahn, K. Lee, and C. Suh, “Community recovery in hypergraphs,” *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6561–6579, Oct. 2019.
- [9] Z. T. Ke, F. Shi, and D. Xia, “Community detection for hypergraph networks via regularized tensor power iteration,” 2019, *arXiv:1909.06503*.
- [10] M. Wang and Y. Zeng, “Multiway clustering via tensor block models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 715–725.
- [11] E. Abbe, “Community detection and stochastic block models: Recent developments,” *J. Mach. Learn. Res.*, vol. 18, pp. 1–86, Jan. 2018.
- [12] E. C. Chi, B. J. Gaines, W. W. Sun, H. Zhou, and J. Yang, “Provable convex co-clustering of tensors,” *J. Mach. Learn. Res.*, vol. 21, no. 214, pp. 1–58, 2020.
- [13] R. Han, Y. Luo, M. Wang, and A. R. Zhang, “Exact clustering in tensor block model: Statistical optimality and computational limit,” *J. Roy. Stat. Soc., Ser. B, Stat. Methodol.*, vol. 84, no. 5, pp. 1666–1698, Nov. 2022.

- 2992 [14] P. J. Bickel and A. Chen, "A nonparametric view of network models and  
2993 Newman–Girvan and other modularities," *Proc. Nat. Acad. Sci. USA*,  
2994 vol. 106, no. 50, pp. 21068–21073, 2009.
- 2995 [15] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Community detection in degree-corrected block models," *Ann. Statist.*, vol. 46, no. 5,  
2996 pp. 2153–2185, Oct. 2018.
- 2998 [16] K. Ahn, K. Lee, and C. Suh, "Hypergraph spectral clustering in the  
2999 weighted stochastic block model," *IEEE J. Sel. Topics Signal Process.*,  
3000 vol. 12, no. 5, pp. 959–974, Oct. 2018.
- 3001 [17] M. Yuan, R. Liu, Y. Feng, and Z. Shang, "Testing community structure  
3002 for hypergraphs," *Ann. Statist.*, vol. 50, no. 1, pp. 147–169, Feb. 2022.
- 3003 [18] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-  
3004 rank matrix factorization: An overview," *IEEE Trans. Signal Process.*,  
3005 vol. 67, no. 20, pp. 5239–5269, Oct. 2019.
- 3006 [19] S.-Y. Yun and A. Proutiere, "Optimal cluster recovery in the labeled  
3007 stochastic block model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29,  
3008 2016, pp. 973–981.
- 3009 [20] C. Kim, A. S. Bandeira, and M. X. Goemans, "Stochastic block  
3010 model for hypergraphs: Statistical limits and a semidefinite programming  
3011 approach," 2018, *arXiv:1807.02884*.
- 3012 [21] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear  
3013 singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21,  
3014 no. 14, pp. 1253–1278, 2006.
- 3015 [22] Z. Zhang, G. I. Allen, H. Zhu, and D. Dunson, "Tensor network  
3016 factorizations: Relationships between brain structural connectomes and  
3017 traits," *NeuroImage*, vol. 197, pp. 330–343, Aug. 2019.
- 3018 [23] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational  
3019 limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338,  
3020 Nov. 2018.
- 3021 [24] M. Brennan and G. Bresler, "Reducibility and statistical-computational  
3022 gaps from secret leakage," in *Proc. 33rd Conf. Learn. Theory*, vol. 125,  
3023 2020, pp. 648–847.
- 3024 [25] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of  
Lloyd's algorithm and its variants," 2016, *arXiv:1612.02099*.
- 3025 [26] E. Abbe, J. Fan, K. Wang, and Y. Zhong, "Entrywise eigenvector analysis  
3026 of random matrices with low expected rank," *Ann. Statist.*, vol. 48, no. 3,  
3027 pp. 1452–1474, Jun. 2020.
- 3028 [27] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower  
3029 bounds and improved relaxations for tensor recovery," in *Proc. 31th Int.  
3030 Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. 73–81.
- 3031 [28] L. Floreescu and W. Perkins, "Spectral thresholds in the bipartite stochastic  
3032 block model," in *Proc. 29th Conf. Learn. Theory*, vol. 49, 2016,  
3033 pp. 943–959.
- 3034 [29] C. Gao and A. Y. Zhang, "Iterative algorithm for discrete structure  
3035 recovery," *Ann. Statist.*, vol. 50, no. 2, pp. 1066–1094, Apr. 2022.
- [30] I. E. Chien, C.-Y. Lin, and I.-H. Wang, "On the minimax misclassification  
3037 ratio of hypergraph community detection," *IEEE Trans. Inf. Theory*,  
3038 vol. 65, no. 12, pp. 8095–8118, Dec. 2019.
- [31] M. Meilă, "Local equivalences of distances between clusterings—A  
3039 geometric perspective," *Mach. Learn.*, vol. 86, no. 3, pp. 369–389,  
3040 Mar. 2012.
- [32] D. C. Van Essen et al., "The WU-minn human connectome project:  
An overview," *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.
- [33] R. S. Desikan et al., "An automated labeling system for subdividing  
3041 the human cerebral cortex on MRI scans into gyral based regions of  
3042 interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, Jul. 2006.
- [34] J. Hu, C. Lee, and M. Wang, "Generalized tensor decomposition with  
3043 features on multiple modes," *J. Comput. Graph. Statist.*, vol. 31, no. 1,  
3044 pp. 204–218, Jan. 2022.
- [35] S. H. Lee, J. M. Magallanes, and M. A. Porter, "Time-dependent com-  
3045 munity structure in legislation cosponsorship networks in the congress  
3046 of the republic of Peru," *J. Complex Netw.*, vol. 5, no. 1, pp. 127–144,  
3047 2017.
- [36] P. Rigollet and J.-C. Hütter, "High dimensional statistics," *Lect. Notes  
Course 18S997*, vol. 813, no. 814, p. 46, 2015.
- [37] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and  
3048 computational framework for generalized tensor estimation," *Ann. Statist.*,  
3049 vol. 50, no. 1, pp. 1–29, Feb. 2022.

**Jiaxin Hu** received the B.S. degree from Wuhan University in 2019 and the  
3060 master's degree in statistics from the University of Wisconsin–Madison in  
3061 2020, where she is currently pursuing the Ph.D. degree with the Department  
3062 of Statistics. Her research interests include tensor methods, network analysis,  
3063 and applications in genetics and social science.

**Miaoyan Wang** received the B.S. degree in mathematics from Fudan Uni-  
3065 versity in 2010 and the Ph.D. degree in statistics from the University of  
3066 Chicago in 2015. She is currently an Assistant Professor of statistics at  
3067 the University of Wisconsin–Madison. She is also a Faculty Affiliate with  
3068 the Mathematical Foundations of Machine Learning and the Institute for  
3069 Foundations of Data Science. Her research interests include the intersection  
3070 of statistics, machine learning, and genetics. She was a recipient of NSF  
3071 CAREER Award in 2022, Best Student Paper Awards (with her as an Advisor)  
3072 from American Statistical Association (three times in 2021, 2022, and 2023),  
3073 and New England Statistical Society in 2022.