Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu and Miaoyan Wang

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 datasets.

Index Terms

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

I. Introduction

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

This paper studies the problem of multiway clustering based on a data tensor. The goal of multiway clustering is to identify a checkerboard structure from a noisy data tensor. Figure 1 illustrates the noisy tensor and the underlying structures discovered by multiway clustering methods. The checkerboard structure serves as a meta tool to many popular structures including the low-rankness (Young et al., 2018), latent space models (Wang and Li, 2020), and isotonic models (Pananjady and Samworth, 2020). In the hypergraph example, the multiway clustering aims to identify the underlying block partition of nodes based on their higher-order connectivities; therefore, we also refer to the clustering as *higher-order clustering*. The most common model for higher-order clustering is called *tensor block model* (TBM) (Wang and Zeng, 2019), which extends the usual matrix stochastic block model (Abbe, 2017) to tensors. The matrix analysis tools, however, are sub-optimal for higher-order clustering. Developing tensor tools for solving block models has received increased interest recently (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019).

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 effects apart from the block effects. However, the exchangeability assumption is often non-realistic. Each node may contribute to the data variation by its own multiplicative effect. Such degree heterogeneity appears commonly in social networks. Ignoring the individual heterogeneity may seriously mislead the clustering results.

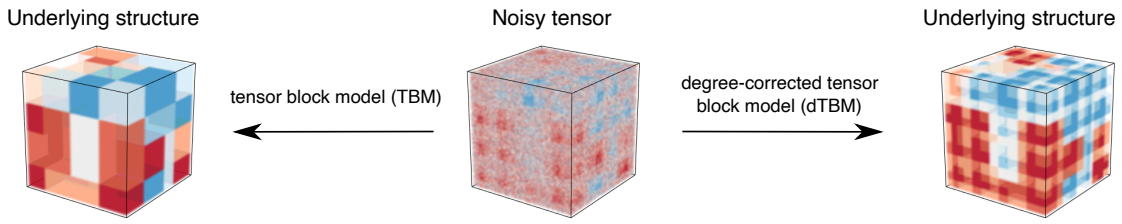

Fig. 1: Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

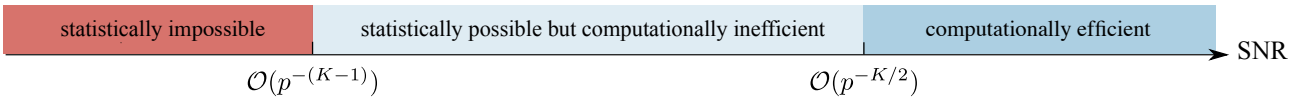

Fig. 2: SNR thresholds for statistical and computational limits in order-K dTBM with dimension $(p, ..., p)$ and $K \geq 2$. The SNR gap between statistical possibility and computational efficiency exists only for tensors with $K \geq 3$.

For example, regular block model fails to model the member affiliation in the Karate Club network (Bickel and Chen, 2009) without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity (Ke et al., 2019). The dTBM combines a higher-order checkerboard structure with degree parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}(1), \ldots, \boldsymbol{\theta}(p))^T$ to allow heterogeneity among p nodes. Figure 1 compares the underlying structures of TBM and dTBM with the same number of communities. The dTBM allows varying values within the same community, thereby allowing a richer structure. To solve dTBM, we project clustering objects to a unit sphere and consider the angle similarity; detailed algorithms are in Section IV. On one hand, the *spherical* clustering avoids the estimation of nuisance degree heterogeneity. On the other hand, the usage of angle similarity brings new challenges to the derivations of theoretical results, and the polar coordinates based techniques are equipped in the proof.

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

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

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

Related work. Our work is closely related to, but also distinct from several lines of existing research.

- *Block model for clustering.* Block models such as stochastic block model (SBM) and degree-corrected SBM has been widely used for matrix clustering problems. The theoretical properties and algorithm performance for block models have been well-studied; see the review paper (Abbe, 2017) and the references therein. However, The tensor counterparts are relatively less understood. Table I summarizes the most relevant models. Specifically,
  - the tensor block model (TBM, Han et al. (2020); Wang and Zeng (2019)), a higher-order extension of SBM, fails to allow degree heterogeneity. The Cartesian coordinates based analysis in Han et al. (2020) is also

	Gao et al. (2018)	Han et al. (2020)	Ghoshdastidar et al. (2017)	Ke et al. (2019)	**Ours**
Allow tensors	$\times$	$\surd$	$\surd$	$\surd$	$\surd$
Allow heterogeneity	$\surd$	$\times$	$\surd$	$\surd$	$\surd$
Eigen free clustering	$\surd$	$\surd$	$\times$	$\times$	$\surd$
Misclustering rate (for order K^*)	-	$\exp(-p^{K/2})$	p^{-1}	p^{-2}	$\exp(-p^{K/2})$

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

non-applicable to handle the extra flexibility brought from heterogeneity. In contrast, our model addresses the degree heterogeneity, and the polar coordinates based tools are adapted for the theoretical analysis;

- the hypergraph degree-corrected block model (hDCBM) proposed by Ke et al. (2019); Yuan et al. (2018) accounts for degree heterogeneity. Whereas, the hDCBM is designed only for binary observations, and the proposed spectral algorithm in Ke et al. (2019) achieves a sub-optimal clustering accuracy in higher-order scenarios. In constrast, our model allows discrete and continuous entries, and achieves exponentially fast rate in clustering tasks. More importantly, to our best knowledge, we are the first to provide the statistical and computational limits analyses for the degree-corrected block model in tensor clustering.

- *Global-to-local algorithm strategy.* Our methods generalize the recent global-to-local strategy for matrix learning (Chi et al., 2019; Gao et al., 2018; Yun and Proutiere, 2016) to tensors (Ahn et al., 2018; Han et al., 2020; Kim et al., 2018). Despite the conceptual similarity, we address several fundamental challenges associated with this non-convex, non-continuous problem. We show the insufficiency of the conventional tensor HOSVD (Kolda and Bader, 2009), and we develop a weighted higher-order initialization that relaxes the eigen-gap separation condition. Furthermore, our local iteration leverages the angle-based clustering in order to avoid explicit estimation of degree heterogeneity. Our bounds reveal the interesting interplay between the computational and statistical errors. We show that our final estimate provably achieves the exact clustering within only polynomial-time complexity.

Notation. We use lower-case letters (e.g., a, b) for scalars, lower-case boldface letters (e.g., $\boldsymbol{a}, \boldsymbol{\theta}$) for vectors, upper-case boldface letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}$) for matrices, and calligraphy letters (e.g., $\mathcal{X}, \mathcal{Y}$) for tensors of order three or greater. We use $|\cdot|$ for the cardinality of a set and $\mathbb{1}\{\cdot\}$ for the indicator function. For an integer $p \in \mathbb{N}_+$, we use the shorthand $[p] = \{1, 2, ..., p\}$. For a length-p vector $\boldsymbol{a}$, we use $a(i) \in \mathbb{R}$ to denote the i-th entry of $\boldsymbol{a}$, and use $\boldsymbol{a}_I$ to denote the sub-vector by restricting the indices in the set $I \subset [p]$. We use $\|\boldsymbol{a}\| = \sqrt{\sum_i a^2(i)}$ to denote the ℓ_2-norm, $\|\boldsymbol{a}\|_1 = \sum_i |a_i|$ to denote the ℓ_1 norm of $\boldsymbol{a}$. For two vector $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension, we denote the angle between $\boldsymbol{a}, \boldsymbol{b}$ by

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

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

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

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K,$$

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

$$\begin{aligned} &\mathcal{X}(i_1, \ldots, i_K) \\ &= \sum_{(j_1, \ldots, j_K)} \mathcal{S}(j_1, \ldots, j_K) \boldsymbol{M}_1(i_1, j_1) \cdots \boldsymbol{M}_K(i_K, j_K). \end{aligned}$$

We use $\mathrm{Ave}(\cdot)$ to denote the operation of taking averages across elements and $\mathrm{Mat}_k(\cdot)$ to denote the unfolding operation that reshapes the tensor along mode k into a matrix. For a symmetric tensor $\mathcal{Y} \in \mathbb{R}^{p \times \cdots \times p}$, we omit the

subscript and use $\mathrm{Mat}(\mathcal{Y}) \in \mathbb{R}^{p\times p^{K-1}}$ to denote the unfolding. For two sequences $\{a_p\}, \{b_p\}$, we denote $a_p \lesssim b_p$ if $\lim_{p\to\infty} a_p/b_p \le c$ and $a_p = \Omega(b_p)$ if $cb_p \le a_p \le Cb_p$, for some constants $c, C \ge 0$. Throughout the paper, we use the terms "community" and "clusters" exchangeably.

Organization. The rest of this paper is organized as follows. Section II introduces the degree-corrected tensor block model (dTBM) with three motivating examples and presents the identifiability of dTBM under the angle gap condition. We show the phase transition and the existence of statistical-computational gaps for the higher-order dTBM in Section III. In Section IV, we provide a polynomial-time two-stage algorithm with misclustering rate guarantees. Numerical studies including the simulations to assess the theoretical results, comparison with other methods, and real data analysis on human brain connectome data and Peru legislation data are in Section V. Last, we conclude our paper in Section VI.

II. Model formulation

A. Degree-corrected tensor block model

Suppose we have an order-K data tensor $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$. For ease of notation, we focus on symmetric tensors in this section; our framework easily extends to general asymmetric tensors. Assume there exist $r \ge 2$ disjoint communities among the p nodes. We represent the community assignment by a function $z\colon [p] \mapsto [r]$, where $z(i) = a$ for i-th node that belongs to the a-th community. Then, $z^{-1}(a) = \{i \in [p]\colon z(i) = a\}$ denotes the set of nodes that belong to the a-th community, and $|z^{-1}(a)|$ denotes the number of nodes in the a-th community. Let $\boldsymbol{\theta} = (\theta(1), \ldots, \theta(p))^T$ denote the degree heterogeneity for p nodes. We consider the order-K dTBM (Ghoshdastidar et al., 2017; Ke et al., 2019),

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

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

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \cdots \times_K \boldsymbol{\Theta}\boldsymbol{M}, \tag{1}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\theta(1), ..., \theta(p)) \in \mathbb{R}^{p\times p}$ is a diagonal matrix, $\boldsymbol{M} \in \{0,1\}^{p\times r}$ is the membership matrix associated with community assignment z such that $\boldsymbol{M}(i,j) = \mathbb{1}\{z(i) = j\}$. By definition, each row of $\boldsymbol{M}$ has one copy of 1's and 0's elsewhere. Note that the discrete nature of $\boldsymbol{M}$ renders our model (1) more challenging than Tucker decomposition. We call a tensor $\mathcal{X}$ an r-block tensor with degree $\boldsymbol{\theta}$ if $\mathcal{X}$ admits dTBM (1). Here, we give two special cases of dTBM.

B. Motivating Examples.

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

a) 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 (Chi et al., 2020; Han et al., 2020; Wang and Zeng, 2019). The theoretical results in TBM serve as benchmarks for dTBM.

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

Similar with the adjacency matrix, it is natural to represent the K-uniform hypergraph by a binary order-K adjacency tensor. Let $\mathcal{Y} \in \{0,1\}^{p\times\cdots\times p}$ denote the adjacency tensor, where the entries encode the presence or absence of

hyperdeges among p nodes. Specifically,

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

for all $(i_1,\ldots,i_K) \in [p]^K$.

Assume there exist r disjoint communities among p nodes. The equation (1) models $\mathbb{E}\mathcal{Y}$ with unknown degree heterogeneity $\boldsymbol{\theta}$ and subgaussianity parameter $\sigma^2 = 1/4$.

c) Multi-layer weighted network: Multi-layer weighted network data consists of multiple networks with the same set of nodes. One representative example is the brain structural connectome data (Zhang et al., 2019). The multi-layer weighted network $\mathcal{Y}$ has dimension of $p \times p \times L$, where p denotes the number of brain regions in interest, and L denotes the number of layer. Each of the L networks describes one aspect of the brain connections, and the tensor $\mathcal{Y}$ include a mixture slices with continuous, binary, and count entries.

Assume there exist r disjoint communities among p nodes and r_l disjoint communities among the L layers. The multi-layer network community detection is modeled by the generalized asymmetric dTBM model (1)

$$\mathbb{E}[\mathcal{Y}] = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \boldsymbol{\Theta}\boldsymbol{M} \times_3 \boldsymbol{\Theta}_l\boldsymbol{M}_l,$$

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

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

Assume there exist r_1, r_2, r_3 disjoint clusters for p genes, n individuals, and t tissues, respectively. We apply the generalized asymmetric dTBM model (1)

$$\mathbb{E}[\mathcal{Y}] = \mathcal{S} \times_1 \boldsymbol{\Theta}_1\boldsymbol{M}_1 \times_2 \boldsymbol{\Theta}_2\boldsymbol{M}_2 \times_3 \boldsymbol{\Theta}_3\boldsymbol{M}_3,$$

where $\{(\boldsymbol{\theta}_k, \boldsymbol{M}_k)\}_{k=1}^3$ refers to the heterogeneity and membership for genes, individuals, and tissues.

Remark 1 (Comparison with non-degree models)**.** Our dTBM uses fewer block parameters than TBM. Let the subscripts "deg" and "non" denote quantities in the models with and without degrees, respectively. Then, every r_{non}-block tensor can be represented by a degree-corrected r_{deg}-block tensor with $r_{\text{deg}} \le r_{\text{non}}$. In particular, there exist tensors with $r_{\text{non}} = p$ but $r_{\text{deg}} = 1$, so the reduction in r 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). We use $\mathcal{P}$ to denote the following parameter space for $(z, \mathcal{S}, \boldsymbol{\theta})$,

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

where $c_i > 0$'s are universal constants. First, the entrywise positivity constraint on $\boldsymbol{\theta} \in \mathbb{R}_+^p$ is imposed to avoid sign ambiguity between entries in $\boldsymbol{\theta}_{z^{-1}(a)}$ and $\mathcal{S}$ and thereof allow the trigonometric cos to describe the angle similarity in the following Assumption 1 and Sub-algorithm 2 in Section IV. This constraint can be achieved without sacrificing model flexibility, by using a slightly larger dimension of $\mathcal{S}$ in the factorization (1); see Example 1. Second, the constants c_1, c_2 in the $|z^{-1}(a)|$ bound assume the roughly balanced size across r communities. Third, the constants c_3, c_4 in the magnitude of $\text{Mat}(\mathcal{S})_{a:}$ requires no purely zero slide in $\mathcal{S}$, so the core tensor $\mathcal{S}$ is not trivially reduced to lower rank. Lastly, the ℓ_1 normalization $\|\boldsymbol{\theta}_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$ is imposed to avoid the scalar ambiguity between $\boldsymbol{\theta}_{z^{-1}(a)}$ and $\mathcal{S}$. This constraint, again, incurs no restriction to model flexibility but makes our presentation cleaner.

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

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

where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$ and $\boldsymbol{M} = (1, 1, 1, 1)^T$. Note that $\mathcal{X} \in \mathbb{R}^{4\times4\times4}$ is a 1-block tensor with mixed-signed degree $\boldsymbol{\theta}$, and the mode-3 slices of $\mathcal{X}$ are

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

Now, instead of original decomposition, we encode $\mathcal{S}$ as a 2-block tensor with positive-signed degree

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

where $\mathcal{S}' \in \mathbb{R}^{2\times2\times2}$ have mode-3 slices

$$\mathcal{S}'_{::1} = -\mathcal{S}'_{::2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{M}' = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_2 \end{bmatrix},$$

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

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

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

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

and set the convention $\Delta_{\min} = 1$ for $r = 1$. Equivalently, none of the two rows in $\boldsymbol{S}$ are parallel, i.e., when $r \geq 2$, $\max_{a\neq b\in[r]} \cos(\boldsymbol{S}_{a:}, \boldsymbol{S}_{b:}) = 1 - \Delta^2_{\min}/2 < 1$.

The quantity $\Delta_{\min}$ characterizes the non-redundancy among clusters measured by angle separation. The definition (3) is well posed because of the lower bound on $\|\boldsymbol{S}_{a:}\|$ in (2). The following theorem shows that the angle gap separation is sufficient and necessary for parameter identifiability under dTBM.

Theorem 1 (Identifiability)**.** Consider the dTBM with $r \geq 2$. 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 more appealing than classical Tucker model. In the Tucker model, the factor matrix $\boldsymbol{M}$ 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 $\boldsymbol{M}$ can be precisely recovered under our algorithm. This property benefits the interpretation of dTBM in practice.

III. Statistical-computational gaps for higher-order tensors

In this section, we study the statistical and computational limits of dTBM. We reserve the term higher-order tensors for tensors of order $K \geq 3$. We propose signal-to-noise ratio (SNR),

$$\mathrm{SNR} := \Delta^2_{\min}/\sigma^2 = p^\gamma, \tag{4}$$

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

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

Note that 1-block dTBM does not belong to the space $\mathcal{P}(\gamma)$ when $\gamma < 0$ by Assumption 1. Our goal is to characterize the clustering accuracy with respect to γ when $r \geq 2$. Let $\hat{z}$ and z be estimated and true clustering functions in the family (2). Define the misclustering error by

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

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

In Sections III-A and III-B, we provide the lower bounds of $\ell(\hat{z}, z)$ for general Gaussian dTBMs without symmetric assumptions. For general (asymmetric) dTBMs, 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 lower bounds should be interpreted as the worst-case results across K modes.

A. Statistical critical values

Our first main result is to show the minimax lower bound of SNR for exact recovery in dTBM.

Theorem 2 (Statistical lower bound)**.** Consider general Gaussian dTBMs under the parameter space $\mathcal{P}(\gamma)$ with $K \geq 1$. Assume $r \lesssim p^{1/3}$. If the signal exponent satisfies $\gamma < -(K-1)$, then, every estimator $\hat{z}_{\text{stat}}$ obeys

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

Theorem 2 demonstrates the impossibility of exact recovery of the assignment when $\gamma < -(K-1)$ in the high-dimensional regime $p \to \infty$ for fixed r. The proof is information-theoretical, and therefore the results apply to all statistical estimators, including but not limited to, maximum likelihood estimation (MLE) (Wang and Zeng, 2019) and trace maximization (Ghoshdastidar and Dukkipati, 2017). Our derived SNR threshold $-(K-1)$ is also a minimax upper bound, because MLE achieves exact recovery when $\gamma > -(K-1)$. Hence, the boundary $\gamma_{\text{stat}} \coloneqq -(K-1)$ is the critical value for statistical performance of dTBM.

B. Computational critical values

An important ingredient to establish the computational limits is the *hypergraphic planted clique (HPC) conjecture* (Brennan and Bresler, 2020; Zhang and Xia, 2018). The HPC conjecture indicates the impossibility of fully recovering the planted cliques with polynomial-time algorithm when the clique size is less than the number of vertices in the hypergraph. The formal statement of HPC detection and conjecture can be found in Definition 1 and Conjecture 1 as following.

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

The order-K HPC detection aims to identify whether there exists a planted clique hidden in a Erdős-Rényi K-hypergraph. We formulate HPC detection as the following hypothesis testing problem

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

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

$$\liminf_{p\to\infty} \mathbb{P}_{H_0}\left(\varphi_p(H)=1\right) + \mathbb{P}_{H_1}\left(\varphi_p(H)=0\right) \ge \frac{1}{2}.$$

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

Theorem 3 (Computational lower bound)**.** Consider general Gaussian dTBMs under the parameter space $\mathcal{P}(\gamma)$ with $K \ge 2$. Assume HPC conjecture holds. If the signal exponent $\gamma < -K/2$, then, every *polynomial-time estimator* $\hat{z}_{\text{comp}}$ obeys

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

Theorem 3 indicates the impossibility of exact recovery by polynomial-time algorithms when $\gamma < -K/2$. Therefore, $\gamma_{\text{comp}} := -K/2$ is the critical value for computational performance of dTBM. In Section IV, we will show the condition $\gamma > -K/2$ suffices for our proposed polynomial-time estimator. Thus, $\gamma_{\text{comp}} := -K/2$ is the critical value for computational performance of dTBM.

Remark 2 (Statistical-computational gaps)**.** Now, we have established the phase transition of exact clustering under order-K dTBM combing Theorems 2 and 3. Figure 2 summarizes our results of critical SNRs when $K \ge 2$. 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

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

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

Remark 3 (Comparison with non-degree models)**.** We compare our results to non-degree tensor models. The allowance of degree heterogeneity $\boldsymbol{\theta}$ makes the model more flexible, but it incurs extra statistical and computational complexity. Fortunately, we find that the extra complexity does not render the estimation of z qualitatively harder; see the comparison of our phase transition with non-degree TBM (Han et al., 2020).

IV. Polynomial-time algorithm under mild SNR

We present a two-stage clustering algorithm. The procedure takes a global-to-local approach. See Figure 3 for illustration. The global step finds the basin of attraction with polynomial miclustering error, whereas the local iterations improve the initial clustering to exact recovery. Both steps are critical to obtain a satisfactory algorithm output.In this section, we first use the symmetric tensor as a working example to describe the algorithm procedures to gain insight. Our theoretical analysis focuses on the noisy tensor with i.i.d. sub-Gaussian noise such as Gaussian and uniform observations. The extensions for asymmetric tensor and Bernoulli observation and other practical issues are in subsection IV-C.

A. Weighted higher-order initialization

We start with weighted higher-order clustering algorithm as initialization. We take the order-3 tensor in illustration for simplicity. Consider noiseless case with $\mathcal{X} = \mathbb{E}\mathcal{Y}$ and $\boldsymbol{X} = \text{Mat}(\mathcal{X})$. By model (1), for all $i \in [p]$, we have

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

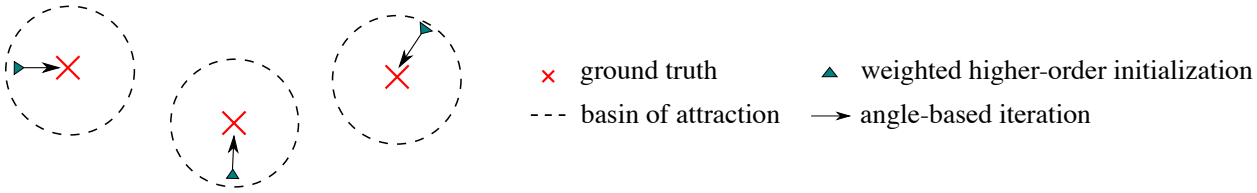

Fig. 3: Illustration of our global-to-local algorithm.

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

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

$$\hat{\boldsymbol{U}} = \text{SVD}_r\left(\text{Mat}\left(\mathcal{Y}\times_1 \boldsymbol{U}_{\text{pre}}^T \times_2 \boldsymbol{U}_{\text{pre}}^T\right)\right).$$

The final denoised tensor $\hat{\mathcal{X}}$ is defined by

$$\hat{\mathcal{X}} = \mathcal{Y}\times_1 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_2 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_3 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T.$$

The double projection improves usual matrix spectral methods in order to alleviate the noise tensor.

The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted k-means clustering. We write $\hat{\boldsymbol{X}} = \text{Mat}(\hat{\mathcal{X}})$, and normalize the rows into $\hat{\boldsymbol{X}}^s_{i:} = \|\hat{\boldsymbol{X}}_{i:}\|^{-1}\hat{\boldsymbol{X}}_{i:}$ as a surrogate of $\theta(i)^{-1}\boldsymbol{X}_{i:}$. Then, a weighted k-means clustering is performed on the normalized rows with weights equal to $\|\hat{\boldsymbol{X}}_{i:}\|^2$. The choice of weights is to bound the k-means objective function by the Frobenius-norm accuracy of $\hat{\mathcal{X}}$. Unlike existing clustering algorithm (Ke et al., 2019), we apply the clustering on the unfolded tensor $\hat{\boldsymbol{X}}$ rather than on the factors $\hat{\boldsymbol{U}}$. This strategy relaxes the eigen-gap separation condition (Gao et al., 2018; Han et al., 2020). We assign degenerate rows with purely zero entries to an arbitrarily random cluster; these nodes are negligible in high-dimensions because of the lower bound on $\|\text{Mat}(\mathcal{S})_{a:}\|$ in (2). The final result gives the initial clustering assignment $\hat{z}^{(0)}$. Full procedures are provided in Sub-algorithm 1.

We now establish the misclustering error rate of initialization. We call $\boldsymbol{\theta}$ is balanced, if the relative extent of heterogeneity is comparable across clusters in that

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

Note that, the assumption (6) does not preclude degree heterogeneity. Indeed, within each of the clusters, the highest degree can be $\theta(i) = \Omega(p)$, whereas the lowest degree can be $\theta(i) = \mathcal{O}(1)$.

Theorem 4 (Error for weighted higher-order initialization)**.** Consider the general sub-Gaussian dTBM with i.i.d. noise under the parameter space $\mathcal{P}$ and Assumption 1. Assume $\boldsymbol{\theta}$ is balanced and $\min_{i\in[p]}\theta(i) \geq c$ for some constant $c > 0$. Let $z^{(0)}$ denote the output of Sub-algorithm 1. With probability going to 1, we have

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}. \tag{7}$$

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

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

Sub-algorithm 1: Weighted higher-order initialization

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

1: Compute factor matrix $\boldsymbol{U}_{\text{pre}} = \text{SVD}_r(\text{Mat}(\mathcal{Y}))$ and the $(K-1)$-mode projection $\mathcal{X}_{\text{pre}} = \mathcal{Y} \times_1 \boldsymbol{U}_{\text{pre}}^T \times_2 \cdots \times_{K-1} \boldsymbol{U}_{\text{pre}}^T$.

2: Compute factor matrix $\hat{\boldsymbol{U}} = \text{SVD}_r(\text{Mat}(\mathcal{X}_{\text{pre}}))$ and denoised tensor $\hat{\mathcal{X}} = \mathcal{Y} \times_1 \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T \times_2 \cdots \times_K \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T$.

3: Let $\hat{\boldsymbol{X}} = \text{Mat}(\hat{\mathcal{X}})$ and $I_0 = \{i \in [p] : \|\hat{\boldsymbol{X}}_{i:}\| = 0\}$. Set $\hat{z}(i)$ randomly in $[r]$ for $i \in I_0$.

4: For all $i \in I_0^c$, compute normalized rows $\hat{\boldsymbol{X}}_{i:}^s := \|\hat{\boldsymbol{X}}_{i:}\|^{-1}\hat{\boldsymbol{X}}_{i:}$.

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

$$\sum_{i\in I_0^c} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{\hat{z}(i)}\|^2 \le \eta \min_{\bar{\boldsymbol{x}}_j, j\in[r], \bar{z}(i), i\in I_0^c} \sum_{i\in S^c} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}_{i:}^s - \bar{\boldsymbol{x}}_{\bar{z}(i)}\|^2.$$

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

Sub-algorithm 2: Angle-based iteration

Input: Observation $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$, initialization $z^{(0)}\colon [p] \to [r]$ from Sub-algorithm 1, iteration number T.

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

7: Update the block tensor $\mathcal{S}^{(t)}$ via $\mathcal{S}^{(t)}(i_1, ..., i_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z^{(t)}(i_k) = j_k, k \in [K]\}$.

8: Calculate reduced tensor $\mathcal{Y}^{\text{d}} \in \mathbb{R}^{p\times r\times\cdots\times r}$ via

$$\mathcal{Y}^{\text{d}}(i, a_2, \ldots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \ldots, i_K) : z^{(t)}(i_k) = a_k, k \ne 1\}.$$

9: Let $\boldsymbol{Y}^{\text{d}} = \text{Mat}(\mathcal{Y}^{\text{d}})$ and $J_0 = \{i \in [p] : \|\boldsymbol{Y}_{i:}^{\text{d}}\| = 0\}$. Set $z^{(t+1)}(i)$ randomly in $[r]$ for $i \in J_0$.

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

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

11: **end for**

Output: Estimated clustering $z^{(T)} \in [r]^p$.

Remark 5 (Failure of conventional tensor HOSVD)**.** If we use conventional HOSVD for tensor denoising; that is, we use $\boldsymbol{U}_{\text{pre}}$ in place of $\hat{\boldsymbol{U}}$ in line 2, then the misclustering rate becomes $\mathcal{O}(p)$ for all $K \ge 2$. This rate is substantially worse than our current rate (7).

Remark 6 (Eigen free clustering)**.** Note that our initialization works on the estimated mean tensor $\hat{\mathcal{X}}$ directly rather than the leading tensor decomposition factors of $\mathcal{X}$. On one hand, clustering on $\hat{\mathcal{X}}$ avoids the eigen-gap assumption in previous work Ke et al. (2019). On the other hand, vanilla spectral methods working on decomposition factors suffers the non-identifiability of the eigenspace with orthogonal transformation when the number of blocks $r \ge 3$. Such ambiguity comes from the possible multiplicity of eigenvalue and causes trouble for effective clustering (Abbe et al., 2020). In contrast, our eigen free strategy working on $\hat{\mathcal{X}}$ avoids such non-identifiability issue regardless the number of blocks.

B. Angle-based iteration

We propose an angle-based local iteration to improve the outputs from Sub-algorithm 1. To gain the intuition, consider an one-dimensional degree-corrected clustering problem with data vectors $\boldsymbol{x}_i = \theta(i)\boldsymbol{s}_{z(i)} + \boldsymbol{\epsilon}_i, i \in [p]$, where $\boldsymbol{s}_i$'s are known cluster centroids, $\theta(i)$'s are unknown positive degrees, and $z\colon [p] \mapsto [r]$ is the clustering assignment of interest. The angle-based k-means algorithm estimates the assignment z by minimizing the angle between data vectors and centroids; i.e.,

$$z(i) = \arg\max_{a\in[r]} \cos(\boldsymbol{x}_i,\ \boldsymbol{s}_a), \quad \text{for all } i \in [p]. \tag{8}$$

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

Our Sub-algorithm 2 shares the same spirit as angle-based k-means. We still take the order-3 tensor for illustration. Specifically, Sub-algorithm 2 updates estimated core tensor and cluster assignment in each iteration. For core tensor, we consider the following update strategy

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

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

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

The row $\text{Mat}(\mathcal{Y}^{\text{d}})_{i:}$ and $\text{Mat}(\mathcal{S}^{(t)})_{a:}$ corresponds to the $\boldsymbol{x}_i$ and $\boldsymbol{s}_a$s in the one-dimensional clustering (8). Then, we obtain the updated assignment as

$$z(i)^{(t+1)} = \underset{a \in [r]}{\arg\min} \sin\left(\boldsymbol{Y}^{\text{d}}_{i:}, \boldsymbol{S}^{(t)}_{a:}\right), \text{ for all } i \in [p],$$

where $z^{(t+1)}(i)$ is randomly assigned for degenerate rows. Full procedures for our angle-based iteration are described in Sub-algorithm 2.

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

Definition 2 (Locally linear stability)**.** Define the ε-neighborhood of z by $\mathcal{N}(z, \epsilon) = \{\bar{z} \colon \ell(\bar{z}, z) \leq \epsilon\}$. We define two cluster-size vectors for $\bar{z}$,

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

We call the degree is ε-locally linearly stable if and only if

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

The local stability holds trivially for $\varepsilon = 0$ based on the construction of parameter space (2). The locally linear stability avoids the concentration of entities with extremely large or small heterogeneity, when a good estimated assignment with a small misclustering error ϵ is given.

Theorem 5 (Error for angle-based iteration)**.** Consider the setup as in Theorem 4. Assume the local linear stability of degree holds in all neighborhoods $\mathcal{N}(z, \epsilon)$ for any $\epsilon \leq \log^{-1} p$. Suppose $r = \mathcal{O}(1)$ and $\text{SNR} \gtrsim p^{-K/2} \log p$. Let $z^{(t)}$ denote the t-th iteration output in Sub-algorithm 2 with initialization $z^{(0)}$ from Sub-algorithm 1. With probability going to 1, there exists a contraction parameter $\rho \in (0, 1)$ such that

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

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 $t \to \infty$.

Theorem 5 implies that, with probability going to 1, our estimate $z^{(T)}$ achieves exact recovery within polynomial iterations; more precisely,

$$z^{(T)} = \pi \circ z, \quad \text{for all } T \gtrsim \log_{1/\rho} p,$$

for some permutation $\pi \in \Pi$. We call our algorithm *computationally efficient* with $\text{SNR} \gtrsim p^{-K/2} \log p$. Note that the minimal SNR requirement, $p^{-K/2} \log p$, coincides with the computational lower bound in Theorem 3 ignoring

the logarithmic term. Therefore, our algorithm is optimal regarding the signal demand and lies in the most right "computationally efficient" regime in Figure 2.

C. Extension and practical issue

Extension for Bernoulli observations. The main difficulty to establish the statistical guarantee for Bernoulli observations lies in the initialization Sub-Algorithm 1. Theorem 5 still holds for Bernoulli observations once the initialization accuracy satisfies the upper bound (7) in Theorem 4.

Specifically, the derivation of Theorem 4 relies on the upper bound of the estimation error for the mean tensor in Lemma 9, i.e., with high probability

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \tag{9}$$

where $\mathcal{X} = \mathbb{E}[\mathcal{Y}]$ and $\hat{\mathcal{X}}$ is defined in Step 2 of Sub-algorithm 1. Unfortunately, the inequality (9) only holds for i.i.d. sub-Gaussian observations while Bernoulli observations are not identically distributed.

One possible remedy is to apply singular value decomposition to the unfolded Bernoulli observation $\mathcal{Y}$. Let the matrix $\text{Mat}_{sq}(\mathcal{Y}) \in \{0,1\}^{\lfloor p^{K/2} \rfloor \times \lceil p^{K/2} \rceil}$ denote the square unfolded binary tensor. We have the estimate

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

Following Lemma 7 in Gao et al. (2018), with high probability, we have

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

Replacing the estimate $\hat{\mathcal{X}}$ by $\hat{\mathcal{X}}'$, the high probability upper bound for Bernoulli initialization is

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-\lfloor K/2 \rfloor}}{\text{SNR}}. \tag{10}$$

The Bernoulli bound (10) is relatively looser than Gaussian bound (7), especially when K is small. A tighter Bernoulli bound will be achieved once a low-rank binary tensor estimation scheme with better accuracy is provided in the future.

Extension for general dTBMs. Our two-stage algorithm is able to be extended for the general (asymmetric) dTBMs. Specifically, in the Sub-Algorithm 1, we make the following changes: (1) Replace the matrcization $\text{Mat}(\mathcal{Y})$ by $\text{Mat}_k(\mathcal{Y})$; (2) Repeat the Steps 1-5 with mode-specified number of cluster r_k; (3) Obtain the collection initialization $\{z_k^{(0)}\}_{k=1}^K$. In the Sub-Algorithm 2, we make the following changes: (1) Take the collection $\{z_k^{(0)}\}_{k=1}^K$ as input, and update the block tensor $\mathcal{S}^{(t)}$ with the collection $\{z_k^{(t)}\}_{k=1}^K$, $\mathcal{S}^{(t)}(i_1, \ldots, i_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z_k^{(t)}(i_k) = j_k, k \in [K]\}$; (2) Calculate reduced tensor $\mathcal{Y}_k^{\text{d}}$ for each mode via

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

(3) Repeat Step 8-10 with $\text{Mat}_k(\cdot), \mathcal{Y}_k^{\text{d}}$ for each $k \in [K]$ and obtain the collection $\{z_k^{(T)}\}_{k=1}^K$.

Correspondingly, Theorems 4 and 5 still hold with $\ell(z_k^{(0)}, z_k)$ and $\ell(z_k^{(t+1)}, z_k)$ for all $k \in [K]$.

Computational Complexity. Our two-stage algorithm has a polynomial computational cost. Specifically, the complexity of Sub-Algorithm 1 is $\mathcal{O}(Kp^{K+1} + Krp^K)$, where the first term is contributed by the first step SVD 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 $\mathcal{Y}^{\text{d}}$ and pr^K comes from the normalization of $\mathcal{Y}^{\text{d}}, \mathcal{S}^{(t)}$ and the cluster assignment update in Step 10.

that comes from the addition operation and comparison in Step 7,8, and 10. Note that the number of iteration is a logarithmic function of p to achieve exact clustering. We conclude that the two-stage algorithm leads to a polynomial complexity in total.

Rank Selection. Note that we assume the number of clusters r is given for the algorithm. We here provide a simple rank selection criteria when r is unknown. With a fixed r, calculate the estimated minimal gap

$$\hat{\Delta}_{\min}(r) = \min_{a \neq b \in [r]} \left\| \frac{\hat{\boldsymbol{S}}_{a:}(r)}{\|\hat{\boldsymbol{S}}_{a:}(r)\|} - \frac{\hat{\boldsymbol{S}}_{b:}(r)}{\|\hat{\boldsymbol{S}}_{b:}(r)\|} \right\|,$$

where $\boldsymbol{S}(r) = \text{Mat}(\hat{\mathcal{S}}(r))$ and $\hat{\mathcal{S}}(r)$ is the estimated signal tensor obtained by the same procedure in Step 7 of Sub-Algorithm 2 with the estimated assignment $\hat{z}(r)$. With a given upper bound for the number of clusters, denoted R, we choose r that maximizes the estimated minimal gap, i.e.,

$$\hat{r} = \arg\max_{r \in [R]} \hat{\Delta}_{\min}(r).$$

V. Numerical studies

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

We generate order-3 tensors with *assortative* (Gao et al., 2018) core tensors to control SNR; i.e., we set $\mathcal{S}_{aaa} = s_1$ for $a \in [r]$ and others be s_2, where $s_1 > s_2 > 0$. Let $\alpha = s_1/s_2$. We set α close to 1 such that $1 - \alpha = o(p)$. In particular, we have $\alpha = 1 + \Omega(p^{\gamma/2})$ by Assumption 1 and definition (4). Hence, we easily adjust SNR via varying α. Note that the assortative setting is proposed for simulations, and our algorithm is applicable for general tensors in practice. The cluster assignment z is randomly generated with equal probability across r clusters for each mode. Without further explanation, we generate degree heterogeneity $\boldsymbol{\theta}$ from absolute normal distribution as $\theta(i) = |X_i| + 1 - 1/\sqrt{2\pi}$ with $|X_i| \overset{\text{i.i.d.}}{\sim} N(0,1), i \in [p]$ and normalize $\boldsymbol{\theta}$ to satisfy (2). We set $\sigma^2 = 1$ for Gaussian data.

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

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

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 ℓ_2-normalized rows of the factor matrix;
- **HLloyd**: High-order Lloyd algorithm and high-order spectral clustering (Han et al., 2020);
- **SCORE**: Tensor-SCORE for clustering (Ke et al., 2019);

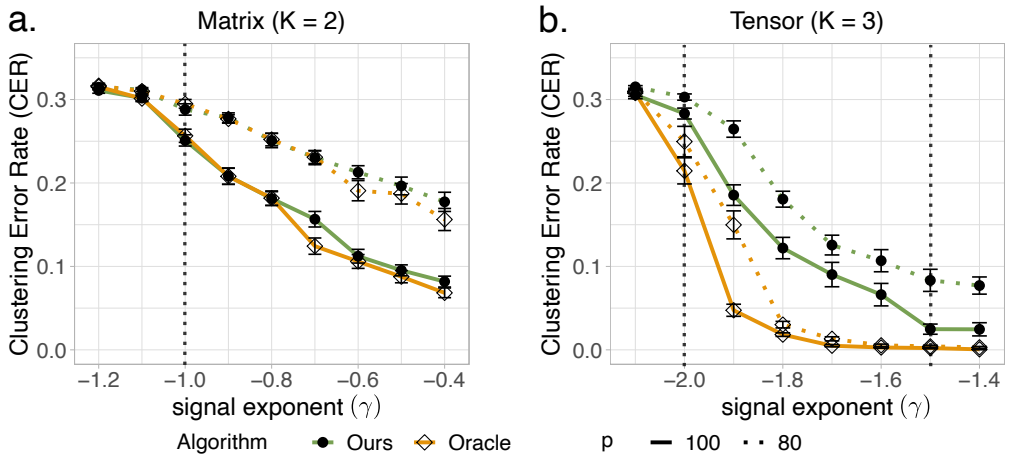

Fig. 4: SNR phase transitions for clustering in dTBM with $p = \{80, 100\}, r = 5$ under (a) matrix case with $\gamma \in [-1.2, -0.4]$ and (b) tensor case with $\gamma \in [-2.1, -1.4]$.

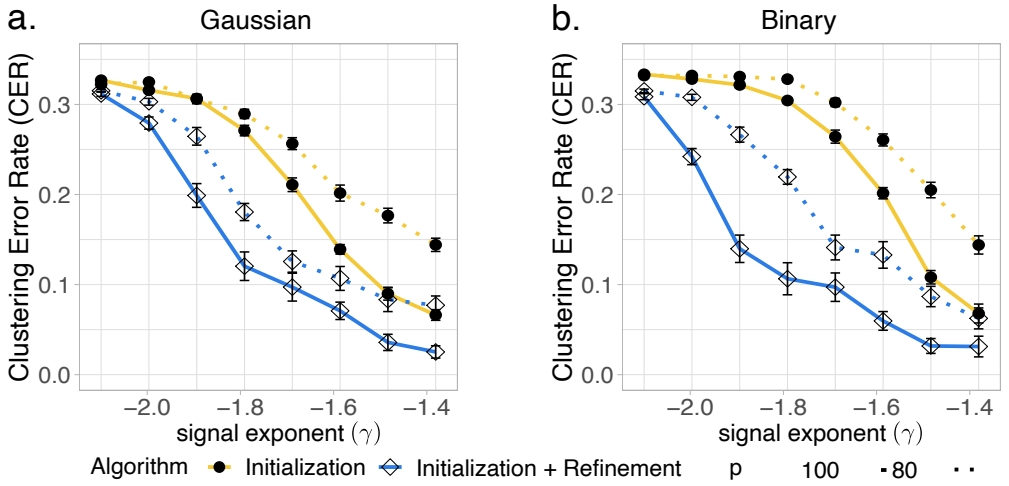

Fig. 5: CER versus signal exponent (γ) for initialization only and for combined algorithm. We set $p = \{80, 100\}, r = 5, \gamma \in [-2.1, -1.4]$ under (a) Gaussian models and (b) Bernoulli models.

Among the four alternative algorithms, the **SCORE** is the closest method to ours. We set the tuning parameters of **SCORE** as in previous literature (Ke et al., 2019). The methods **SCORE** and **HOSVD+** are designed for degree models, whereas **HOSVD** and **HLloyd** are designed for non-degree models. We conduct two experiments to assess the impacts of (i) signal strength and (ii) degree heterogeneity, based on Gaussian and Bernoulli models with $p = 100, r = 5$. We refer to our algorithm as **dTBM** in the comparison.

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

The only exception in Figure 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 V-B, except that

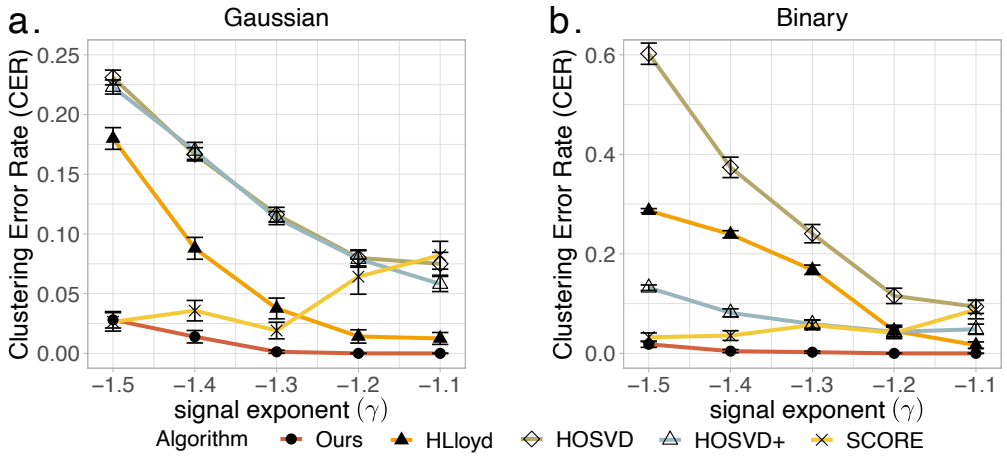

Fig. 6: CER versus signal exponent (denoted γ) for different methods. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under (a) Gaussian and (b) Bernoulli models.

we now generate the degree heterogeneity $\boldsymbol{\theta}$ from Pareto distribution with shape parameter a prior to normalization. We consider the Gaussian model under low $(a = 6)$ and high $(a = 2)$ degree heterogeneity. Figure 8 shows that the errors for non-degree algorithms (**HLloyd**, **HOSVD**) increases with degree heterogeneity. In addition, the advantage of **HLloyd** over **HOSVD+** disappears with higher degree heterogeneity.

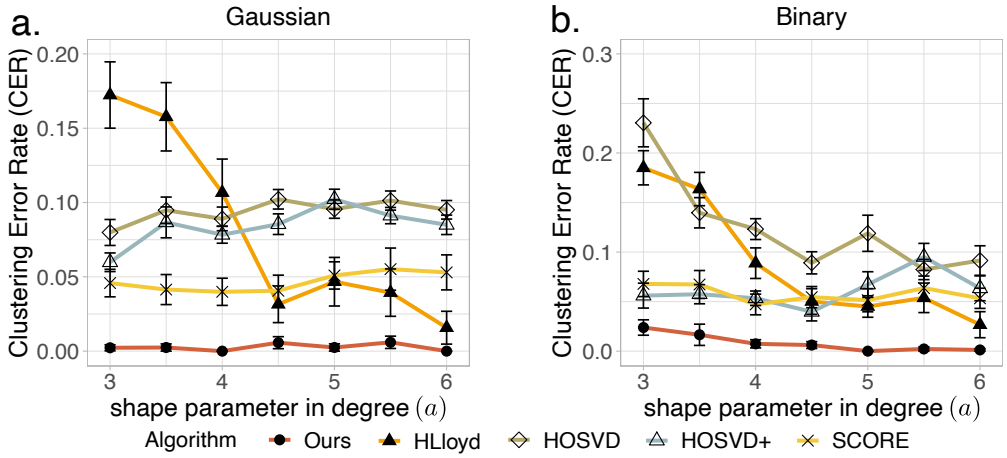

Fig. 7: CER versus shape parameter in degree (denoted $a \in [3, 6]$) for different methods. We set $p = 100, r = 5, \gamma = -1.2$ under (a) Gaussian and (b) Bernoulli models.

The last experiment investigates the effects of degree heterogeneity to clustering performance. We fix the signal exponent $\gamma = -1.2$ and vary the extent of degree heterogeneity. In this experiment, we generate $\boldsymbol{\theta}$ 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 a is called *shape* parameter. We vary $a \in [3, 6]$ and choose b such that $\mathbb{E}[X] = a(a-1)^{-1}b = 1$ for X following Pareto(a, b). Note that a smaller a leads to a larger variance in $\boldsymbol{\theta}$ and hence a larger degree heterogeneity. Figure 7 demonstrates the stability of degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over the entire range of degree heterogeneity under consideration. In contrast, non-degree algorithms (**HLloyd**, **HOSVD**) show poor performance with large heterogeneity, especially in Bernoulli cases. This experiment, again, highlights the benefit of addressing degree heterogeneity in higher-order clustering.

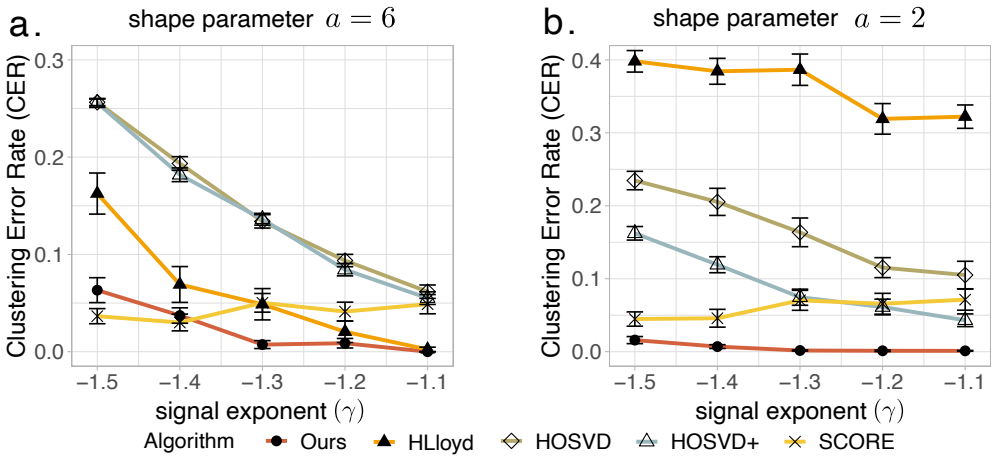

Fig. 8: CER comparison versus signal exponent (denoted γ) under (a) low (shape parameter $a = 6$) (b) high (shape parameter $a = 2$) degree heterogeneity. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under Gaussian model.

C. Human brain connectome data analysis

The Human Connectome Project (HCP) aims to construct the structural and functional neural connections in human brains (Van Essen et al., 2013). We preprocess the original dataset following Desikan et al. (2006) and partition the brain into 68 regions. The cleaned data includes brain networks for 136 individuals. Each brain network is represented by a 68-by-68 binary symmetric matrix, where the entries with value 1 refers to the presence of connection among 68 nodes while value 0 refers to the absence. We use $\mathcal{Y} \in \{0, 1\}^{68\times68\times136}$ to denote the binary observation. Individual attributes such as gender and sex are recorded.

We apply our generalized Algorithm to the HCP data with the number 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 chosen to be small following previous analysis (Hu et al., 2021). The estimated brain node clustering results on the first and second mode are the same. Figure 9 indicates 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 consistent with the common sense that the left and right hemispheres play different roles in human brain.

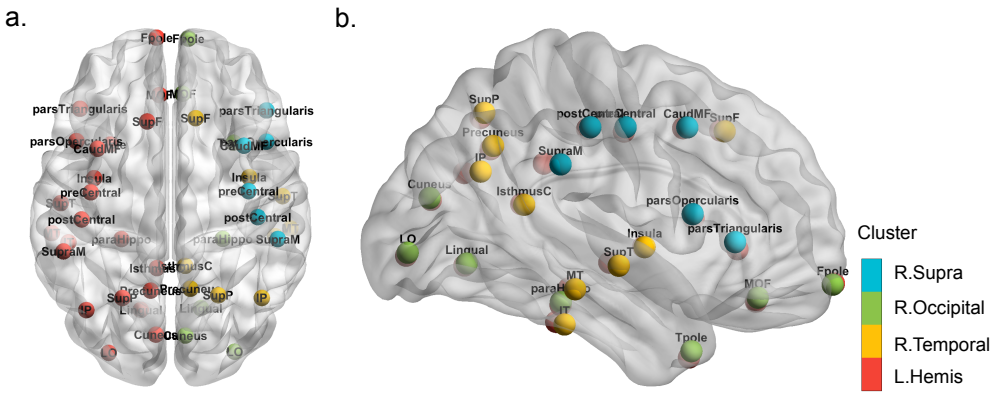

Fig. 9: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

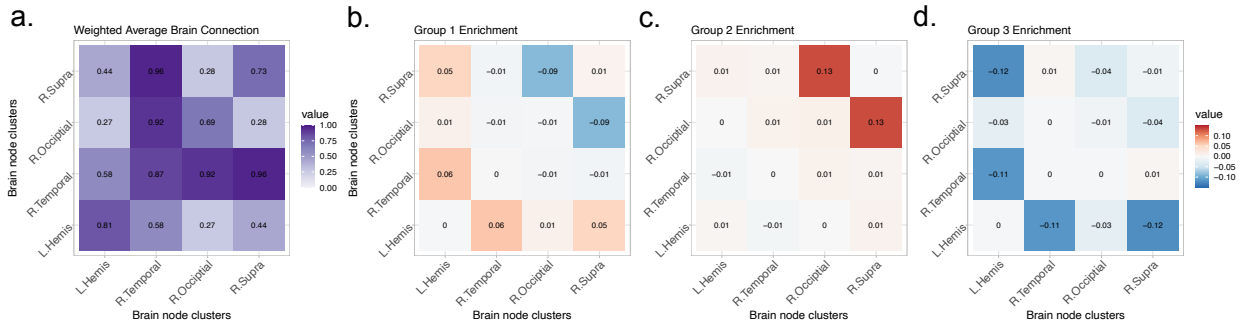

Fig. 10: Mode 3 slices of estimated core tensor $\hat{\mathcal{S}}$. (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the subtraction between each slice of $\hat{\mathcal{S}}$ and the averaged slice.

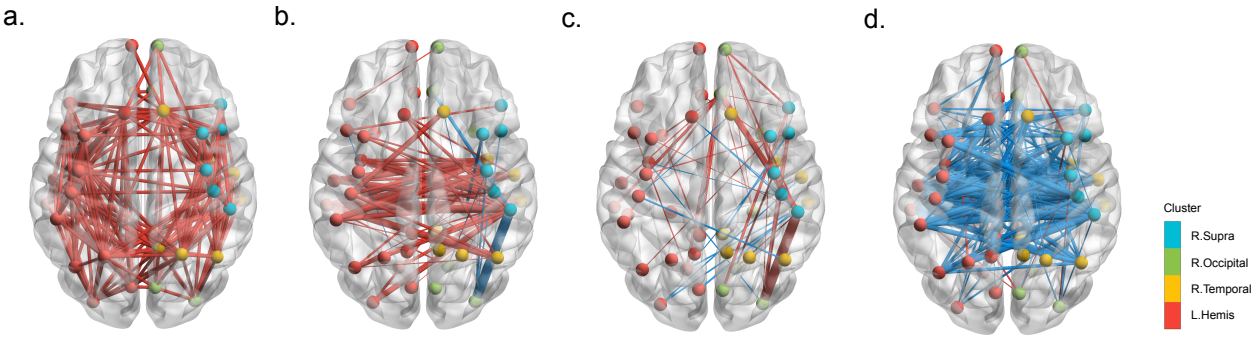

Fig. 11: Observed brain connections in the population and each group of individuals.(a) Average brain network; (b)-(d) Group-specified brain networks enrichment in Groups 1-3. Red edges refer to the positive enrichment and blue edges refer the negative reduction.

Figure 10 illustrates the estimated core tensor $\hat{\mathcal{S}}$ with estimated clustering, and Figure 11 visualizes the average observed brain connections and the connection enrichment with average observed networks in each group. In general, we find that the inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Figure 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 implies an enrichment on the inter-hemisphere connections (Figure 10b) while group 4 with only 36% females exhibits a reduction (Figure 10d). This result agrees with previous findings in Hu et al. (2021). The enrichment on the back-front connection is also recognized in group 3 (Figure 10c). The interpretive pattern in results demonstrate the usefulness of our clustering methods in the human brain connectome data application.

D. Peru Legislation data analysis

We also apply our method to the legislation networks in the Congress of the Republic of Peru (Lee et al., 2017). Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the data for the first half of 2006-2007 year. The dataset records the co-sponsorship of 116 legislators from top 5 parties and 802 bill proposals. We reconstruct legislation network as an order-3 binary tensor $\mathcal{Y} \in \{0,1\}^{116\times116\times116}$, 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 ground truth. We apply various higher-order clustering methods to $\mathcal{Y}$ with $r = 5$. Table II shows that our **dTBM** achieves the best performance compared to others. The second best method is the two-stage algorithm **HLloyd**, followed by the spectral methods **SCORE** and **HOSVD+**. This result is consistent with our simulations under strong signal and moderate degree heterogeneity. The comparison suggests that our method **dTBM** is more appealing in real-world applications.

Method	**dTBM**	**HOSVD**	**HOSVD+**	**HLloyd**	**SCORE**
CER	**0.116**	0.22	0.213	0.149	0.199

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

VI. Conclusion

We have developed a general degree-corrected tensor block model with a two-step angle-based polynomial-times algorithm. We have, for the first time, characterized the statistical and computational behaviors of the degree-corrected tensor block model under different signal-to-noise ratio regimes. Simulations and Peru Legislation and Human brain connection data analysis confirm the potential of our method for practical applications.

VII. Proof Sketches

1) **Angle-based techniques**
   - Padding of the vectors won't extremely change the angle gap.
   - Using geometry property to upper bound I_{11}, J_{11}. (have not found a good name for this technique yet)
2) **Proof idea of Theorem 4**
3) **Proof idea of Theorem 5**

References

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

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Ahn, K., Lee, K., and Suh, C. (2018). Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974.

Ahn, K., Lee, K., and Suh, C. (2019). Community recovery in hypergraphs. *IEEE Transactions on Information Theory*, 65(10):6561–6579.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR.

Chi, E. C., Gaines, B. J., Sun, W. W., Zhou, H., and Yang, J. (2020). Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Ghoshdastidar, D. and Dukkipati, A. (2017). Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *The Journal of Machine Learning Research*, 18(1):1638–1678.

Ghoshdastidar, D. et al. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315.

Han, R., Luo, Y., Wang, M., and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094.

Hu, J., Lee, C., and Wang, M. (2021). Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, pages 1–15.

Ke, Z. T., Shi, F., and Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.

Kim, C., Bandeira, A. S., and Goemans, M. X. (2018). Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Koniusz, P. and Cherian, A. (2016). Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5395–5403.

Lee, S. H., Magallanes, J. M., and Porter, M. A. (2017). Time-dependent community structure in legislation cosponsorship networks in the congress of the republic of peru. *Journal of Complex Networks*, 5(1):127–144.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

Meilă, M. (2012). Local equivalences of distances between clusterings—a geometric perspective. *Machine Learning*, 86(3):369–389.

Pananjady, A. and Samworth, R. J. (2020). Isotonic regression with unknown permutations: Statistics, computation, and adaptation. *arXiv preprint arXiv:2009.02609*.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., and WU-Minn HCP Consortium (2013). The WU-Minn human connectome project: An overview. *Neuroimage*, 80:62–79.

Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network–response regression. *Bioinformatics*, 33(12):1859–1866.

Wang, M., Fischer, J., and Song, Y. S. (2019). Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics*, 13(2):1103–1127.

Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21(154):1–38.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723.

Young, J.-G., St-Onge, G., Desrosiers, P., and Dubé, L. J. (2018). Universality of the stochastic block model. *Physical Review E*, 98(3):032309.

Yuan, M., Liu, R., Feng, Y., and Shang, Z. (2018). Testing community structures for hypergraphs. *arXiv preprint arXiv:1810.04617*.

Yun, S.-Y. and Proutiere, A. (2016). Optimal cluster recovery in the labeled stochastic block model. *Advances in Neural Information Processing Systems*, 29:965–973.

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338.

Zhang, Z., Allen, G. I., Zhu, H., and Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *Neuroimage*, 197:330–343.

APPENDIX

We provide the proofs for all the theorems in our main paper. In each sub-section, we show the proof of main theorem and attach the useful lemmas in the end.

NOTATION

Before the proofs, we first introduce the notations used throughout the following sections and the generalized dTBM without symmetric assumptions for the proofs of Theorem 1 and Theorem 2. The parameter space and minimal gap assumption for are also extended the generalized dTBM. The conclusions of Theorem 1 and 2 in the main paper are obtained by simply setting $z_k = z, \boldsymbol{S}_k = \boldsymbol{S}, \boldsymbol{\theta}_k = \boldsymbol{\theta}, k \in [K]$ for the generalized dTBM.

a) Notations.:

1) For all $k \in [K]$, denote the tensor matricizations as

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

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

3) For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m} \in \mathbb{R}^{n^K \times m^K}$, let $\boldsymbol{A}^{\otimes K}$ denotes the kronecker product of K matrices $\boldsymbol{A} \otimes \cdots \otimes \boldsymbol{A}$,

4) For two terms a and b, let $a \asymp b$ if there exist two positive constants c, C such that $cb \le a \le Cb$.

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

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where} \quad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \boldsymbol{M}_K, \tag{11}$$

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 mean-zero sub-Gaussian entries with variance bounded by σ^2, $\boldsymbol{\Theta}_k = \mathrm{diag}(\boldsymbol{\theta}_k)$, and $\boldsymbol{M}_k \in \{0,1\}^{p_k \times r_k}$ is the membership matrix corresponding to the assignment $z_k : [p_k] \mapsto [r_k]$, for all $k \in [K]$.

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

$$\mathcal{P}(\{r_k\}) = \Bigg\{ (\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) : \boldsymbol{\theta}_k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} |z_k^{-1}(a)| \le \frac{c_2 p_k}{r_k}, c_3 \le \|\boldsymbol{S}_{k,a:}\| \le c_4, \\ \|\boldsymbol{\theta}_{k, z_k^{-1}(a)}\|_1 = |z_k^{-1}(a)|, \text{ for all } a \in [r_k], k \in [K] \Bigg\}.$$

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

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

We also consider the generalized Assumption 1 on angle gap.

Assumption 2 (Generalized angle gap)**.** We assume the minimal gap between normalized rows of $\boldsymbol{S}_k$ is bounded away from zero, i.e.,

$$\Delta_{\min} := \min_{k \in [K]} \min_{a \ne b \in [r_k]} \left\| \boldsymbol{S}_{k,a:}^s - \boldsymbol{S}_{k,b:}^s \right\| > 0$$

for all $k \in [K]$. Similarly, let $\mathrm{SNR} = \Delta_{\min}^2 / \sigma^2$ with generalized minimal gap $\Delta_{\min}^2$ in Assumption 2. We define the regime

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

PROOF OF THEOREM 1

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

$$\mathcal{X} = \mathcal{S} \times_1 \Theta_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \Theta_K \boldsymbol{M}_K' = \mathcal{S}' \times_1 \Theta_1' \boldsymbol{M}_1' \times_2 \cdots \times_K \Theta_K' \boldsymbol{M}_K'$$

with $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$ and $(\{z_k'\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\}) \in \mathcal{P}(\{r_k'\})$. We prove the sufficient and necessary conditions separately.

($\Leftarrow$) For the necessity, it is equivalent to show that there exists a triplet $(\{z_k'\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\})$ is not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation if the model (11) violates Assumption 2. Without loss of generality, we assume $\left\|\boldsymbol{S}_{1,1:}^s - \boldsymbol{S}_{1,2:}^s\right\| = 0$.

If $\boldsymbol{S}_{1,1:}$ is a zero vector, consider $\boldsymbol{\theta}_1'$ such that $\boldsymbol{\theta}'_{z_1^{-1}(1)} \neq \boldsymbol{\theta}_{z_1^{-1}(1)}$. Let $\{z_k'\} = \{z_k\}$, $\mathcal{S}' = \mathcal{S}$, and $\boldsymbol{\theta}_k' = \boldsymbol{\theta}_k$ for all $k = 2, \ldots, K$. Then the triplet $(\{z_k'\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\})$ is not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation. We have a similar story when $\boldsymbol{S}_{1,2:}$ is a zero vector.

If neither $\boldsymbol{S}_{1,1:}$ nor $\boldsymbol{S}_{1,2:}$ is a zero vector, there exists a positive constant c such that $\boldsymbol{S}_{1,1:} = c\boldsymbol{S}_{1,2:}$. Thus, there exists a core tensor $\mathcal{S}_0 \in \mathbb{R}^{r_1 - 1 \times \cdots \times r_K}$ such that

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

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

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

and $\boldsymbol{M}_k' = \boldsymbol{M}_k, \boldsymbol{\theta}_k' = \boldsymbol{\theta}_k$ for all $k = 2, \ldots, K$. Then we have constructed a triplet $(\{z_k'\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\})$ that is not identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation.

($\Rightarrow$) For the sufficiency, it is equivalent 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 (11) satisfies Assumption (2). We show the uniqueness of the parameters separately.

First, we show the uniqueness of $\boldsymbol{M}_k$ for all $k \in [K]$. Specifically, we show the uniqueness of the first mode membership matrix, i.e., $\boldsymbol{M}_1' = \boldsymbol{M}_1 \boldsymbol{P}_1$ where $\boldsymbol{P}_1$ is a permutation matrix. The uniqueness of $\boldsymbol{M}_k$ for $k = 2, \ldots, K$ can be showed in the same way, and we omit the repeated procedures.

On one hand, consider the pair of nodes (i, j) such that $z_1(i) = z_1(j)$. We have $\|\boldsymbol{X}_{1,z_1(i):}^s - \boldsymbol{X}_{1,z_1(j):}^s\| = 0$ and thus $\|(\boldsymbol{S}')_{1,z_1'(i):}^s - (\boldsymbol{S}')_{1,z_1'(j):}^s\| = 0$ by Lemma 1. Then, by Assumption (2), we have $z_1'(i) = z_1'(j)$. On the other hand, consider the pair of nodes (i, j) such that $z_1(i) \neq z_1(j)$. We have $\left\|\boldsymbol{X}_{1,i:}^s - \boldsymbol{X}_{1,j:}^s\right\| \neq 0$ and thus $\left\|(\boldsymbol{S}')_{1,z_1'(i):}^s - (\boldsymbol{S}')_{1,z_1'(j):}^s\right\| \neq 0$ by Lemma 1. Hence, we have $z_1'(i) \neq z_1'(j)$. Therefore, we have proven that z_1' is equal to z_i up to label permutation.

Next, we show the uniqueness of $\boldsymbol{\theta}_k$ for all $k \in [K]$ given that $z_k = z_k'$. Similarly, we show the detailed proof of the uniqueness of $\boldsymbol{\theta}_1$, i.e., $\boldsymbol{\theta}_1' = \boldsymbol{\theta}_1$, and omit the repeated procedures for $\boldsymbol{\theta}_k, k = 2, \ldots, K$.

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

$$\frac{\boldsymbol{X}_{1,z_1(i):}}{\boldsymbol{X}_{1,z_1(j):}} = \frac{\boldsymbol{X}'_{1,z_1(i):}}{\boldsymbol{X}'_{1,z_1(j):}}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta_1'(j)}{\theta_1'(i)}. \tag{12}$$

Let $\theta_1'(j) = c\theta_1(j)$ for some constant c. By the equation (12), we have $\theta_1'(i) = c\theta_1(i)$ for all $i \in z_1^{-1}(a)$. Note that $(\{z_k\}, \mathcal{S}', \{\boldsymbol{\theta}_k'\}) \in \mathcal{P}(\{r_k\})$. We have

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

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

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

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

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

Therefore, we finish the proof of Theorem 1. □

c) Useful Lemma for the Proof of Theorem 1:

Lemma 1 (Motivation of angle-based clustering)**.** Consider the signal tensor $\mathcal{X}$ in the generalized dTBM (11) with $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\{r_k\})$ and $r_k \geq 2$. Then, for any $k \in [K]$ and index pair $(i, j) \in [p_k]^2$, we have

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

Proof of Lemma 1. For simplicity, we show the detailed proof for $k = 1$ and drop the subscript k in $\boldsymbol{X}_k, \boldsymbol{S}_k$.The repeated proofs for $k = 2, \ldots, K$ are omitted.

By tensor matricization, we have

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

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

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

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

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

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

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

where $\tilde{\boldsymbol{M}}^T \tilde{\boldsymbol{M}}$ is an invertiable diagonal matrix with positive diagonal elements. Thus, we have $\boldsymbol{S}_{z_1(i):} = c \boldsymbol{S}_{z_1(j):}$, which implies $\left\| \boldsymbol{S}^s_{z_1(i):} - \boldsymbol{S}^s_{z_1(j):} \right\| = 0$.

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

Therefore, we finish the proof of Lemma 1. □

PROOF OF THEOREM 2

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

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

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

$$\sup_{(\{z_k\},\mathcal{S},\{\boldsymbol{\theta}_k\})\in\mathcal{P}(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{k,\text{stat}}, z_k)\right] \geq 1. \tag{13}$$

Noticed that inequality (13) is a minimax lower bound, it suffices to show the inequality holds for a particular $(\{z_k\},\mathcal{S},\{\boldsymbol{\theta}_k\}) \in \mathcal{P}(\gamma)$. Specifically, we consider the estimation problem based on a particular parameter point $(\{z_k\},\mathcal{S},\{\boldsymbol{\theta}_k\})$ with the following three properties:

$$\text{(i) } \theta_k(i) = 1 \text{ for all } i \in [p]; \text{ (ii) } \Delta_{\min} \lesssim \left(\frac{p}{r}\right)^{-\frac{K-1}{2}} \sigma; \text{ (iii) } |z_k^{-1}(a)| = \frac{p}{r} \in \mathbb{Z}_+ \text{ for all } a \in [r], \tag{14}$$

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

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

for all $k \in [K]$. The construction of T is precisely the same as Han et al. (2020, Proof of Theorem 6). Then, following the proof of Gao et al. (2018, Theorem 2), for all $k \in [K]$, we have

$$\inf_{\hat{z}_k} \sup_{z_k} \mathbb{E}\ell(\hat{z}_k, z_k) \gtrsim \frac{1}{r^3|T_k^c|} \sum_{i\in T_k^c} \inf_{\hat{z}_k} \left\{\mathbb{P}[\hat{z}_k(i) = 2|z_k(i) = 1] + \mathbb{P}[\hat{z}_k(i) = 1|z_k(i) = 2]\right\}, \tag{15}$$

where $\hat{z}_k, z_k$ on the left hand side denote the generic clustering functions in $\mathcal{P}(\gamma)$, z_k on the right hand side denotes a particular parameter satisfying properties (i)-(vi), and the infimum on the right hand side is taken over the restricted family of $\hat{z}$ satisfying (iv)-(vi). Here, the factor $r^3 = r \cdot r^2$ in (15) 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 $|T_k^c| \asymp \frac{p}{r}$ to be clustered.

Next, we need to find the lower bound of the rightmost side in (15). For simplicity, we only show the bound for the mode-1 case $k = 1$. We drop the subscripts 1 in $z_1, T_1, \boldsymbol{S}_1, \boldsymbol{\theta}_1$ and omit the repeated procedures for the cases of $k = 2, \ldots, K$.

We consider the hypothesis test based on model (11). First, we reparameterize the model under the construction (14)

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

where $\boldsymbol{x}_a$'s are centroids in $\mathbb{R}^{p^{K-1}}$. Without loss of generality, we consider the lower bound for the summand in (15) for $i = 1$. The analysis for other $i \in T^c$ are similar. For notational simplicity, we suppressed the subscript i and write $\boldsymbol{y}, \theta, z$ in place of $\boldsymbol{y}_1, \theta_1$ and $z(1)$, respectively. The equivalent vector problem for assessing the summand in (15) is

$$\boldsymbol{y} = \theta \boldsymbol{x}_z + \boldsymbol{e}, \tag{16}$$

where $\theta \in \mathbb{R}_+$ and $z \in \{1, 2\}$ are unknown parameters, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{p^{K-1}}$ are given centroids, and $\boldsymbol{e} \in \mathbb{R}^{p^{K-1}}$ consists of i.i.d. $N(0, \sigma^2)$ entries. Then, we consider the hypothesis testing under the model (16):

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

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

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

and the maximum likelihood estimators (MLE) of θ and z are

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

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

$$\hat{z}_{\text{MLE}} = \begin{cases} 1 & \text{if } \cos(\boldsymbol{y}, \boldsymbol{x}_1) \geq \cos(\boldsymbol{y}, \boldsymbol{x}_2) \text{ and } \langle \boldsymbol{y}, \boldsymbol{x}_1 \rangle > 0, \\ 2 & \text{if } \cos(\boldsymbol{y}, \boldsymbol{x}_1) < \cos(\boldsymbol{y}, \boldsymbol{x}_2) \text{ and } \langle \boldsymbol{y}, \boldsymbol{x}_2 \rangle > 0, \\ 1 \text{ or } 2 \text{ with equal probability} & \text{otherwise.} \end{cases} \tag{17}$$

Neyman-Pearson Lemma implies

$$\inf_{\hat{z}} \{\mathbb{P}[\hat{z} = 2 | z = 1] + \mathbb{P}[\hat{z} = 1 | z = 2]\} = \mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] + \mathbb{P}[\hat{z}_{\text{MLE}} = 2 | z = 1]. \tag{18}$$

By symmetric, it suffices to bound $\mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2]$. Using (17), we obtain

$$\begin{aligned} \mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] &= \mathbb{P}\left[\cos(\theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_1) \geq \cos(\theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_2) \text{ and } \langle \theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_1 \rangle > 0\right] \\ &\overset{(*)}{\geq} \mathbb{P}\left[\left\langle \boldsymbol{e}, \frac{\boldsymbol{x}_1^s - \boldsymbol{x}_2^s}{\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|} \right\rangle \geq \frac{\theta}{2} \|\boldsymbol{x}_2\| \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\| \rangle\right] - \\ &\qquad \mathbb{P}\left[\langle \boldsymbol{e}, \boldsymbol{x}_1^s \rangle \leq -\frac{\theta}{2} \|\boldsymbol{x}_2\| \left(2 - \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2\right)\right] \\ &\overset{(**)}{=} \Phi\left(\frac{\theta}{2} \|\boldsymbol{x}_2\| \left(2 - \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2\right)\right) - \Phi\left(\frac{\theta}{2} \|\boldsymbol{x}_2\| \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|\right), \end{aligned} \tag{19}$$

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

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

By construction (14) of $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ with three properties and lower bound $\min_{a \in [r]} \|\boldsymbol{S}_{a:}\| \geq c_3$ in the definition of $\mathcal{P}(\gamma)$, we have $\theta^* = 1$, $\|\boldsymbol{x}_2\| \geq \|\boldsymbol{S}_{2:}\| \min_{a \in [r]} \left\|\boldsymbol{\theta}_{\boldsymbol{z}^{-1}(a)}\right\|^{K-1} \gtrsim \left(\frac{p}{r}\right)^{(K-1)/2}$. Also, note that under the construction (14)

$$\cos(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\boldsymbol{x}_1, \boldsymbol{x}_2}{\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|} = \frac{(p/r)^{K-1} \langle \boldsymbol{S}_{1:}, \boldsymbol{S}_{2:} \rangle}{\sqrt{(p/r)^{K-1} \|\boldsymbol{S}_{1:}\|^2} \sqrt{(p/r)^{K-1} \|\boldsymbol{S}_{2:}\|^2}} = \cos(\boldsymbol{S}_{1:}, \boldsymbol{S}_{2:}),$$

which implies $\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\| = \|\boldsymbol{S}_{1:}^s - \boldsymbol{S}_{2:}^s\| = \Delta_{\min} \leq 1$. Therefore, the equation (19) is lower bounded by

$$\mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] \geq \mathbb{P}\left[\left(\frac{p}{r}\right)^{(K-1)/2} \Delta_{\min} \lesssim N(0, 1) \lesssim \left(\frac{p}{r}\right)^{(K-1)/2}\right] \geq C > 0, \tag{20}$$

where the existence of strictly positive constant C is based on the SNR assumption (14). Combining (15), (18) and (20) yields

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

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

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

□

PROOF OF THEOREM 3

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

$$\mathcal{S} = \mathcal{S}(\gamma) = \begin{bmatrix}1\\1\\1\end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2} \begin{bmatrix}1\\-1\\0\end{bmatrix}^{\otimes K}. \tag{21}$$

Then, we consider the signal tensor family

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

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

(i) For every $\gamma \in \mathbb{R}$, $\mathcal{P}_{\text{shifted}}(\gamma) \subset \mathcal{P}(\gamma)$, where $\mathcal{P}(\gamma)$ is the degree-corrected cluster tensor family (5).

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

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

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

$$\liminf_{p\to\infty} \sup_{\mathcal{X}\in\mathcal{P}_{\text{non-degree}}(\gamma)} \mathbb{P}(\exists k \in [K],\ \hat{\boldsymbol{M}}_k \neq \boldsymbol{M}_k) \geq 1/2, \tag{22}$$

under the HPC Conjecture 1. The inequality (22) implies

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

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

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

We complete the proof by verifying the properties (i)-(ii). For (i), we verify that the angle gap for the core tensor $\mathcal{S}$ in (21) is on the order of $\sigma p^{-\gamma/2}$. Specifically, write $\mathbf{1} = (1,1,1)$ and $\boldsymbol{e} = (1,-1,0)$. We have

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

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

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

Therefore, we have shown that $\mathcal{P}_{\text{shifted}}(\gamma) = \mathcal{P}(\gamma)$. Finally, the property (ii) follows directly by comparing the definition of $\mathcal{S}$ in (21) with that in the proof of Han et al. (2020, Theorem 7). □

d) Useful Definitions and Lemmas for the Proof of Theorem 3:

Definition 3 (Principal angles)**.** For two matrices $\boldsymbol{U}, \hat{\boldsymbol{U}} \in \mathbb{O}^{p\times r}$, we define the principal angles between $\boldsymbol{U}$ and $\hat{\boldsymbol{U}}$ by the following diagonal matrix

$$\text{Angle}(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \text{diag}(\arccos(\sigma_1), \cdots, \arccos(\sigma_r)),$$

where $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are singular values of $\boldsymbol{U}^T\hat{\boldsymbol{U}}$. We define the norm of the principal angles as

$$\left\|\sin(\boldsymbol{U}, \hat{\boldsymbol{U}})\right\|_\sigma = \max_{i\in[r]} \sin(\arccos(\sigma_i)).$$

We restate the Lemma 3 from Zhang and Xia (2018) for self-contentedness.

Lemma 2 (Bounds for principal angle spectral norm.)**.** For any $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{O}_{p,r}$ and all $1 \leq q \leq +\infty$, we have

$$\frac{1}{4}\left\|\boldsymbol{U}_1\boldsymbol{U}_1^T - \boldsymbol{U}_2\boldsymbol{U}_2^T\right\|_\sigma \leq \left\|\sin(\boldsymbol{U}_1, \boldsymbol{U}_2)\right\|_\sigma \leq \left\|\boldsymbol{U}_1\boldsymbol{U}_1^T - \boldsymbol{U}_2\boldsymbol{U}_2^T\right\|_\sigma.$$

Proof of Lemma 2. See Zhang and Xia (2018, Proof of Lemma 3). □

Lemma 3 (Computational lower bound of tensor decomposition. (Theorem 4 in Zhang and Xia (2018)))**.** Consider the tucker decomposition model

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K + \mathcal{E},$$

where $\mathcal{Y} \in \mathbb{R}^{p^{\otimes^K}}, \mathcal{S} \in \mathbb{R}^{r^{\otimes^K}}$ with $r \geq 1$ is a full rank core tensor, $\boldsymbol{M}_k \in \mathbb{R}^{p\times r}$ has orthonormal columns for all $k \in [K]$, and $\mathcal{E}$ has independent entries following $N(0, \sigma^2)$. Let $\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K$. Define the singular signal $\lambda = \min_{k\in[K]} \lambda_{\min}(\mathrm{Mat}_k(\mathcal{X}))$, where $\lambda_{\min}(\cdot)$ denotes the minimal non-zero singular value. Suppose Conjecture 1 holds for some $\tau \in (0, 1)$. If $\lambda/\sigma \lesssim p^{K(1-\tau)/4}$, then for any polynomial times estimator $\hat{\boldsymbol{M}}_k$, the following inequality holds

$$\liminf_{p\to\infty} \sup_{(\mathcal{S},\boldsymbol{M}_1,\ldots,\boldsymbol{M}_K) \text{ has singular signal } \lambda} \mathbb{E}\left\|\sin(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_k)\right\|_\sigma^2 \geq c,$$

where $c > 0$ is a positive constant.

Proof of Lemma 3. See Zhang and Xia (2018, Proof of Theorem 4). □

Lemma 4 (Angle distance and misclassification error)**.** Consider two membership matrices $\boldsymbol{M}, \hat{\boldsymbol{M}} \in \mathbb{R}^{p\times r}$ with corresponding assignments $z, \hat{z} : [p] \mapsto [r]$. If the angle distance $\left\|\sin(\boldsymbol{M}, \hat{\boldsymbol{M}})\right\|_\sigma \geq c$ for a constant $c > 0$, then the misclassification error $\min_{\pi\in\Pi} p\ell(\hat{z}, \pi \circ z) \geq 1$.

Proof of Lemma 4. By Lemma 2, we know that for any permutation matrix $\boldsymbol{P} \in \mathbb{R}^{r\times r}$,

$$\left\|\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T - \boldsymbol{M}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{M}^T\right\|_F \geq \left\|\hat{\boldsymbol{M}}\hat{\boldsymbol{M}}^T - \boldsymbol{M}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{M}^T\right\|_\sigma \geq \left\|\sin(\boldsymbol{M}, \hat{\boldsymbol{M}})\right\|_\sigma \geq c > 0, \tag{23}$$

where the second inequality follows from the invariance of sin distance to permutation. The inequality (23) implies that there exists at least one disagreement in z and $\hat{z}$; i.e., there exists a pair (i, j) such that $z(i) = z(j)$ but $\hat{z}(i) \neq \hat{z}(j)$. Therefore, we have $\min_{\pi\in\Pi} p\ell(\hat{z}, \pi \circ z) \geq 1$. □

PROOF OF THEOREM 4

Proof of Theorem 4. Recall that we prove the Theorem 4 under the dTBM (1) with parameters $(z, \mathcal{S}, \boldsymbol{\theta})$. We drop the subscript k in the matricizations $\boldsymbol{X}_k, \boldsymbol{S}_k, \boldsymbol{X}_k$. For simplicity, let $\hat{z}$ denote the output of Sub-Algorithm 1 $\hat{z}^{(0)}$.

First, by Lemma 7, there exists a positive constant such that $\min_{z(i)\neq z(j)} \left\|\boldsymbol{X}_{i:}^s - \boldsymbol{X}_{j:}^s\right\| \geq c_0\Delta_{\min}$. Also, by balanced assumption on $\boldsymbol{\theta}$ and Lemma 10, we have

$$\min_{\pi\in\Pi} \sum_{i:\hat{z}(i)\neq\pi(z(i))} \theta(i)^2 \leq \sum_{i\in S_0} \theta(i)^2 + 4\sum_{i\in S} \theta(i)^2, \tag{24}$$

where

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

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

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

where the last inequality follows Lemma 8, the assumption that $\min_{i\in[p]}\theta(i)\geq c>0$ and $\min_{a\in[r]}\|\boldsymbol{S}_{a:}\|\geq c_3$ in the parameter space (2). Thus, we have

$$\sum_{i\in P}\theta(i)^2 \lesssim \sum_{i\in P}\|\boldsymbol{X}_{i:}\|^2 p^{-(K-1)}r^{K-1}. \tag{25}$$

On the other hand, note that

$$\begin{aligned}\sum_{i\in S}\|\boldsymbol{X}_{i:}\|^2 &\leq 2\sum_{i\in S}\left\|\hat{\boldsymbol{X}}_{i:}\right\|^2 + 2\sum_{i\in S}\left\|\hat{\boldsymbol{X}}_{i:}-\boldsymbol{X}_{i:}\right\|^2 &(26)\\ &\leq \frac{8}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\left\|\hat{\boldsymbol{X}}_{i:}\right\|^2\|\hat{\boldsymbol{x}}_{\hat{z}(i)}-\boldsymbol{X}^s_{i:}\|^2 + 2\left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_F^2 &(27)\\ &\leq \frac{16}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\left\|\hat{\boldsymbol{X}}_{i:}\right\|^2\left[\left\|\hat{\boldsymbol{x}}_{\hat{z}(i)}-\hat{\boldsymbol{X}}^s_{i:}\right\|^2+\left\|\hat{\boldsymbol{X}}^s_{i:}-\boldsymbol{X}^s_{i:}\right\|^2\right] + 2\left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_F^2 &(28)\\ &\leq \frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\left\|\hat{\boldsymbol{X}}_{i:}\right\|^2\left\|\hat{\boldsymbol{X}}^s_{i:}-\boldsymbol{X}^s_{i:}\right\|^2 + 2\left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_F^2 &(29)\\ &\leq \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}+2\right)\left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_F^2 &(30)\\ &\lesssim \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}+2\right)\left(p^{K/2}r+pr^2+r^K\right)\sigma^2, &(31)\end{aligned}$$

where inequalities (26) and (28) follow the triangle inequality, (27) follows the definition of S, (29) follows the update rule of k-means in Sub-Algorithm 1, (30) follows Lemma 5, and the last inequality (31) follows Lemma 9. Also, note that

$$\sum_{i\in S_0}\|\boldsymbol{X}_{i:}\|^2 = \sum_{i\in S_0}\left\|\hat{\boldsymbol{X}}_{i:}-\boldsymbol{X}_{i:}\right\|^2 \leq \left\|\hat{\mathcal{X}}-\mathcal{X}\right\|_F^2 \lesssim \left(p^{K/2}r+pr^2+r^K\right)\sigma^2, \tag{32}$$

where the equation follows by the definition of S_0. Therefore, combining the inequalities (24), (25), (31), and (32), we have

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

With the assumption that $\min_{i\in[p]}\theta(i)\geq c>0$, we finally obtain the result

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

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

e) Useful Corollary of Theorem 4:

Corollary 1. Suppose we have the initial assignment $z^{(0)}$ from Sub-Algorithm 1 and the identity permutation minimizes the misclassification error, i.e., $\pi^{(0)} = \arg\min_{\pi\in\Pi} \sum_{i\in[p]} \mathbb{1}\left\{z^{(0)}(i) \neq \pi \circ z(i)\right\}$ and $\pi^{(0)}(a) = a$ for all $a \in [r]$. Suppose SNR $\gtrsim p^{-K/2}\log p$. Then, the misclassification loss for the initialization is upper bounded as

$$L^{(0)} \lesssim \frac{\Delta_{\min}^2}{r\log p}.$$

Proof. Note that

$$\begin{aligned}
\sum_{i\in[p]} \|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{x}}_a\|^2 &\leq 2\sum_{i\in[p]} \left\|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{X}}_{i:}^s\right\|^2 + \sum_{i\in[p]} \left\|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_a\right\|^2 \\
&\leq \frac{2}{\min_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2}\left[\sum_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2\left\|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{X}}_{i:}^s\right\|^2 + \sum_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2\left\|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_a\right\|^2\right] \\
&\lesssim \frac{r^{K-1}(1+\eta)}{p^{K-1}}\sum_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2\left\|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{X}}_{i:}^s\right\|^2 \\
&\lesssim \frac{r^{K-1}(1+\eta)}{p^{K-1}}\left\|\hat{\mathcal{X}} - \mathcal{X}\right\|_F^2 \\
&\lesssim \frac{r^{K-1}(1+\eta)}{p^{K-1}}p^{-K/2}.
\end{aligned}$$

Suppose $z^{(0)} = 1$. Note that $\boldsymbol{X}_{i:}^s$ only have r different values. Let $\boldsymbol{X}_a^s = \boldsymbol{X}_{i:}^s$ with $z(i) = a, a \in [r]$. Note that

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

when $\Delta_{\min}^2 \geq p^{-K/2}\log p$. Then

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

Note that

$$\begin{aligned}
L^{(0)} &= \frac{1}{p}\sum_{i\in[p], z^{(0)}(i)\neq z(i)} \theta(i)\left\|\boldsymbol{S}_{z^{(0)}(i),:}^s - \boldsymbol{S}_{z(i),:}^s\right\|^2 \\
&\lesssim \frac{1}{p}\sum_{i\in[p], z^{(0)}(i)\neq z(i)} \theta(i)\left\|\boldsymbol{X}_{z^{-1}(z^{(0)}):}^s - \boldsymbol{X}_{i,:}^s\right\|^2 \\
&\lesssim \frac{1}{p}\sum_{i\in[p], z^{(0)}(i)\neq z(i)} \theta(i)\left[\left\|\boldsymbol{X}_{z^{-1}(z^{(0)}):}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}\right\|^2 + \left\|\hat{\boldsymbol{x}}_{z^{(0)}(i)} - \boldsymbol{X}_{i,:}^s\right\|^2\right] \\
&\lesssim \frac{1}{p}\sum_{i\in[p], z^{(0)}(i)\neq z(i)} \theta(i)\frac{\left\|\hat{\boldsymbol{X}}_{i:} - \boldsymbol{X}_{i:}\right\|^2}{\|\boldsymbol{X}_{i:}\|^2} + \frac{1}{p}\sum_{i\in[p], z^{(0)}(i)\neq z(i)} \theta(i)\frac{r^K(1+\eta)}{p^K}p^{-K/2} \\
&\lesssim \frac{r^K}{p^K}p^{-K/2}
\end{aligned}$$

$$\lesssim \frac{\Delta_{\min}^2}{r \log p}.$$

where $\frac{1}{p}\sum_{i\in[p],z^{(0)}(i)\neq z(i)}\theta(i) \lesssim \frac{1}{p}\sum_{i\in[p],z^{(0)}(i)\neq z(i)}\theta(i)^2$ due to the boundedness. □

f) Useful Definitions and Lemmas for the Proof of Theorem 4:

Lemma 5 (Basic inequality)**.** For any two nonzero vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ of same dimension, we have

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

Proof of Lemma 5. For the first inequality, let $\alpha \in [0, \pi]$ denote the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. We have

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

where the equations follows by the properties of trigonometric function and the inequality follows by 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 $\|\boldsymbol{v}_1\| \geq \|\boldsymbol{v}_2\|$. Then

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

Therefore, Lemma 5 is proved. □

Definition 4 (Weighted padding vectors)**.** For a vector $\boldsymbol{a} = [\![a_i]\!] \in \mathbb{R}^d$, we define the padding vector of $\boldsymbol{a}$ with the weight collection $\boldsymbol{w} = \{\boldsymbol{w}_i\}_{i=1}^d$ and $\boldsymbol{w}_i = [\![w_{ik}]\!] \in \mathbb{R}^{p_i}, i \in [d]$ as

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

and $\text{Pad} : \mathbb{R}^d \mapsto \mathbb{R}^{\sum_{i\in[d]} p_i}$ can be viewed as an operator, and by definition, we have $\text{Pad}(\boldsymbol{a}) \in \mathbb{R}^{\sum_{i\in[d]} p_i}$. Then , we have the bounds of the weighted padding vector

$$\min_{i\in[d]} \|\boldsymbol{w}_i\|^2 \|\boldsymbol{a}\|^2 \leq \|\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a})\|^2 \leq \max_{i\in[d]} \|\boldsymbol{w}_i\|^2 \|\boldsymbol{a}\|^2. \tag{33}$$

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

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

Lemma 6 (Angle for weighted padding vectors)**.** Suppose we have two non-zero vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$. Given the weight collection $\boldsymbol{w}$, we have

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

Proof of Lemma 6. We prove the two inequalities separately with the similar idea.

First, we prove the inequality $\sin(\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a}), \text{Pad}_{\boldsymbol{w}}(\boldsymbol{b})) \leq \frac{\max_{i\in[d]} \|\boldsymbol{w}_i\|^2}{\min_{i\in[d]} \|\boldsymbol{w}_i\|^2} \sin(\boldsymbol{a}, \boldsymbol{b})$. Decomposing $\boldsymbol{b}$, we have

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

where $\boldsymbol{a}^\perp \in \mathbb{R}^d$ is the orthogonal vector of $\boldsymbol{a}$. By the Definition 4, we have

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

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

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

where the second inequality follows by applying the property (33) to vectors $\boldsymbol{b},\boldsymbol{a}^\perp$.

Next, we prove $\frac{\min_{i\in[d]}\|\boldsymbol{w}_i\|^2}{\max_{i\in[d]}\|\boldsymbol{w}_i\|^2}\sin(\boldsymbol{a},\boldsymbol{b}) \leq \sin(\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{a}),\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{b}))$. With the decomposition of $\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{b})$ and the inverse weighted padding operator, we have

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

Therefore, we obtain

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

where the second inequality follows by applying the property (33) to vectors $\boldsymbol{b},\mathrm{Pad}_{\boldsymbol{w}}^{-1}((\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{a}))^\perp)$.

□

Lemma 7 (Angle gap in $\mathcal{X}$)**.** Consider the dTBM model (1). Suppose Assumption 1 holds and $\boldsymbol{\theta}$ is balanced satisfying (6). Then the angle gap in $\mathcal{X}$ is approximate to angle gap in $\mathcal{S}$, i.e.,

$$\Delta_{\min} \lesssim \min_{(i,j):z(i)\neq z(j)}\left\|\boldsymbol{X}_{i:}^s - \boldsymbol{X}_{j:}^s\right\|.$$

Proof of Lemma 7. Note that the vector $\boldsymbol{S}_{z(i):}$ can be folded to a tensor $\mathcal{S}' = [\![\mathcal{S}'_{a_2,\ldots,a_K}]\!] \in \mathbb{R}^{r^{K-1}}$, i.e., $\mathrm{vec}(\mathcal{S}') = \boldsymbol{S}_{z(i):}$. Let the weight vector $\boldsymbol{w}_{a_2\cdots a_K}$ correspond to the elements $\mathcal{S}'_{a_2,\ldots,a_K}$ as

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

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

$$\max_{(a_2,\ldots,a_K)}\|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 = (1+o(1))\min_{(a_2,\ldots,a_K)}\|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2. \tag{34}$$

Consider the inner product of $\boldsymbol{X}_{i:}$ and $\boldsymbol{X}_{j:}$ for $z(i)\neq z(j)$. By the definition of weighted padding operator and the balanced assumption (34), we have

$$\begin{aligned}\langle \boldsymbol{X}_{i:},\boldsymbol{X}_{j:}\rangle &= \theta(i)\theta(j)\left\langle \mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{S}_{z(i):}),\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{S}_{z(j):})\right\rangle \\ &= \theta(i)\theta(j)\min_{(a_2,\ldots,a_K)}\|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2\left\langle \boldsymbol{S}_{z(i):},\boldsymbol{S}_{z(j):}\right\rangle(1+o(1)).\end{aligned}$$

Thus, when p large enough, the inner product $\langle \boldsymbol{X}_{i:}, \boldsymbol{X}_{j:} \rangle$ as the same sign with $\langle \boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):} \rangle$. Next, we discuss the angle between $\boldsymbol{X}_{i:}$ and $\boldsymbol{X}_{j:}$ by cases.

1) **Case 1:** Suppose $\langle \boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):} \rangle \leq 0$. Then, we also have $\langle \boldsymbol{X}_{i:}, \boldsymbol{X}_{j:} \rangle \leq 0$, which implies $\left\| \boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:} \right\| \geq \sqrt{2}$. Note that $\left\| \boldsymbol{S}^s_{z(i):} - \boldsymbol{S}^s_{z(j):} \right\| \leq 2$. We have $\left\| \boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:} \right\| \gtrsim \left\| \boldsymbol{S}^s_{z(i):} - \boldsymbol{S}^s_{z(j):} \right\|$.

2) **Case 2:** Suppose $\langle \boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):} \rangle > 0$. Then, we have $\cos(\boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):}) > 0$. Note that the fact $\sqrt{1-\cos\alpha} = 2\sin\frac{\alpha}{2} \lesssim \sin\alpha$ for the angle $\alpha \in [0, \frac{\pi}{2})$. Then, we have

$$\begin{aligned}
\left\| \boldsymbol{S}^s_{z(i):} - \boldsymbol{S}^s_{z(j):} \right\| &= \sqrt{1 - \cos(\boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):})} \\
&\lesssim \sin(\boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):}) \\
&\leq \frac{\max_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|}{\min_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|} \sin(\mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{S}_{z(i):}), \mathrm{Pad}_{\boldsymbol{w}}(\boldsymbol{S}_{z(j):})) \\
&\leq (1 + o(1)) \left\| \boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:} \right\|,
\end{aligned}$$

where the second inequality follows by Lemma 6, and the last inequality follows by the balanced assumption (34) and Lemma 5.

Hence, we conclude that

$$\min_{(i,j):z(i)\neq z(j)} \left\| \boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:} \right\| \gtrsim \min_{(i,j):z(i)\neq z(j)} \left\| \boldsymbol{S}^s_{z_1(i):} - \boldsymbol{S}^s_{z_1(j):} \right\| \gtrsim \Delta_{\min}.$$

□

Lemma 8 (Singular value of weighted membership matrix)**.** Under the assumption that $\min_{i\in[p]} \theta(i) \geq c_0 > 0$, we have the minimal singular value of $\boldsymbol{\Theta M}$ bounded as

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

Proof of Lemma 8. Note that

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

By the definition of singular value, we have

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

Given that $\min_{i\in[p]} \theta(i) \geq c_0$, we have

$$\min_{a\in[r]} \left\| \boldsymbol{\theta}_{z^{-1}(a)} \right\|^2 \geq c_0^2 \min_{a\in[r]} |z^{-1}(a)| \gtrsim \frac{p}{r},$$

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

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

Therefore, we complete the proof of Lemma 8. □

Lemma 9 (Singular-value-gap-free tensor estimation error bound)**.** Let an order-K tensor $\mathcal{A} = \mathcal{X} + \mathcal{Z} \in \mathbb{R}^{p\times\cdots\times p}$, and $\mathcal{X}$ has tucker rank $(r, \ldots r)$ and $\mathcal{Z}$ has independent sub-Gaussian entries with parameter σ^2. Let $\hat{\mathcal{X}}$ denote

the two-step estimated tensor in line 2 of Sub-Algorithm 1 in the main paper. Then with probability at least $1 - C\exp(-cp)$, we have

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

Proof. See Proposition 1 in Han et al. (2020). □

Lemma 10 (Upper bound of misclassification error)**.** Let $z : [p] \mapsto [r]$ be a clustering assignment such that $|z^{-1}(a)| \asymp p/r$ for all $a \in [r]$. Let node i corresponds to a vector $\boldsymbol{x}_i = \theta(i)\boldsymbol{v}_{z(i)} \in \mathbb{R}^d$ where $\{\boldsymbol{v}_a\}_{a=1}^r$ 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 such that $\frac{\max_{a\in[r]}\left\|\boldsymbol{\theta}_{z^{-1}(a)}\right\|^2}{\min_{a\in[r]}\left\|\boldsymbol{\theta}_{z^{-1}(a)}\right\|^2} = 1 + o(1)$. Consider an arbitrary estimate $\hat{z}$ with $\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{v}}_{\hat{z}(i)}$. Then, if

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

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

$$S_0 = \{i : \hat{z}(i) = 0\} \quad \text{and} \quad S = \{i \in S_0^c : \|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\| \geq c\}.$$

Proof of Lemma 10. For each cluster $u \in [r]$, we use C_u to collect the subset of points for which the estimated and true positions $\hat{\boldsymbol{x}}_i, \boldsymbol{x}_i$ are within distance c. Specifically, define For each cluster $u \in [r]$, we define

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

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

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

Then, we have

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

Thus, we only need to bound $\sum_{i\in\cup_{u\in R_3} C_u} \theta(i)^2$ to finish the proof. Note that we have $|R_3| \leq |R_1|$ by Lemma 6 in (Gao et al., 2018). We have

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

where the second inequality follows the conclusion $|R_3| \leq |R_1|$ from the proof of Lemma 6 in (Gao et al., 2018), and the last inequality holds by the balanced assumption on $\boldsymbol{\theta}$ when p is large enough, and the fact that $\cup_{u\in R_1} z^{-1}(u) \subset S$.

□

PROOF OF THEOREM 5

Proof of Theorem 5. Recall that we prove Theorem 5 under the dTBM 1 with parameters $(z, \mathcal{S}, \boldsymbol{\theta})$. We drop the subscript k in the matricizations $\boldsymbol{M}_k, \boldsymbol{S}_k, \boldsymbol{X}_k$. Without loss of generality, we assume that the noise variance $\sigma = 1$ and the identity permutation minimizes the initial misclassification error, i.e., $\pi^{(0)} = \arg\min_{\pi\in\Pi} \sum_{i\in[p]} \mathbb{1}\left\{z^{(0)}(i) \neq \pi \circ z(i)\right\}$ and $\pi^{(0)}(a) = a$ for all $a \in [r]$.

g) Step 1:: We first introduce additional notations and the necessary condition to complete the proof. We will verify that the conditions hold in our context under high probability in the last step of the proof.

Notations.

1) Norms. Let $\|\boldsymbol{A}\|_\sigma$ denote the spectral norm of matrix $\boldsymbol{A}$, which is equal to the maximal singular value of $\boldsymbol{A}$; let $\|\boldsymbol{A}\|_F$ denote the Frobenius norm of matrix $\boldsymbol{A}$.

2) Projection. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, let $\mathrm{Proj}(\boldsymbol{v}) \in \mathbb{R}^{d\times d}$ denote the projection matrix to $\boldsymbol{v}$. Then $\boldsymbol{I}_d - \mathrm{Proj}(\boldsymbol{v})$ is the projection matrix to the orthogonal complement $\boldsymbol{v}^\perp$.

3) Normalized membership matrix

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

and dual normalized membership matrix

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

4) Estimator of $\mathcal{S}$ in the t-th iteration, denoted $\mathcal{S}^{(t)}$ and the oracle estimator of $\mathcal{S}$ given true assignment z, denoted $\tilde{\mathcal{S}}$,

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

5) Matricization of tensors

$$\boldsymbol{S} = \mathrm{Mat}_1(\mathcal{S}), \quad \boldsymbol{S}^{(t)} = \mathrm{Mat}_1(\mathcal{S}^{(t)}), \quad \tilde{\boldsymbol{S}} = \mathrm{Mat}_1(\tilde{\mathcal{S}})$$
$$\boldsymbol{Y} = \mathrm{Mat}_1(\mathcal{Y}), \quad \boldsymbol{X} = \mathrm{Mat}_1(\mathcal{X}), \quad \boldsymbol{E} = \mathrm{Mat}_1(\mathcal{E}).$$

6) The angle-based misclassification loss in the t-th iteration

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

and the oracle loss

$$\begin{aligned}\xi = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}\mathbb{1}\Big\{&\left\langle\boldsymbol{E}_{i:}\boldsymbol{V}, \left[\tilde{\boldsymbol{S}}_{z(i):}\right]^s - \left[\tilde{\boldsymbol{S}}_{b:}\right]^s\right\rangle \\ &\le -\frac{\theta(i)m}{4}\left\|\left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s\right\|^2\Big\}\left\|\left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s\right\|^2,\end{aligned}$$

where m is a positive constant related to c_3 and is defined in (40).

Condition 1. Let $\mathbb{O}_{p,r}$ denote the collection of all the p-by-r matrices with orthonormal columns. For all $a \in [r]$, we have

$$\|\boldsymbol{E}\boldsymbol{V}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}}\left(p^{1/2} + r^{(K-1)/2}\right), \quad \|\boldsymbol{E}\boldsymbol{V}\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}}, \quad \left\|\boldsymbol{W}_{a:}^T\boldsymbol{E}\boldsymbol{V}\right\| \lesssim \frac{r^K}{p^{K/2}}, \tag{35}$$

$$\sup_{\boldsymbol{U}_k\in\mathbb{O}_{p,r},k=2,\ldots,K}\|\boldsymbol{E}(\boldsymbol{U}_2\otimes\cdots\otimes\boldsymbol{U}_K)\|_\sigma\lesssim\left(\sqrt{r^{K-1}}+K\sqrt{pr}\right),\tag{36}$$

$$\sup_{\boldsymbol{U}_k\in\mathbb{O}_{p,r},k=2,\ldots,K}\|\boldsymbol{E}(\boldsymbol{U}_2\otimes\cdots\otimes\boldsymbol{U}_K)\|_F\lesssim\left(\sqrt{pr^{K-1}}+K\sqrt{pr}\right),\tag{37}$$

$$\xi\lesssim\exp\left(-\frac{\Delta_{\min}^2p^{K-1}}{r^{K-1}}\right),\tag{38}$$

$$L^{(t)}\lesssim\frac{\Delta_{\min}^2}{r\log p},\quad\text{for}\quad t=0,1,\ldots,T.\tag{39}$$

Particularly, the inequality (35) holds by replacing $\boldsymbol{V}$ to $\boldsymbol{V}^{(t)}$ and $\boldsymbol{W}_{:a}$ to $\boldsymbol{W}_{:a}^{(t),T}$ when initialization condition (39) holds.

h) Step 2:: Next, we derive upper bound of $L^{(t+1)}$ for $t=0,1,\ldots,T-1$. By Sub-Algorithm 2, we update the assignment in t-th iteration as

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

following the fact that $\mathrm{Mat}_1(\mathcal{Y}^{\mathrm{d}})=\boldsymbol{Y}\boldsymbol{V}^{(t)}$, where $\mathcal{Y}^{\mathrm{d}}$ is the reduced tensor defined in Sub-Algorithm 2. Then the event $z^{(t+1)}(i)=b$ implies

$$\left\|\left[\boldsymbol{Y}_{j:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{S}_{b:}^{(t)}\right]^s\right\|^2\le\left\|\left[\boldsymbol{Y}_{j:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{S}_{z(j):}^{(t)}\right]^s\right\|^2.$$

Arranging the terms, we obtain the decomposition

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

where

$$\begin{aligned}
F_{ib}^{(t)}&=2\left\langle\boldsymbol{E}_{i:}\boldsymbol{V}^{(t)},\left(\left[\tilde{\boldsymbol{S}}_{z(i):}\right]^s-\left[\boldsymbol{S}_{z(i):}^{(t)}\right]^s\right)-\left(\left[\tilde{\boldsymbol{S}}_{b:}\right]^s-\left[\boldsymbol{S}_{b:}^{(t)}\right]^s\right)\right\rangle\\
&\quad+2\left\langle\boldsymbol{E}_{i:}\left(\boldsymbol{V}-\boldsymbol{V}^{(t)}\right),\left[\tilde{\boldsymbol{S}}_{z(i):}\right]^s-\left[\tilde{\boldsymbol{S}}_{b:}\right]^s\right\rangle,\\
G_{ib}^{(t)}&=\left(\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{S}_{z(i):}^{(t)}\right]^s\right\|^2-\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}\right]^s\right\|^2\right)\\
&\quad-\left(\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{S}_{b:}^{(t)}\right]^s\right\|^2-\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}\right]^s\right\|^2\right),\\
H_{ib}^{(t)}&=\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}\right]^s\right\|^2-\left\|\left[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right]^s-\left[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}\right]^s\right\|^2+\left\|\left[\boldsymbol{S}_{z(i):}\right]^s-\left[\boldsymbol{S}_{b:}\right]^s\right\|^2.
\end{aligned}$$

Then, we have the following upper bound

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

Note that

$$\left\|\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\right\|=\left\|\boldsymbol{X}_{i:}\boldsymbol{W}^{(t),\otimes^{K-1}}\right\|$$

$$
\begin{aligned}
&\geq \left\| \boldsymbol{S}_{i:}(\boldsymbol{\Theta}\boldsymbol{M})^{\otimes(K-1),T} \right\| \lambda_r^{K-1}(\boldsymbol{W}^{(t)}) \\
&\geq \theta(i) \left\| \boldsymbol{S}_{z(i):} \right\| \lambda_r^{K-1}(\boldsymbol{\Theta}\boldsymbol{M}) \lambda_r^{K-1}(\boldsymbol{W}^{(t)}) \\
&\geq \theta(i) m, \qquad (40)
\end{aligned}
$$

where the first two inequalities follow by the property of eigenvalues, the last inequality follows by Lemmas 8 and 12, the assumption that $\min_{a\in[r]} \left\| \boldsymbol{S}_{z(i):} \right\| \geq c_3$, and $m > 0$ is a positive constant related to c_3. Plugging the lower bound of $\left\| \boldsymbol{X}_{i:}\boldsymbol{V}^{(t)} \right\|$ (40) into the inequality 0h, we obtain that

$$
\mathbb{1}\left\{ z^{(t+1)}(i) = b \right\} \leq I_1 + I_2, \qquad (41)
$$

where

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

On one hand, the event A_{ib} is the critical component in ξ, where

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

We will investigate $\mathbb{P}(A_{ib} = 1)$ when verifying the Condition 1.

On the other hand, the terms $F_{ib}^{(t)}, G_{ib}^{(t)}$, and $H_{ib}^{(t)}$ in B_{ib} describe the misclassification error in t-th iteration, which intuitively can be represented by $L^{(t)}$. Specifically, under Condition 1, Lemma 14 and Han et al. (2020, Step 4, Proof of Theorem 2) imply that

$$
\frac{1}{p} \sum_{i\in[p]} \theta(i) \sum_{b\in[r]/z(i)} \left\| \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\|^2 B_{ib} \leq c' L^{(t+1)} + \rho L^{(t)},
$$

where $c', \rho \in (0,1)$ are two small positive constants. Hence, by the upper bound (41) and the defintion of $L^{(t+1)}$, we obtain the upper bound for the misclassification loss in the $t+1$-th iteration

$$
L^{(t+1)} \leq M\xi + \rho L^{(t)}, \qquad (42)
$$

where $M \geq 1$ and ρ is called the contraction parameter.

i) Step 3:: Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (35), (36), and (37) describes the property of the noise tensor $\mathcal{E}$, and the readers can find the proof directly in Han et al. (2020, Step 5, Proof of Theorem 2).

Now, we verify the oracle loss condition (38). Recall the definition of ξ

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

Let $e_i = \boldsymbol{E}_{i:}\boldsymbol{V}$ denote the aggregated noise vector for $i \in [p]$, and e_i are independent mean-zero sub-Gaussian vector in $\mathbb{R}^{r^{K-1}}$. Also, each coordinate in e_i is also independent mean-zero sub-Gaussian variable with sub-Gaussian norm upper bounded by $C\sqrt{r^{K-1}/p^{K-1}}$ with some positive constant C. Note that probability

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

where

$$
P_1 = \mathbb{P}\left(\left\langle e_i, \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\rangle \leq -\frac{m}{8} \left\| \left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s \right\|^2 \right),
$$

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

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

$$P_1 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\left\|\left[\boldsymbol{S}_{z(j):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^2\right). \tag{43}$$

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

$$P_2 \leq P_{21} + P_{22}, \tag{44}$$

where with some positive constant t

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

For P_{21}, note that the term $\boldsymbol{W}^T_{:z(i)}\boldsymbol{E}\boldsymbol{V} = \frac{\sum_{j\neq i, j\in[p]}\mathbb{1}\{z(j)=z(i)\}e_j}{\sum_{j\in[p]}\mathbb{1}\{z(j)=z(i)\}}$ is a sub-Gaussian vector with sub-Gaussian norm bounded by $C_1\sqrt{r^K/p^K}$. Then, we have

$$\begin{aligned} P_{21} &\leq \mathbb{P}\left(t\left\|\boldsymbol{S}_{z(i):}\right\| < \left\|\tilde{\boldsymbol{S}}_{z(i):} - \boldsymbol{S}_{z(i):}\right\|\right) \\ &\leq \mathbb{P}\left(c_3 t < \left\|\boldsymbol{W}^T_{:z(i)}\boldsymbol{E}\boldsymbol{V}\right\|\right) \\ &\lesssim \exp\left(-\frac{p^K t^2}{r^K}\right), \end{aligned} \tag{45}$$

where the first inequality follows by the basic inequality in Lemma 5, the second inequality follows by the assumption that $\min_{a\in[r]}\left\|\boldsymbol{S}_{z(i):}\right\| \geq c_3$, and the last inequality follows by the Bernstein inequality.

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

$$P_{22} \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}t^2}\left\|\left[\boldsymbol{S}_{z(j):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^4\right). \tag{46}$$

Take $t = \left\|\left[\boldsymbol{S}_{z(j):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|$ in P_{21} and P_{22}, and plug the inequalities (45) and (46) into to the upper bound for P_2 (44). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\left\|\left[\boldsymbol{S}_{z(j):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^2\right). \tag{47}$$

Therefore, combining the upper bounds (43) and (47), we have

$$\mathbb{P}\left(\left\langle e_i, \left[\tilde{\boldsymbol{S}}_{z(i):}\right]^s - \left[\tilde{\boldsymbol{S}}_{b:}\right]^s\right\rangle \leq -\frac{m}{4}\left\|\left[\boldsymbol{S}_{z(i):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^2\right) \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\left\|\left[\boldsymbol{S}_{z(j):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^2\right).$$

Hence, we have

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

$$\lesssim \frac{1}{p}\sum_{i\in[p]}\sqrt{p}\left\|\left[\boldsymbol{S}_{z(i):}\right]^s-[\boldsymbol{S}_{b:}]^s\right\|^2\exp\left(-\frac{p^{K-1}}{r^{K-1}}\left\|\left[\boldsymbol{S}_{z(j):}\right]^s-[\boldsymbol{S}_{b:}]^s\right\|^2\right)$$
$$\lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right),$$

where the first inequality follows by the fact that $\max_{i\in[p]}\theta(i)\lesssim\sqrt{p}$.

By Markov's inequality, we have

$$\mathbb{P}\left(\xi\lesssim\mathbb{E}[\xi]+\exp\left(-\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right)\right)\geq 1-C_2\exp\left(-\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right),$$

and thus the condition (38) holds with probability at least $1-C_2\exp\left(-\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right)$ for some constant $C_2>0$.

Finally, we verify the bounded loss condition (39) by induction. By Corollary 1, we have

$$L^{(0)}\lesssim\frac{\Delta_{\min}^2}{r\log p}$$

when p is large enough by Theorem 4, and thus condition (39) holds for $t=0$. Assume the condition (39) also holds for all $t\leq t_0$. Then, by the decomposition (42), we have

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

where the second inequality follows by condition (38) and the last inequality follows by the assumption that $\Delta_{\min}^2\gtrsim p^{-K/2}$. Thus, the condition (39) holds for t_0+1, and the condition (39) is proved by induction. □

j) Useful Definitions and Lemmas for the Proof of Theorem 5:

Lemma 11 (Misclassification error and loss)**.** Define the misclustering error in the t-th iteration as $\ell^{(t)}=\ell(z^{(t)},z)$. We have

$$\ell^{(t)}\lesssim\frac{1}{p}\sum_{i\in[p]}\theta(i)\mathbb{1}\left\{z^{(t)}(i)\neq z(i)\right\}\leq\frac{L^{(t)}}{\Delta_{\min}^2}.$$

Lemma 12 (Membership matrices)**.** Suppose the condition (39) holds. Then, for any $a\in[r]$, we have $|\left(z^{(t)}\right)^{-1}(a)|\asymp p/r$. Moreover, we have

$$\lambda_r(\boldsymbol{M})\asymp\|\boldsymbol{M}\|_\sigma\asymp\sqrt{p/r},\quad\lambda_r(\boldsymbol{W})\asymp\|\boldsymbol{W}\|_\sigma\asymp\sqrt{r/p}.\tag{48}$$

The inequalities (48) also hold by replacing $\boldsymbol{M}$ and $\boldsymbol{W}$ to $\boldsymbol{M}^{(t)}$ and $\boldsymbol{W}^{(t)}$ respectively. Further, we have

$$\lambda_{\min}(\boldsymbol{W}\boldsymbol{W}^T)\asymp\left\|\boldsymbol{W}\boldsymbol{W}^T\right\|_\sigma\asymp r/p,\tag{49}$$

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

Proof of Lemma 12. See Han et al. (2020, Proof of Lemma 4) for inequality (48).

For inequality (49), note that

$$\lambda(\boldsymbol{W}\boldsymbol{W}^T)=\sqrt{\text{eigen}(\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{W}^T)}\asymp\sqrt{\frac{r}{p}\text{eigen}(\boldsymbol{W}\boldsymbol{W}^T)}=\sqrt{\frac{r}{p}\lambda^2(\boldsymbol{W})}\asymp\frac{r}{p},$$

where the first inequality follows the fact that $\boldsymbol{W}^T\boldsymbol{W}$ is a diagonal matrix with elements of order $\frac{r}{p}$, and the second equation follows by the definition of singular value. □

Lemma 13 (Relationship between error and intermediate parameters)**.** Under the Condition 1, we have

$$\|\boldsymbol{V}-\boldsymbol{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}}\frac{r(K-1)}{\Delta_{\min}^2}L^{(t)}, \tag{50}$$

$$\|\boldsymbol{E}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1}+pr)}{p^{K-1}}}\frac{r(K-1)}{\Delta_{\min}^2}L^{(t)}. \tag{51}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s\| \lesssim \frac{rL^{(t)}}{\Delta_{\min}}+\sqrt{\frac{r^{2K}+pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}} \tag{52}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\| \lesssim \sqrt{\frac{rr^{2K}+pr^{K+2}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}+\frac{rL^{(t)}}{\Delta_{\min}} \tag{53}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\| \lesssim \frac{rL^{(t)}}{\Delta_{\min}}+\sqrt{\frac{rr^{2K}+pr^{K+2}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}, \tag{54}$$

In addition, the inequality (53) also holds by replacing $\boldsymbol{W}_{:b}^{(t)}$ to $\boldsymbol{W}_{:b}$.

Proof of Lemma 13. We follow and use several intermediate conclusions in Han et al. (2020, Proof of Lemma 5). We prove each inequality separately.

1) Inequality (50). By Han et al. (2020, Proof of Lemma 5), we have

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

Then, we complete the proof of inequality (50) by applying Lemma 11 to the above inequality.

2) Inequality (51). By Han et al. (2020, Proof of Lemma 5), we have

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

Also, we complete the proof of inequality (50) by applying Lemma 11 to the above inequality.

3) Inequality (52). We upper bound the desired quantity by triangle inequality as

$$\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s\| \le I_1+I_2+I_3,$$

where

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

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

For I_1, we further bound I_1 by triangle inequality as

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

where

$$I_{11} = \left\| \frac{\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}\|} - \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} \right\| \quad \text{and} \quad I_{12} = \left\| \frac{\boldsymbol{W}_{:b}^T \boldsymbol{E}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}\|} - \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} \right\|$$

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

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

By Lemma 12, we have $\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} \gtrsim p/r$. Then, we have

$$\sum_{a \neq b, a, b \in [r]} D_{ab} \lesssim \frac{r}{p} \sum_{i:\ z^{(t)}(i) \neq z(i)} \theta(i) \lesssim \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \frac{1}{\log p}, \tag{55}$$

and for all $b \in [r]$

$$D_{bb} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\left\{z(i) = z^{(t)}(i) = b\right\}}{\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \geq \frac{c_0(\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} - p\ell^{(t)})}{\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \gtrsim 1 - \frac{1}{\log p}, \tag{56}$$

under the inequality (39) in Condition 1. By the definition of $\boldsymbol{W}, \boldsymbol{W}^{(t)}, \boldsymbol{V}$, we have

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

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

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

where C is a positive constant, and the last inequality holds when p is large enough following the constraint of $\|\boldsymbol{S}_{b:}\|$ in parameter space (2) and the bounds of $\boldsymbol{D}$ (55) and (56).

The positive inner product between $\boldsymbol{S}_{b:}$ and $D_{bb}\boldsymbol{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab}\boldsymbol{S}_{a:}$ 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

$$\begin{aligned}
\|[D_{bb}\boldsymbol{S}_{b:} + \sum_{a \neq b, a \in [r]} D_{ab}\boldsymbol{S}_{a:}] \sin \alpha\| &= \|(\boldsymbol{I}_d - \text{Proj}(\boldsymbol{S}_{b:})) \sum_{a \neq b, a \in [r]} D_{ab}\boldsymbol{S}_{a:}\| \\
&= \sum_{a \neq b, a \in [r]} D_{ab} \|(\boldsymbol{I}_d - \text{Proj}(\boldsymbol{S}_{b:}))\boldsymbol{S}_{a:}\| \\
&= \sum_{a \neq b, a \in [r]} D_{ab} \|\boldsymbol{S}_{a:} \sin(\boldsymbol{S}_{b:}, \boldsymbol{S}_{a:})\| \\
&\leq \sum_{a \neq b, a \in [r]} D_{ab} \|\boldsymbol{S}_{a:}\| \|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\|,
\end{aligned} \tag{57}$$

where the last inequality follows by Lemma 5. Also, note that with bounds (55) and (56), when p is large

enough we have

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\| = \|D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}\| \geq D_{bb}\|\boldsymbol{S}_{b:}\| - \sum_{a\neq b, a\in[r]} D_{ab}\|\boldsymbol{S}_{a:}\| \geq C_1, \tag{58}$$

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

$$\begin{aligned} I_{11} &\leq \sqrt{2}\sin\alpha \\ &= \frac{\|[D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}]\sin\alpha\|}{\|D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}\|} \\ &\leq \frac{1}{C_1}\sum_{a\neq b, a\in[r]} D_{ab}\|\boldsymbol{S}_{a:}\|\,\|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\| \\ &\lesssim \frac{r}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}\mathbb{1}\left\{z^{(t)}(i) = b\right\}\|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\| \\ &= \frac{rL^{(t)}}{\Delta_{\min}}, \end{aligned} \tag{59}$$

where the second inequality follows by (57) and (58), and the last inequality follows by the definition of D_a and the constraint of $\|\boldsymbol{S}_{b:}\|$ in parameter space (2).

Consider I_{12}. By triangle inequality, we have

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

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

$$\|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}\| \lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{60}$$

Notice that

$$\|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}\| \leq \|\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T}\|\,\|\boldsymbol{X}\boldsymbol{V}\|_F \lesssim \frac{r^{3/2}L^{(t)}}{\sqrt{p}\Delta_{\min}^2}\|\boldsymbol{S}\|\|\boldsymbol{\Theta}\boldsymbol{M}\|_\sigma \lesssim \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}, \tag{61}$$

where the second inequality follows by Han et al. (2020, Proof of Lemma 5) and the last inequality follows by Lemma 8 and (39) in Condition 1. Note that $\left\|\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}\right\| = \|\boldsymbol{S}_{b:}\| \geq c_3$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\| \geq C_1$ by inequality (58). Therefore, we have

$$\begin{aligned} I_{12} &\lesssim \|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}\| + \|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}\|\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}\sqrt{\frac{r^{2K}}{p^K}} \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}, \end{aligned} \tag{62}$$

where second inequality follows the inequalities (60), (61), and (35) in Condition 1.

Hence, combining inequalities (59) and (62), we have

$$I_1 \lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{63}$$

For I_2 and I_3, recall that $\left\|\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}\right\| = \|\boldsymbol{S}_{b:}\| \geq c_3$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\| \geq C_1$ by inequality (58). By triangle

inequality and (35) in Condition 1, we have

$$I_2 \leq \frac{\|\boldsymbol{W}_{:b}^T \boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^T \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{64}$$

and

$$I_3 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{65}$$

Therefore, combining the inequalities (63), (64), and (65), we finish the proof of inequality (52).

4) Inequality (53). Here we only show the detailed proof of inequality (53) with $\boldsymbol{W}_{:b}^{(t)}$. The proof also holds by replacing $\boldsymbol{W}_{:b}^{(t)}$ to $\boldsymbol{W}_{:b}$, and we omit the repeated procedures.

We upper bound the desired quantity by triangle inequality as

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

where

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

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

For J_1, by triangle inequality and Lemma 5, we have

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

where

$$J_{11} = \left\| \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} - \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^{(t)}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^{(t)}\|} \right\|, \quad J_{12} = \left\| \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} - \frac{\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}^{(t)}}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^{(t)}\|} \right\|.$$

Consider J_{11}. Define the matrix $\boldsymbol{V}^k \coloneqq \boldsymbol{W}^{\otimes(k-1)} \otimes \boldsymbol{W}^{(t),\otimes(K-k)}$ for $k = 2, \ldots, K-1$, and denote $\boldsymbol{V}^1 = \boldsymbol{V}^{(t)}, \boldsymbol{V}^K = \boldsymbol{V}$. Then, define the quantity

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

for $k = 1, \ldots, K-1$. Let β_k denote the angle between $\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^k$ and $\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^{k+1}$. With the same idea to prove I_{11}, we bound J_{11}^k by considering the trigonometric function of β_k.

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

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

where

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

Recall the definition of confusion matrix $\boldsymbol{D} = \boldsymbol{M}^T\boldsymbol{\Theta}^T\boldsymbol{W}^{(t)} = [\![D_{ab}]\!] \in \mathbb{R}^{r\times r}$. Then, we have

$$\begin{aligned}\left\langle \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^k, \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{k+1}\right\rangle &= \langle \mathrm{Mat}_{k+1}(\mathcal{T}_k), \mathrm{Mat}_{k+1}(\mathcal{T}_{k+1})\rangle \\ &= \left\langle \boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^k, \boldsymbol{S}\boldsymbol{Z}^k\right\rangle \\ &= \sum_{b\in[r]}\left(D_{bb}\|\boldsymbol{S}_{b:}\boldsymbol{Z}^k\|^2 + \sum_{a\neq b, a\in[r]} D_{ab}\left\langle \boldsymbol{S}_{a:}\boldsymbol{Z}^k, \boldsymbol{S}_{b:}\boldsymbol{Z}^k\right\rangle\right) \\ &\gtrsim (1-\log p^{-1})\min_{a\in[r]}\|\boldsymbol{S}_{a:}\boldsymbol{Z}^k\|^2 - \log p^{-1}\max_{a\in[r]}\|\boldsymbol{S}_{a:}\boldsymbol{Z}^k\|^2, \end{aligned} \tag{66}$$

where $\boldsymbol{Z}^k = \boldsymbol{D}_{:b}\otimes \boldsymbol{I}_r^{\otimes(k-1)}\otimes \boldsymbol{D}^{\otimes(K-k-1)}$, the equations follow by the tensor algebra and the definitions, and the last inequality follows by the bounds of $\boldsymbol{D}$ (55) and (56).

Note that

$$\|\boldsymbol{D}\|_\sigma \le \|\boldsymbol{D}\|_F \le \sqrt{\sum_{b\in[r]} D_{bb}^2 + (\sum_{a\neq b, a,b\in[r]} D_{ab})^2} \lesssim \sqrt{r+\log^2 p^{-1}} \lesssim 1, \tag{67}$$

where the third inequality follows by 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

$$\lambda_r(\boldsymbol{D}) \ge \lambda_r(\boldsymbol{W}^{(t)})\lambda_r(\boldsymbol{\Theta}\boldsymbol{M}) \gtrsim 1, \tag{68}$$

following the Lemma 8 and Lemma 12. Then, for all $k\in[K]$, we have

$$1 \lesssim \|\boldsymbol{D}_{:b}\|\,\lambda_r(\boldsymbol{D})^{K-k-1} \le \lambda_{r^{K-2}}(\boldsymbol{Z}^k) \le \|\boldsymbol{Z}^k\|_\sigma \le \|\boldsymbol{D}_{:b}\|\,\|\boldsymbol{D}\|_\sigma^{K-k-1} \lesssim 1. \tag{69}$$

Thus, we have bounds

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

Hence, when p is large enough, the inner product (66) is positive, which implies $\beta_k\in[0,\pi/2)$ and thus $2\sin\frac{\beta_k}{2} \le \sqrt{2}\sin\beta_k$.

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

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

where

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

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

On one hand, recall the definitions for any cluster assignment $\bar{z}$

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

Then, we have $\mathrm{diag}(\boldsymbol{p}(z^{(t)}))\boldsymbol{D} = \mathrm{diag}(\boldsymbol{p}_{\boldsymbol{\theta}}(z^{(t)}))\tilde{\boldsymbol{D}}$. By Condition 1 and Lemma 11, we have $\ell^{(t)} \le \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \log^{-1}(p)$. Then, by the stability assumption, we have

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

Note that $\sin(\boldsymbol{a},\boldsymbol{b}) = \min_{c\in\mathbb{R}} \frac{\|\boldsymbol{a}-c\boldsymbol{b}\|}{\|\boldsymbol{a}\|}$ for vectors $\boldsymbol{a},\boldsymbol{b}$ of same dimension, and let $c_0 = \arg\min_{c\in\mathbb{R}} \frac{\|\boldsymbol{p}(z^{(t)})-c\boldsymbol{p}_{\boldsymbol{\theta}}(z^{(t)})\|}{\|\boldsymbol{p}(z^{(t)})\|}$. We have

$$\begin{aligned}
\min_{c\in\mathbb{R}}\|\boldsymbol{D}-c\tilde{\boldsymbol{D}}\|_F &\le \|\boldsymbol{I}_r - c_0\mathrm{diag}(\boldsymbol{p}(z^{(t)}))\mathrm{diag}^{-1}(\boldsymbol{p}_{\boldsymbol{\theta}}(z^{(t)}))\|_F\|\boldsymbol{D}\|_F \\
&\lesssim \frac{\|\boldsymbol{p}(z^{(t)}) - c_0\boldsymbol{p}_{\boldsymbol{\theta}}(z^{(t)})\|}{\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{(t),-1}(a)}\|_1} \\
&\lesssim \frac{\|\boldsymbol{p}(z^{(t)})\|}{\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{(t),-1}(a)}\|_1}\sin(\boldsymbol{p}(z^{(t)}),\boldsymbol{p}_{\boldsymbol{\theta}}(z^{(t)})) \\
&\lesssim \frac{L^{(t)}}{\Delta_{\min}},
\end{aligned}$$

where the last inequality follows Lemma 12 that $\|\boldsymbol{p}(z^{(t)})\| \lesssim p$ and $\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{(t),-1}(a)}\|_1 \gtrsim p$. By the geometry property of trigonometric function, we have

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

where the second inequality follows by the singular property of $\boldsymbol{D}$ in (67) and (68).

On the other hand, let $\boldsymbol{C} = \mathrm{diag}(\{\|\boldsymbol{S}_{a:}\|\}_{a\in[r]})$. We have

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

where the third inequality follows by the singular property of $\boldsymbol{D}$ and the boundedness of $\boldsymbol{S}$, and the fourth inequality follows by the definition of $\tilde{\boldsymbol{D}}$, boundedness of $\boldsymbol{S}$, lower bound of $\boldsymbol{\theta}$, and the singular property of $\boldsymbol{Z}^k$ in inequality (69).

Therefore, we have shown the

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

Finally, by triangle inequality, we obtain

$$J_{11} \le \sum_{k=1}^{K-1} J_{11}^k \lesssim (K-1)\frac{rL^{(t)}}{\Delta_{\min}}. \tag{70}$$

Consider J_{12}. By triangle inequality, we have

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

Note that

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\| = \|\boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^1\| \geq \lambda_r(\boldsymbol{D})\,\|\boldsymbol{S}\|\,\lambda_{r^{K-2}}(\boldsymbol{Z}^1) \gtrsim 1, \tag{71}$$

where inequality follows by the bounds (68) and (69).

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

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\| \lesssim \sqrt{\frac{r^{2K+1}+pr^{2+K}}{p^K}}\frac{(K-1)\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{72}$$

Notice that

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

where the first inequality follows by the tensor algebra in equation , and the second inequality follows by the fact that $\boldsymbol{I}=\boldsymbol{W}^T\boldsymbol{\Theta}\boldsymbol{M}$, and the last inequality follows by Han et al. (2020, Proof of Lemma 5). Thus, we have

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\| \leq \|\boldsymbol{W}_{:b}^{(t),T}\|\sum_{k=1}^{K-1}\|\boldsymbol{X}(\boldsymbol{V}^k-\boldsymbol{V}^{k+1})\|_F \lesssim \frac{(K-1)\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}}. \tag{73}$$

Note that $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|$ are lower bounded by inequalities (58) and (71), respectively. We have

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

where the second inequality follows by inequalities (72), (73), and the inequality (35) in Condition 1.

For J_2 and J_3, recall that $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|$ are lower bounded by inequalities (58) and (71), respectively. By triangle inequality and inequality (35) in Condition 1, we have

$$J_2 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{74}$$

and

$$J_3 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}^{(t)}\|}{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{75}$$

Therefore, combining the inequalities (70), (74), and (75), we finish the proof of inequality (53).

5) Inequality (54). By triangle inequality, we upper bound the desired quantity

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

following the inequalities (52) and (53). Therefore, we finish the proof of inequality (54).

□

Lemma 14 (Upper bound for $F_{ib}^{(t)}, G_{ib}^{(t)}$ and $H_{ib}^{(t)}$)**.** Under the Condition 1, we have

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left(F_{ib}^{(t)}\right)^2}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2}\|\boldsymbol{E}_{i:}\boldsymbol{V}\|^2 + \left(1+\frac{rL^{(t)}}{\Delta_{\min}^2}\right)\|\boldsymbol{E}_{i:}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\|^2, \tag{76}$$

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left(G_{ib}^{(t)}\right)^2}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \lesssim \left(\Delta_{\min}^2 + L^{(t)}\right), \tag{77}$$

where c is a small positive constant near to 0.

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left|H_{ib}^{(t)}\right|}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \le \frac{1}{4}. \tag{78}$$

Proof of Lemma 14. We prove Lemma 14 by the following order:

1) Upper bound for $F_{ib}^{(t)}$ (76). Recall the definition of $F_{ib}^{(t)}$

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

By Cauchy-Schwartz inequality, we have

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

Note that for any $a \in [r]$

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

$$\lesssim rL^{(t)},$$

where the second inequality follows the inequalities (52) and (53) in Lemma 13, the third inequality follows the initialization condition (39) in Condition 1, and the last inequality follows the assumption that $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$.

Also note that

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

where the second inequality follows by Lemma 5, and the last inequality follows by the assumptions on $\|\boldsymbol{S}_{a:}\|$ in the parameter space (2), the inequality (35) Condition 1 and the assumption $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$.

Therefore, we finish the proof of inequality (76).

2) Upper bound for $G_{ib}^{(t)}$ (77). Recall the definition of $G_{ib}^{(t)}$ and rearrange the terms

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

where

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

For G_1, we have

$$\begin{aligned}
|G_1|^2 &\le \left|\|[\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2 - \|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2\right|^2 \\
&\le \max_{a\in[r]}\|[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^4 \\
&\lesssim \frac{r^4}{\Delta_{\min}^4}(L^{(t)})^4 + \frac{r^2r^{4K} + p^2r^{2K+4}}{p^{2K}}\frac{(L^{(t)})^2}{\Delta_{\min}^4} \\
&\le c\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right),
\end{aligned} \tag{79}$$

where the third inequality follows by the inequality (54) in Lemma 12 and the last inequality follows by the assumption that $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$.

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

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

$$\lesssim \frac{r^{2K-1}+Kpr^{K+1}}{p^K}\left(\frac{r^2}{\Delta_{\min}^2}(L^{(t)})^2+\frac{rr^{2K}+pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\right)$$
$$\leq c\Delta_{\min}^2 L^{(t)}, \tag{80}$$

where the second inequality follows by Lemma 5, the third inequality follows by the bound (71), inequality (36) in Condition 1, inequality (54) in Lemma 13, and the last inequality follows by the assumption $\Delta_{\min}^2 \geq p^{-K/2}\log p$.

For G_3, note that by triangle inequality

$$\begin{aligned}\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2 &\leq \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + 2\max_{a\in[r]}\|[\boldsymbol{W}_{:a}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^T\boldsymbol{X}\boldsymbol{V}]^s\|^2\\ &\lesssim \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2},\end{aligned} \tag{81}$$

where the last inequality follows by the derivation of J_{11} in the proof of Lemma 13.

Then we have

$$\begin{aligned}|G_3|^2 &\leq 2\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\max_{a\in[r]}\|[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\\ &\leq 2\left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2 + \|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2\right)\\ &\quad\cdot\max_{a\in[r]}\left\|[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\right\|^2\\ &\lesssim \left(\|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2}\right)\left(\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2}+\frac{rr^{2K}+pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\right) + c\Delta_{\min}^2 L^{(t)}\\ &\lesssim \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2(\Delta_{\min}^2 + L^{(t)}) + c\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right),\end{aligned} \tag{82}$$

where the third inequality follows by the procedures to derive (79) and (80), the inequality (54) in Lemma 13, and the last inequality follows by the assumption $\Delta_{\min}^2 \geq p^{-K/2}\log p$.

Combining the inequalities (79), (80), and (82), we finish the proof of inequality (77).

3) Upper bound for $H_{ib}^{(t)}$ (78). Recall the definition of H_{ib} and rearrange the terms

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

where

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

For H_1, we have

$$|H_1| \leq \frac{4\max_{a\in[r]}\|\boldsymbol{W}_{:a}^T\boldsymbol{E}\boldsymbol{V}^{(t)}\|^2}{\|\boldsymbol{W}_{z(i):}^T\boldsymbol{X}\boldsymbol{V}^{(t)}\|^2} \leq \frac{r^{2K-1}+Kpr^{K+1}}{p^K} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2, \tag{83}$$

following the derivation of G_2 in inequality (80) and the assumption that $\Delta_{\min}^2 \geq p^{-K/2}\log p$.

For H_2, by the inequality (81), we have

$$|H_2| \lesssim 2\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2 + \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2, \tag{84}$$

where the last inequality follows by the condition (39) in Condition 1.

For H_3, by Cauchy-Schwartz inequality, we have

$$|H_3| \lesssim \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\||H_1|^{1/2} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2, \tag{85}$$

following the inequalities (81) and (83).

Therefore, combining inequalities (83), (84), and (85), we have finish the proof of inequality (78).

□