Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu and Miaoyan Wang
University of Wisconsin - Madison

Abstract

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

Index Terms

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

I. Introduction

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

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

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 degree heterogeneity may seriously mislead the clustering results. For

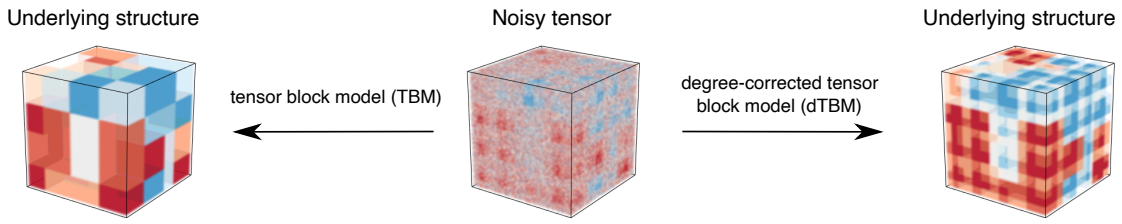

Fig. 1: Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

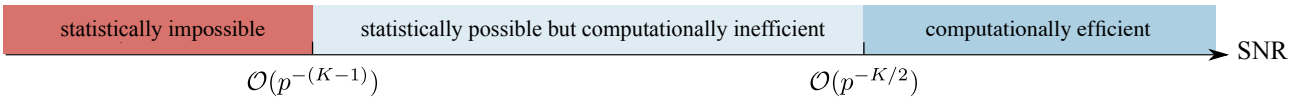

Fig. 2: SNR thresholds for statistical and computational limits in order-K dTBM with dimension $(p, ..., p)$ and $K \geq 2$. The SNR gap between statistical possibility and computational efficiency exists only for tensors with $K \geq 3$.

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 perform iterative clustering based on angle similarity. We refer to the algorithm as the *spherical* clustering; detailed procedures are in Section IV. The spherical clustering avoids the estimation of nuisance degree heterogeneity. The usage of angle similarity brings new challenges to the theoretical results, and we develop new polar-coordinate based techniques in the proofs.

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

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

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

	Gao et al. (2018)	Han et al. (2020)	Ghoshdastidar et al. (2017)	Ke et al. (2019)	**Ours**
Allow tensors of arbitrary order	$\times$	$\surd$	$\surd$	$\surd$	$\surd$
Allow degree heterogeneity	$\surd$	$\times$	$\surd$	$\surd$	$\surd$
Singular-value gap-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 $r = \mathcal{O}(1)$.

- *Tensor block model.* The tensor block model (TBM, Ghoshdastidar et al. (2017); Han et al. (2020); Wang and Zeng (2019)) is a higher-order extension of SBM, but it fails to allow degree heterogeneity. The Cartesian coordinates based analysis in Han et al. (2020) is non-applicable to handle the extra complexity brought by degree heterogeneity. In contrast, our model addresses the degree heterogeneity, and we develop polar-coordinate based tools for the theoretical analysis.
- *Degree-corrected block model.* The hypergraph degree-corrected block model (hDCBM) proposed by Ke et al. (2019); Yuan et al. (ress) accounts for degree heterogeneity. However, the hDCBM is designed only for binary observations, and the proposed spectral algorithm in Ke et al. (2019) achieves sub-optimal polynomial clustering rate 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. See Fig 2 for overview of our results.
- *Global-to-local algorithm strategy.* Our methods generalize the recent global-to-local strategy for matrix learning (Chi et al., 2019; Gao et al., 2018; Yun and Proutiere, 2016) to tensors (Ahn et al., 2018; Han et al., 2020; Kim et al., 2018). Despite the conceptual similarity, we address several fundamental challenges associated with this non-convex, non-continuous problem. We show the insufficiency of the conventional tensor HOSVD (De Lathauwer et al., 2000), and we develop a weighted higher-order initialization that relaxes the singular-value gap separation condition. Furthermore, our local iteration leverages the angle-based clustering in order to avoid explicit estimation of degree heterogeneity. Our bounds reveal the interesting interplay between the computational and statistical errors. We show that our final estimate *provably* achieves the exact clustering within only polynomial-time complexity.

Notation. We use lower-case letters (e.g., a, b) for scalars, lower-case boldface letters (e.g., $\boldsymbol{a}, \boldsymbol{\theta}$) 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 $\mathbf{1}_p$ to denote a vector of length p with all entries to be 1. We use $|\cdot|$ for the cardinality of a set and $\mathbb{1}\{\cdot\}$ 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 $\cos(\boldsymbol{a}, \boldsymbol{b}) = \cos\left(\boldsymbol{a}^T, \boldsymbol{b}^T\right)$.

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

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

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

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

II. Model formulation and motivations

A. Degree-corrected tensor block model

Suppose we have an order-K data tensor $\mathcal{Y} \in \mathbb{R}^{p\times\cdots\times p}$. For ease of notation, we focus on symmetric tensors in this section; the extension to general asymmetric tensors is provided in Section IV-C. 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),

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

where $\mathcal{S} \in \mathbb{R}^{r\times\cdots\times r}$ is an order-K tensor collecting the block means among communities, and $\mathcal{E} \in \mathbb{R}^{p\times\cdots\times p}$ is a noise tensor consisting of independent zero-mean sub-Gaussian entries with variance bounded by 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:

$$\mathbb{E}\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \cdots \times_K \boldsymbol{\Theta}\boldsymbol{M}, \tag{2}$$

where $\boldsymbol{\Theta} = \text{diag}(\theta(1), ..., \theta(p)) \in \mathbb{R}^{p\times p}$ 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 (2) more challenging than Tucker decomposition. We call a tensor $\mathcal{Y}$ an r-block tensor with degree $\boldsymbol{\theta}$ if $\mathcal{Y}$ admits dTBM (2). The goal of clustering is to estimate z from a single noisy tensor $\mathcal{Y}$. We are particularly interested in the high-dimensional regime where p grows whereas $r = \mathcal{O}(1)$.

B. Motivating examples

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

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

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

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

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

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

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

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

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

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 (2)

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

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

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

C. Identifiability under angle gap condition

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

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

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

where $c_i > 0$'s are universal constants. We briefly describe the rationale of the constraints in (3). First, the entrywise positivity constraint on $\boldsymbol{\theta} \in \mathbb{R}_+^p$ is imposed to avoid sign ambiguity between entries in $\boldsymbol{\theta}_{z^{-1}(a)}$ and $\mathcal{S}$. This constraint allows the trigonometric cos to describe the angle similarity in the Assumption 1 below and Sub-algorithm 2 in Section IV. Note that the positivity constraint can be achieved without sacrificing model flexibility, by using a slightly larger dimension of $\mathcal{S}$ in the factorization (2); see Example 1 below. Second, recall that the quantity $|z^{-1}(a)|$ denotes the number of nodes in a-th community. The constants c_1, c_2 in the $|z^{-1}(a)|$ bound assume the roughly balanced size across r communities. Third, the 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 a 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. Consider an order-3 dTBM with core tensor $\mathcal{S} = 1$ and degree $\boldsymbol{\theta} = (1, 1, -1, -1)^T$. We have the mean tensor

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

where $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta})$ 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 $\mathcal{X}$ as a 2-block tensor with *positive-signed* degree. Specifically, we write

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

where $\boldsymbol{\Theta}' = \text{diag}(\boldsymbol{\theta}') = \text{diag}(1, 1, 1, 1)$, the core tensor $\mathcal{S}' \in \mathbb{R}^{2\times2\times2}$ has mode-3 slices, and the membership matrix $\boldsymbol{M}' \in \{0, 1\}^{2\times4}$ defines the clustering $z' \colon [4] \to [2]$,

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

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

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

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

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

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

bound on $\|\boldsymbol{S}_{a:}\|$ in (3).

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

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

The identifiability guarantee for the dTBM is 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 limits for higher-order tensors

In this section, we study the statistical and computational limits of dTBM. We propose signal-to-noise ratio (SNR),

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

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 (5), we consider the following parameter space,

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

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

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

where $\pi : [r] \mapsto [r]$ is a permutation of cluster labels, $\circ$ denotes the composition operation, and 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 (1) without symmetric assumptions. For general (asymmetric) Gaussian dTBMs, we assume Gaussian noise $\mathcal{E}(i_1, \ldots, i_K) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and we extend the parameter space (3) to allow K clustering functions $(z_k)_{k \in [K]}$, one for each mode. For notational simplicity, we still use z and $\mathcal{P}(\gamma)$ for this general (asymmetric) model. All lower bounds should be interpreted as the worst-case results across K modes.

A. Statistical critical values

The statistical limit means the minimal SNR required for solving dTBMs with *unlimited computational cost.* Our following result shows 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). As we will show in Section IV, the SNR threshold $-(K-1)$ is also a minimax upper bound, because MLE achieves exact recovery when $\gamma > -(K-1)$. Hence, the boundary $\gamma_{\text{stat}} := -(K-1)$ is the critical value for statistical performance of dTBM.

B. Computational critical values

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

Definition 1 (Hypergraphic planted clique (HPC) detection)**.** Consider an order-K hypergraph $H = (V, E)$ where $V = [p]$ collects vertices and E collects all the order-K edges. Let $\mathcal{H}_k(p, 1/2)$ denote the Erdős-Rényi K-hypergraph where the edge $(i_1, \ldots, i_K)$ 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 an Erdős-Rényi K-hypergraph. The HPC detection is formulated as the following hypothesis testing problem

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

Conjecture 1 (HPC conjecture)**.** Consider the HPC detection problem in Definition 1. Suppose the sequence $\{\kappa_p\}$ such that $\limsup_{p\to\infty} \log \kappa_p / \log \sqrt{p} \le (1-\tau)$. Then, for every sequence of polynomial-time 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

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

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 by combing Theorems 2 and 3. Figure 2 summarizes our results of critical SNRs when $K \ge 2$. In the weak SNR region $\gamma < -(K-1)$, no statistical estimator succeeds in degree-corrected higher-order clustering. In the strong SNR region $\gamma > -K/2$, our proposed algorithm precisely recovers the clustering in polynomial time. In the moderate SNR regime, $-(K-1) \le \gamma \le -K/2$, the degree-corrected clustering problem is statistically easy but computationally hard. Particularly, dTBM reduces to matrix degree-corrected model when $K = 2$, and the statistical and computational bounds show the same critical value. When $K = 1$, dTBM reduces to the degree-corrected sub-Gaussian mixture model (GMM) with model

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

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 that

polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Therefore, we conclude that the statistical-to-computational gap emerges only for higher-order tensors with $K \geq 3$. The result reveals the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering.

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

In this section, we present an efficient polynomial-time clustering algorithm under mild SNR. The procedure takes a global-to-local approach. See Figure 3 for illustration. The global step finds the basin of attraction with polynomial miclustering error, whereas the local iterations improve the initial clustering to exact recovery. Both steps are critical to obtain a satisfactory algorithm output. In what follows, we first use the symmetric tensor as a working example to describe the algorithm procedures to gain insight. Our theoretical analysis focuses on the 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 Section IV-C.

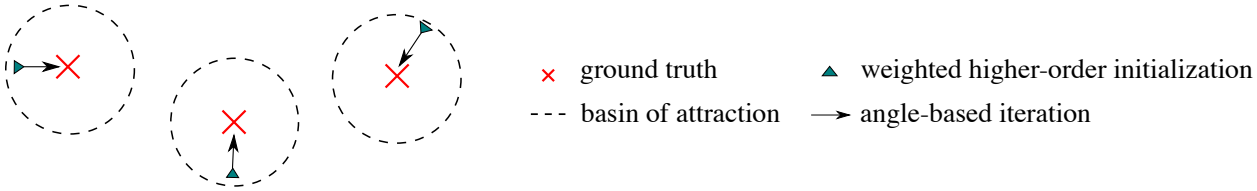

Fig. 3: Illustration of our global-to-local algorithm.

A. Weighted higher-order initialization

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

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

This implies that, all node i belonging to a-th community (i.e., $z(i) = a$) share the same normalized mean vector $\theta(i)^{-1}\boldsymbol{X}_{i:}$, 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 $\boldsymbol{U}_{\text{pre}} = \mathrm{SVD}_r(\mathrm{Mat}(\mathcal{Y}))$, where $\mathrm{SVD}_r(\cdot)$ returns the top-r left singular vectors. The second projection performs HOSVD on the projected $\mathcal{Y}$ onto the multilinear Kronecker space $\boldsymbol{U}_{\text{pre}} \otimes \boldsymbol{U}_{\text{pre}}$; i.e.,

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

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 effects for $K \geq 3$ (Han et al., 2020).

The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted k-means clustering. We write $\hat{\boldsymbol{X}} = \mathrm{Mat}(\hat{\mathcal{X}})$, 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 singular-value gap condition (Gao et al., 2018; Han et al., 2020). We assign degenerate rows with purely zero entries to an arbitrarily random cluster; these nodes are negligible in high-dimensions because of the lower bound on $\|\mathrm{Mat}(\mathcal{S})_{a:}\|$ in (3). The final result gives the initial cluster assignment $\hat{z}^{(0)}$. Full procedures are provided in Sub-algorithm 1.

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

Sub-algorithm 1: Weighted higher-order initialization

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

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

2: Compute factor matrix $\hat{\boldsymbol{U}} = \mathrm{SVD}_r(\mathrm{Mat}(\mathcal{X}_{\text{pre}}))$ 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 $\hat{\boldsymbol{X}} = \mathrm{Mat}(\hat{\mathcal{X}})$ and $S_0 = \{i \in [p] : \|\hat{\boldsymbol{X}}_{i:}\| = 0\}$. Set $\hat{z}(i)$ randomly in $[r]$ for $i \in S_0$.

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

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

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

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 $\mathcal{S}^{(t)}(i_1, ..., i_K) = \mathrm{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z^{(t)}(i_k) = j_k, k \in [K]\}$.

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

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

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

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

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

11: **end for**

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

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

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

Note that, the assumption (7) 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 $\min_{i\in[p]} \theta(i) \ge c$ for some constant $c > 0$. Let $z^{(0)}$ denote the output of Sub-algorithm 1. With probability going to 1, we have

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-K/2}}{\mathrm{SNR}}. \tag{8}$$

Remark 4 (Comparison to previous results)**.** For fixed SNR, our initialization error rate with $K = 2$ agrees with the

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

Remark 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 $\mathcal{O}(p^{-1})$ for all $K \geq 2$. This rate is substantially worse than our current rate (8).

Remark 6 (Singular-value gap-free clustering)**.** Note that our clustering directly applies to the estimated mean tensor $\hat{\mathcal{X}}$ rather than the leading tensor factors $\hat{\boldsymbol{U}}$. Applying clustering to the tensor factors suffers from the non-identifiability issue due to the infinitely many orthogonal rotations when the number of blocks $r \geq 3$ in the absence of singular-value gaps. Such ambiguity causes the trouble for effective clustering (Abbe et al., 2020). In contrast, our initialization algorithm applies the clustering to the overall mean tensor $\hat{\mathcal{X}}$. This strategy avoids the non-identifiability issue regardless of the number of blocks and singular-value gaps.

B. Angle-based iteration

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

$$z(i) = \underset{a\in[r]}{\arg\max} \cos(\boldsymbol{x}_i,\ \boldsymbol{s}_a), \quad \text{for all } i \in [p]. \tag{9}$$

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

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

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

We use $\boldsymbol{Y}^d$ and $\boldsymbol{S}^{(t)}$ to denote the Mat($\mathcal{Y}^{\text{d}}$) and Mat($\mathcal{S}^{(t)}$). The rows $\boldsymbol{Y}^d_{i:}$ and $\boldsymbol{S}^{(t)}_{a:}$ correspond to the $\boldsymbol{x}_i$ and $\boldsymbol{s}_a$ in the one-dimensional clustering (9). Then, we obtain the updated assignment by

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

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

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

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

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

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

$$\sin(\boldsymbol{p}(\bar{z}),\ \boldsymbol{p}_{\boldsymbol{\theta}}(\bar{z})) \lesssim \varepsilon\Delta_{\min}, \quad \text{for all } \bar{z} \in \mathcal{N}(z,\varepsilon). \tag{10}$$

Roughly speaking, the vector $\boldsymbol{p}(\bar{z})$ represents the raw cluster sizes, and $\boldsymbol{p}_\theta(\bar{z})$ represents the relative cluster sizes weighted by degrees. The local stability holds trivially for $\varepsilon = 0$ based on the construction of parameter space (3). The condition (10) controls the impact of node degree to the $\boldsymbol{p}_\theta(\cdot)$ with respect to the misclassification rate ε and angle gap.

Theorem 5 (Error for angle-based iteration)**.** Consider the setup as in Theorem 4. Assume the local linear stability of degree holds in the neighborhood $\mathcal{N}(z,\epsilon)$ for $\epsilon \geq \log^{-1} p$. Suppose $r = \mathcal{O}(1)$ and 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

$$\ell(z,\hat{z}^{(t+1)}) \lesssim \underbrace{\mathrm{SNR}^{-1}\exp\left(-\frac{p^{K-1}\mathrm{SNR}}{r^{K-1}}\right)}_{\text{statistical error}} + \underbrace{\rho^t \ell(z,z^{(0)})}_{\text{computational error}}. \tag{11}$$

From the conclusion (11), we find that the iteration error is decomposed into two parts: statistical error and computational error. The statistical error is unavoidable with noisy data regardless t, whereas the computational error decays in an exponential rate as the number of iterations $t \to \infty$.

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

C. Extensions and practical issues

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 (8) in Theorem 4.

We now provide a high-level explanation for the technical difficulty when applying Theorem 4 to Bernoulli observations. The derivation of Theorem 4 relies on the upper bound of the estimation error for the mean tensor in Lemma 6; i.e., with high probability

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2}, \tag{12}$$

where $\mathcal{X} = \mathbb{E}\mathcal{Y}$ and $\hat{\mathcal{X}}$ is defined in Step 2 of Sub-algorithm 1. Unfortunately, the inequality (12) holds only for i.i.d. sub-Gaussian observations while Bernoulli observations are generally not identically distributed.

One possible remedy is to apply singular value decomposition to the *square unfolding* (Mu et al., 2014) of Bernoulli tensor $\mathcal{Y}$. Let the matrix $\mathrm{Mat}_{sq}(\mathcal{Y}) \in \{0,1\}^{\lfloor p^{K/2}\rfloor \times \lceil p^{K/2}\rceil}$ denote the nearly square unfolded Bernoulli tensor. Define a new estimate

$$\hat{\mathcal{X}}' = \mathop{\arg\min}_{\mathrm{rank}(\mathrm{Mat}_{sq}(\mathcal{X}))\leq r^{\lceil K/2\rceil}} \|\mathrm{Mat}_{sq}(\mathcal{X}) - \mathrm{Mat}_{sq}(\mathcal{Y})\|_F^2. \tag{13}$$

The optimization (13) is simply a matrix SVD problem. Following Lemma 7 in Gao et al. (2018), with high probability, the new estimate satisfies

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

Replacing the estimate $\hat{\mathcal{X}}$ by $\hat{\mathcal{X}}'$ in Theorem 4, the high probability upper bound for Bernoulli initialization is

$$\ell(z^{(0)}, z) \lesssim \frac{r^K p^{-\lfloor K/2 \rfloor}}{\text{SNR}}. \tag{14}$$

The Bernoulli bound (14) is relatively looser than Gaussian bound (8), especially when K is small. Nevertheless, our bound (14) is already tighter than the pervious work (Ke et al., 2019). The investigation of the gap between upper bound $p^{-\lfloor K/2 \rfloor}$ and the lower bound $p^{-K/2}$ for Bernoulli tensors will be left as future work.

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 clusters 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 $k \in [K]$. The detailed model extension for general asymmetric dTBMS can be found in Appendix.

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

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

$$\begin{aligned} \hat{r} &= \mathop{\arg\min}_{r \in \mathbb{Z}_+} \left(p^K \log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_F^2) + p_e(r) K \log p \right), \\ \text{where} \quad \hat{\mathcal{X}} &= \hat{\mathcal{S}}(r) \times_1 \hat{\boldsymbol{\Theta}}(r) \hat{\boldsymbol{M}}(r) \times_2 \cdots \times_K \hat{\boldsymbol{\Theta}}(r) \hat{\boldsymbol{M}}(r), \end{aligned} \tag{15}$$

where the triplet $(\hat{z}(r), \hat{\boldsymbol{S}}(r), \hat{\boldsymbol{\theta}}(r))$ are estimated parameters with cluster number r, and $p_e(r) = r^K + p(\log r + 1) - r$ is the effective number of parameters. Note that we have added the argument (r) to related quantities as functions of r. In particular, the estimate $\hat{\boldsymbol{\theta}}(r)$ in (15) is obtained by first calculating the reduced tensor $\hat{\mathcal{Y}}^{\text{d}}$ with $\hat{z}(r)$, and then normalizing the row norms $\|\hat{\boldsymbol{Y}}_{i:}^{\text{d}}\|$ to 1 in each cluster; i.e.,

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

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

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 $\alpha = 1 + \Omega(p^{\gamma/2})$ with $\gamma < 0$ by Assumption 1 and definition (5). 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 by $\theta(i) = |X_i| + 1 - 1/\sqrt{2\pi}$ with $|X_i| \overset{\text{i.i.d.}}{\sim} N(0,1), i \in [p]$ and normalize $\boldsymbol{\theta}$ to satisfy (3). Also, we set $\sigma^2 = 1$ for Gaussian data without further specification.

A. *Verification of theoretical results*

The first experiment verifies statistical-computational gap described in Section III. Consider the Gaussian model with $p = \{80, 100\}$, $r = 5$. We vary 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.

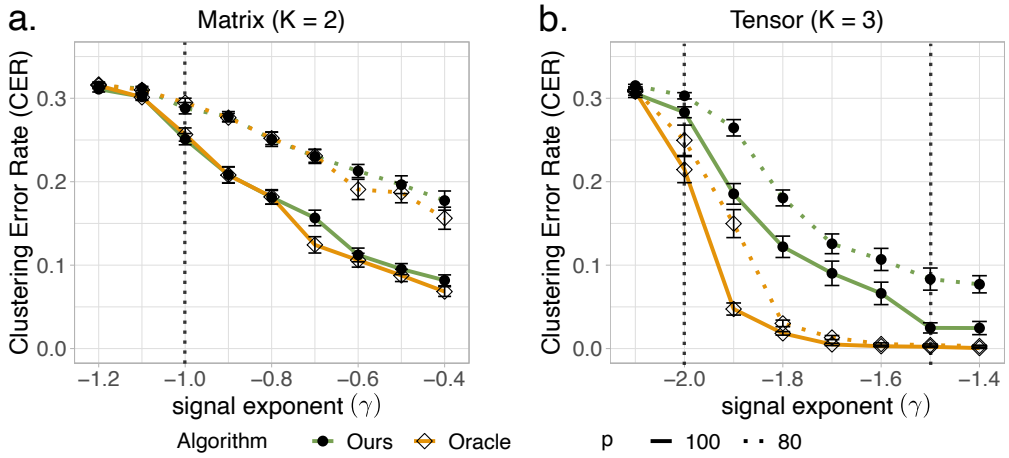

Fig. 4: SNR phase transitions for clustering in dTBM with $p = \{80, 100\}, r = 5$ under (a) matrix case with $\gamma \in [-1.2, -0.4]$ and (b) tensor case with $\gamma \in [-2.1, -1.4]$.

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

The third experiment evaluates the empirical performance of the BIC criterion to select unknown number of clusters. We generate the data from an order-3 Gaussian model with $p = \{50, 80\}$, $r = \{2, 4\}$, and noise level $\sigma^2 \in \{0.25, 1\}$. Table II shows that our BIC criterion well chooses the true r under most settings. Note that the BIC slightly underestimates the true number of clusters ($r = 4$) with smaller dimension and higher noise ($p = 50, \sigma = 1$), and

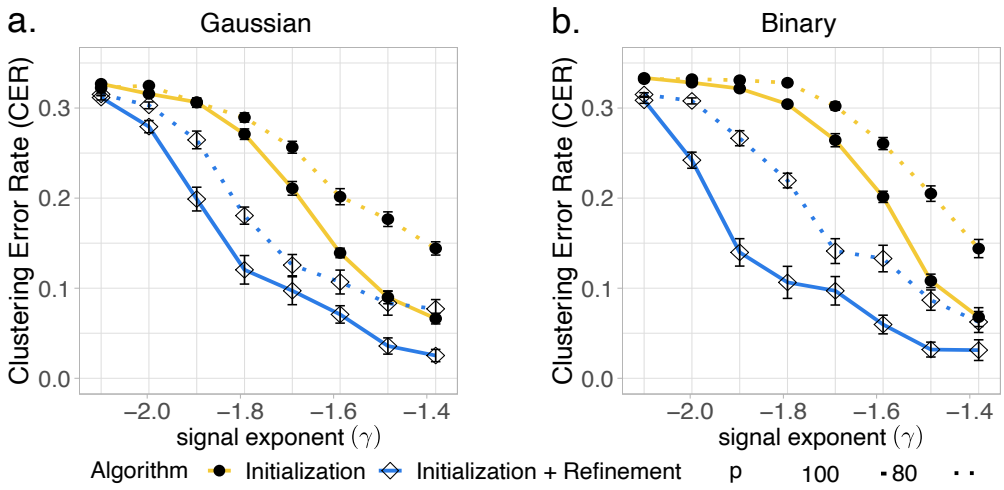

Fig. 5: CER versus signal exponent (γ) for initialization only and for combined algorithm. We set $p = \{80, 100\}, r = 5, \gamma \in [-2.1, -1.4]$ under (a) Gaussian models and (b) Bernoulli models.

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

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

the accuracy immediately increases with larger dimension $p = 80$. The improvement follows from the fact that a larger dimension p indicates a larger sample size in the tensor block model. Therefore, we conclude that BIC criterion is a reasonable way to tune the number of clusters.

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** (Han et al., 2020): High-order clustering algorithm developed for non-degree tensor block models;
- **SCORE** (Ke et al., 2019): Tensor-SCORE for clustering developed for binary tensors.

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

We investigate the effects of signal to clustering performance by varying $\gamma \in [-1.5, -1.1]$. Figure 6 shows 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

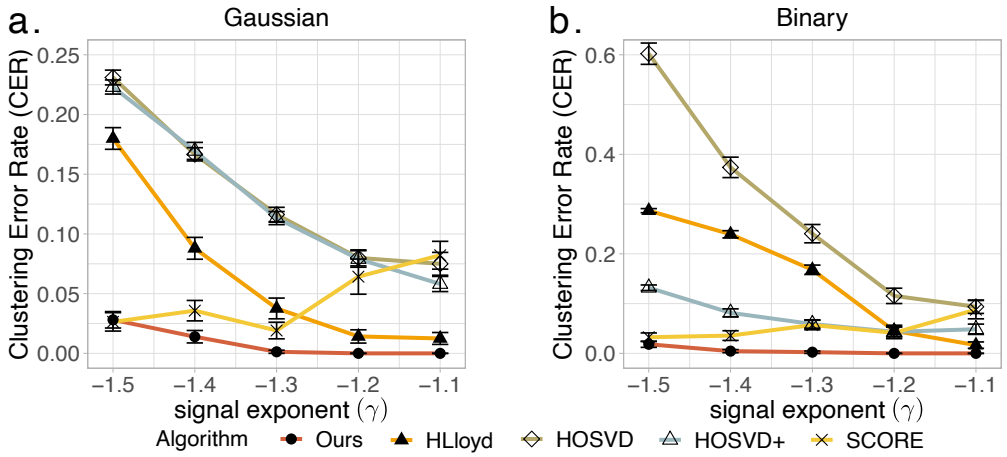

Fig. 6: CER versus signal exponent (denoted γ) for different methods. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under (a) Gaussian and (b) Bernoulli models.

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

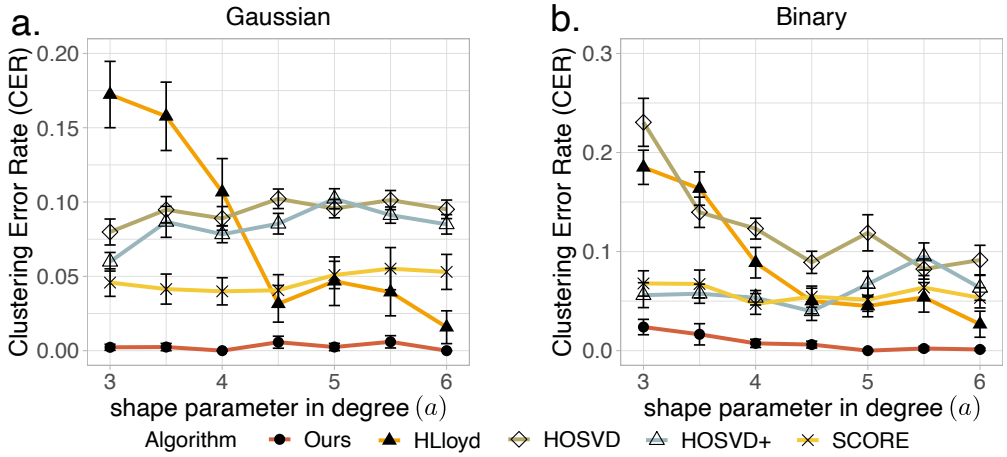

Fig. 7: CER versus shape parameter in degree (denoted $a \in [3, 6]$) for different methods. We set $p = 100, r = 5, \gamma = -1.2$ under (a) Gaussian and (b) Bernoulli models.

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. We vary the shape parameter $a \in [3, 6]$ in the Pareto distribution to investigate a range of degree heterogeneities. Figure 7 demonstrates the stability of degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over the entire range of degree heterogeneity under consideration. In contrast, non-degree algorithms (**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.

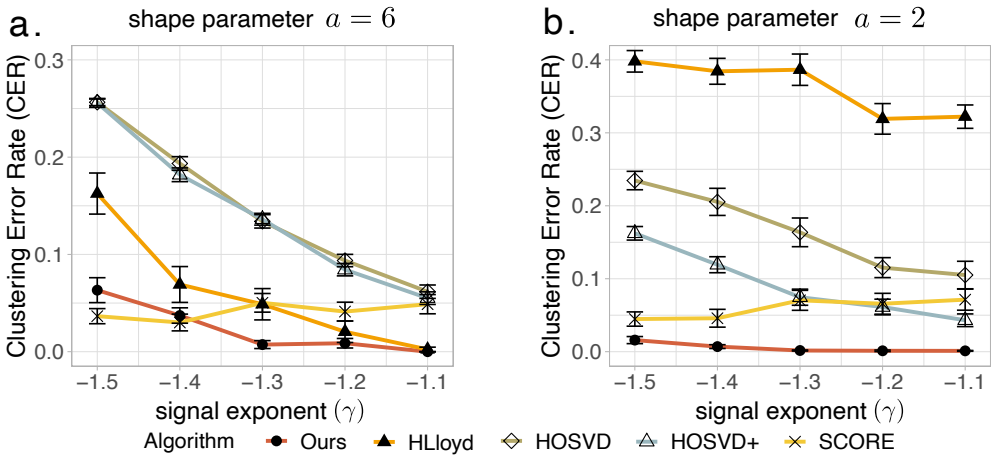

Fig. 8: CER comparison versus signal exponent (denoted γ) under (a) low (shape parameter $a = 6$) (b) high (shape parameter $a = 2$) degree heterogeneity. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under Gaussian model.

VI. Real data applications

A. Human brain connectome data analysis

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

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

Figure 10 illustrates the estimated core tensor $\hat{\mathcal{S}}$ with estimated clustering, and Figure 11 visualizes the average brain connections and the connection enrichment in contrast to average networks in each group. In general, we find that the inner-hemisphere connection has stronger connection compared to inter-hemisphere connections (Figure 10a). Also, the back and front parts (*R.Occipital*, *R.Supra*) are shown to have more interactions with temporal tissues than inner-cluster connections. In addition, the group 1 with 54% females shows an enrichment on the inter-hemisphere connections (Figure 10b), while group 4 with only 36% females exhibits a reduction (Figure 10d). This result agrees with previous findings in Hu et al. (2021). The enrichment on the back-front connection is also recognized in group 3 (Figure 10c). The interpretive patterns in our results demonstrate the usefulness of our clustering methods in the human brain connectome data application.

B. Peru Legislation data analysis

We also apply our method to the legislation networks in the Congress of the Republic of Peru (Lee et al., 2017). Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the

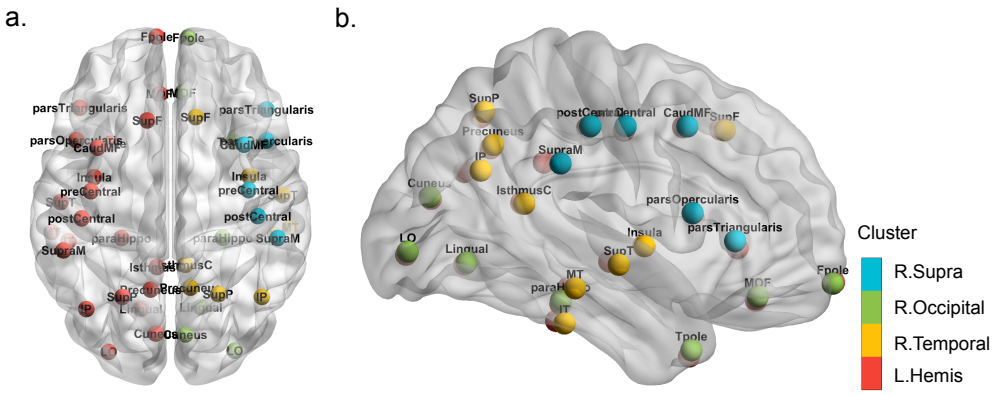

Fig. 9: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

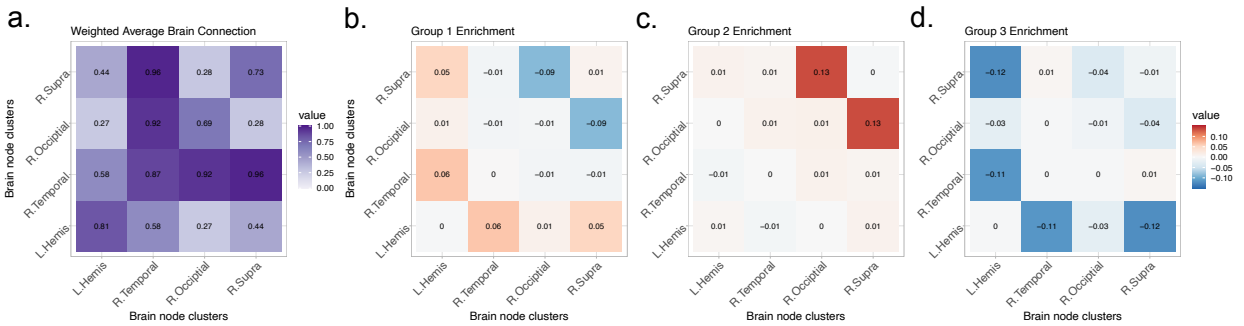

Fig. 10: Mode 3 slices of estimated core tensor $\hat{\mathcal{S}}$. (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the difference between each slice of $\hat{\mathcal{S}}$ and the averaged slice.

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 the ground truth. We apply various higher-order clustering methods to $\mathcal{Y}$ with $r = 5$. Table III shows that our **dTBM** achieves the best performance compared to others. The second best

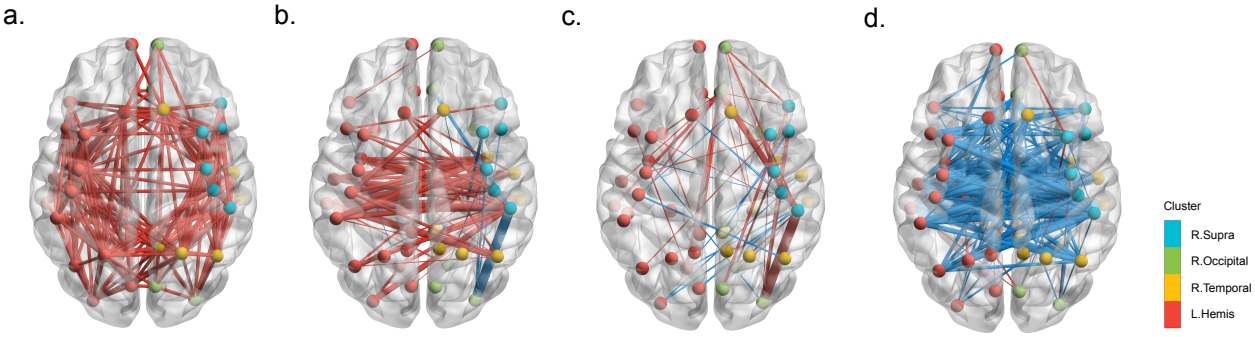

Fig. 11: Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain network enrichments in Groups 1-3. Red edges represent the positive enrichment and blue edges represent the negative enrichment.

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

VII. Proof Sketches

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

A. Proof sketch of Theorem 4

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

$$\ell(z^{(0)}, z) \lesssim \frac{1}{p} \min_{\pi \in \Pi} \sum_{i: z^{(0)}(i) \neq \pi(z(i))} \theta(i)^2 \lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K} \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim \frac{r^K p^{-K/2}}{\text{SNR}}, \tag{16}$$

where $\mathcal{X} = \mathbb{E}\mathcal{Y}$ is the true mean. The first inequality in (16) holds with the assumption $\min_{i \in [p]} \theta(i) \geq c > 0$ in Theorem 4. The second inequality relies on an important conclusion that the angle gap of mean tensor $\mathcal{X}$ is lower bounded by that of core tensor $\mathcal{S}$, i.e., the minimal angle gap $\Delta_{\min}$ defined in Assumption 1. Let $\boldsymbol{a}^s := \boldsymbol{a}/\|\boldsymbol{a}\|$ denote the normalized vector with the convention that $\boldsymbol{a}^s = 0$ if $\boldsymbol{a} = 0$. We want to show that

$$\min_{z(i) \neq z(j)} \|[\boldsymbol{X}_{i:}]^s - [\boldsymbol{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \tag{17}$$

where $\boldsymbol{X} = \text{Mat}(\mathcal{X})$. The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (17), in which the proof of Gao et al. (2018) is no longer applicable due to different angle gap assumption in our dTBM. We develop the extra padding technique in Lemma 3 and balance assumption (7) to derive (17). Last, we finish the proof of Theorem 4 by showing the third inequality of (16) using Han et al. (2020, Proposition 1).

B. Proof sketch of Theorem 5

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

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

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

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

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

$$\mathbb{1}\left\{z^{(t+1)}(i) = b\right\} = \mathbb{1}\left\{\|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \le \|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2\right\} \le A_{ib} + B_{ib}, \tag{18}$$

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

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

The terms $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ are controlled by $z^{(t)}, \mathcal{S}^{(t)}$; see the detailed definitions in (57), (58), (59). Note that the event A_{ib} only involves the oracle estimator independent of t, while all the terms related to the t-th iteration are in B_{ib}. Thus, the inequality (18) decomposes the misclustering loss in the $(t+1)$-th iteration into the oracle loss and the loss in t-th iteration. This decomposition leads to the separation of statistical error and computational error in the final upper bound of Theorem 5.

Specifically, we prove the contraction inequality

$$L^{(t+1)} \lesssim \xi + \rho L^{(t)}, \quad \xi = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}A_{ib}\left\|\left[\boldsymbol{S}_{z(i):}\right]^s - [\boldsymbol{S}_{b:}]^s\right\|^2, \tag{19}$$

where $\rho \in (0,1)$ is the contraction parameter, and we call ξ the oracle loss. Controlling the probability of event B_{ib} and obtaining the $\rho L^{(t)}$ term in the right hand side of (19) are the most challenging parts in the proof of Theorem 5. Note that the true and estimated core tensors are involved via their normalized rows such as $\boldsymbol{S}_{a:}^s, \tilde{\boldsymbol{S}}_{a:}^s, [\boldsymbol{S}_{a:}^{(t)}]^s$. The Cartesian coordinate based analysis in Han et al. (2020) is no longer applicable in our case. Instead, we use the polar-coordinate based analysis and the geometry property of trigonometric functions to derive the high probability upper bounds for $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$.

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

$$\xi \lesssim \exp\left(-\frac{p^{K-1}\text{SNR}}{r^{K-1}}\right). \tag{20}$$

Combining the decomposition (19) and the oracle bound (20), we finish the proof of Theorem 5.

ACKNOWLEDGMENTS

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

REFERENCES

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *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. *The 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. *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., Sgonne, 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. *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*.

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 clusteringsa geometric perspective. *Machine Learning*, 86(3):369–389.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81.

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

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

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

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, volume 32, pages 713–723.

Yuan, M., Liu, R., Feng, Y., and Shang, Z. (In Press). Testing community structures for hypergraphs. *The Annals of Statistics, arXiv preprint arXiv:1810.04617*.

Yun, S.-Y. and Proutiere, A. (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in*

Neural Information Processing Systems, volume 29, pages 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 first show the proof of main theorem and then collect the useful lemmas in the end.

NOTATION

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

Preliminaries.

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

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

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

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

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

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

Model extension to 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

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \quad \text{where} \quad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{\Theta}_K \boldsymbol{M}_K, \tag{21}$$

where $\mathcal{Y} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is the data tensor, $\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is the mean tensor, $\mathcal{S} \in \mathbb{R}^{r_1\times\cdots\times r_K}$ is the core tensor, $\mathcal{E} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ is the noise tensor consisting of independent zero-mean sub-Gaussian entries with variance bounded by σ^2, $\boldsymbol{\Theta}_k = \text{diag}(\boldsymbol{\theta}_k)$, 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 $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$,

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

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

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

We also consider the generalized Assumption 1 on angle gap.

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

$$\Delta_{\min} \coloneqq \min_{k\in[K]} \min_{a\neq b\in[r_k]} \left\|\boldsymbol{S}^s_{k,a:} - \boldsymbol{S}^s_{k,b:}\right\| > 0.$$

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

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

Proof of Theorem 1

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

$$\mathcal{X} = \mathcal{S} \times_1 \Theta_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \Theta_K \boldsymbol{M}'_K = \mathcal{S}' \times_1 \Theta'_1 \boldsymbol{M}'_1 \times_2 \cdots \times_K \Theta'_K \boldsymbol{M}'_K, \tag{22}$$

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

($\Leftarrow$) For the necessity, it suffices to construct two distinct parameters up to cluster label permutation, if the model (21) violates Assumption 2. Without loss of generality, we assume $\left\|\boldsymbol{S}^s_{1,1:} - \boldsymbol{S}^s_{1,2:}\right\| = 0$.

If $\boldsymbol{S}_{1,1:}$ is a zero vector, construct $\boldsymbol{\theta}'_1$ such that $\boldsymbol{\theta}'_{1,z_1^{-1}(1)} \neq \boldsymbol{\theta}_{1,z_1^{-1}(1)}$. 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 distinct from $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation. Similar conclusion holds 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

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

Let $\boldsymbol{D} = \mathrm{diag}(1 + c, 1, ..., 1) \in \mathbb{R}^{r_1-1\times r_1-1}$. 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 distinct from $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation.

($\Rightarrow$) For the sufficiency, it suffices to show that all possible triplets $(\{z'_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\})$ are identical to $(\{z_k\}, \mathcal{S}, \{\boldsymbol{\theta}_k\})$ up to label permutation if the model (21) satisfies Assumption (2). We show the uniqueness of the three parameters, $\{\boldsymbol{M}_k\}, \{\mathcal{S}\}, \{\boldsymbol{\theta}_k\}$ separately.

First, we show the uniqueness of $\boldsymbol{M}_k$ for all $k \in [K]$. Without loss of generality, we consider $k = 1$ and show 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 conclusion for $k \geq 2$ can be showed similarly and thus omitted.

Consider an arbitrary node pair (i, j). If $z_1(i) = z_1(j)$, then we have $\|\boldsymbol{X}^s_{1,z_1(i):} - \boldsymbol{X}^s_{1,z_1(j):}\| = 0$ and thus $\|(\boldsymbol{S}')^s_{1,z'_1(i):} - (\boldsymbol{S}')^s_{1,z'_1(j):}\| = 0$ by Lemma 1. Then, by Assumption (2), we have $z'_1(i) = z'_1(j)$. Conversely, if $z_1(i) \neq z_1(j)$, then we have $\left\|\boldsymbol{X}^s_{1,i:} - \boldsymbol{X}^s_{1,j:}\right\| \neq 0$ and thus $\left\|(\boldsymbol{S}')^s_{1,z'_1(i):} - (\boldsymbol{S}')^s_{1,z'_1(j):}\right\| \neq 0$ by Lemma 1. Hence, we have $z'_1(i) \neq z'_1(j)$. Therefore, we have proven that z'_1 is identical z_i up to label permutation.

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

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

$$\frac{\boldsymbol{X}_{1,z_1(i):}}{\boldsymbol{X}_{1,z_1(j):}} = \frac{\boldsymbol{X}'_{1,z_1(i):}}{\boldsymbol{X}'_{1,z_1(j):}}, \text{ which implies } \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta'_1(j)}{\theta'_1(i)}. \tag{23}$$

Let $\theta'_1(j) = c\theta_1(j)$ for some positive constant c. By equation (23), we have $\theta'_1(i) = c\theta_1(i)$ for all $i \in z_1^{-1}(a)$. By the constraint $(\{z_k\}, \mathcal{S}', \{\boldsymbol{\theta}'_k\}) \in \mathcal{P}(\{r_k\})$, we have

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

which implies $c = 1$. Hence, we have proven $\boldsymbol{\theta}_1 = \boldsymbol{\theta}'_1$ provided 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 $\boldsymbol{P}_k$'s are permutation matrices for all $k \in [K]$. Provided $z'_k = z_k, \boldsymbol{\theta}'_k = \boldsymbol{\theta}_k$, 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 (22), 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. □

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 (21) 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. Without loss of generality, we prove $k = 1$ only and drop the subscript k in $\boldsymbol{X}_k, \boldsymbol{S}_k$ for notational convenience. 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 suffices to show the following statement holds 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 (21) 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

$$\sup_{(\{z_k\},\mathcal{S},\{\boldsymbol{\theta}_k\})\in\mathcal{P}(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{k,\text{stat}}, z_k)\right] \geq 1. \tag{24}$$

Since the inequality (24) 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:

$$\text{(i) } \theta_k(i) = 1 \text{ for all } i \in [p]; \quad \text{(ii) } \Delta_{\min} \lesssim \left(\frac{p}{r}\right)^{-\frac{K-1}{2}} \sigma; \quad \text{(iii) } |z_k^{-1}(a)| = \frac{p}{r} \in \mathbb{Z}_+ \text{ for all } a \in [r], \tag{25}$$

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:

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

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

$$\inf_{\hat{z}_k} \sup_{z_k} \mathbb{E}\ell(\hat{z}_k, z_k) \gtrsim \frac{1}{r^3|T_k^c|} \sum_{i\in T_k^c} \inf_{\hat{z}_k} \left\{\mathbb{P}[\hat{z}_k(i) = 2 | z_k(i) = 1] + \mathbb{P}[\hat{z}_k(i) = 1 | z_k(i) = 2]\right\}, \tag{26}$$

where $\hat{z}_k$ and z_k on the left hand side denote the generic clustering functions in $\mathcal{P}(\gamma)$, z_k on the right hand side denotes a particular parameter satisfying properties (i)-(vi), and the infimum on the right hand side is taken over the restricted family of $\hat{z}$ satisfying (iv)-(vi). Here, the factor $r^3 = r \cdot r^2$ in (26) comes from two sources: $r^2 \asymp \binom{r}{2}$ comes from the multiple testing burden for all pairwise comparisons among r clusters; and another r comes from the number of elements $|T_k^c| \asymp p/r$ to be clustered.

Next, we need to find the lower bound of the rightmost side in (26). For simplicity, we show the bound for the mode-1 case $k = 1$ only. 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 (21). First, we reparameterize the model under the construction (25)

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

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 (26) for $i = 1$. The analysis for other $i \in T^c$ are similar. For notational simplicity, we suppress 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 (26) is

$$\boldsymbol{y} = \theta \boldsymbol{x}_z + \boldsymbol{e}, \tag{27}$$

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 (27):

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

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 MLE's 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

$$\hat{z}_{\text{MLE}} = \begin{cases} 1 & \text{if } \cos(\boldsymbol{y}, \boldsymbol{x}_1) \geq \cos(\boldsymbol{y}, \boldsymbol{x}_2) \text{ and } \langle \boldsymbol{y}, \boldsymbol{x}_1 \rangle > 0, \\ 2 & \text{if } \cos(\boldsymbol{y}, \boldsymbol{x}_1) < \cos(\boldsymbol{y}, \boldsymbol{x}_2) \text{ and } \langle \boldsymbol{y}, \boldsymbol{x}_2 \rangle > 0, \\ 1 \text{ or } 2 \text{ with equal probability} & \text{otherwise.} \end{cases} \tag{28}$$

The Neyman-Pearson Lemma implies

$$\inf_{\hat{z}} \{\mathbb{P}[\hat{z} = 2 | z = 1] + \mathbb{P}[\hat{z} = 1 | z = 2]\} = \mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] + \mathbb{P}[\hat{z}_{\text{MLE}} = 2 | z = 1]. \tag{29}$$

By symmetric, it suffices to bound $\mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2]$. Using (28), we obtain

$$\begin{aligned} \mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] &= \mathbb{P}\left[\cos(\theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_1) \geq \cos(\theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_2) \text{ and } \langle \theta \boldsymbol{x}_2 + \boldsymbol{e}, \boldsymbol{x}_1 \rangle > 0\right] \\ &\overset{(*)}{\geq} \mathbb{P}\left[\left\langle \boldsymbol{e}, \frac{\boldsymbol{x}_1^s - \boldsymbol{x}_2^s}{\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|} \right\rangle \geq \frac{\theta}{2} \|\boldsymbol{x}_2\| \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\| \rangle\right] - \\ &\qquad \mathbb{P}\left[\langle \boldsymbol{e}, \boldsymbol{x}_1^s \rangle \leq -\frac{\theta}{2} \|\boldsymbol{x}_2\| \left(2 - \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2\right)\right] \\ &\overset{(**)}{=} \Phi\left(\frac{\theta}{2} \|\boldsymbol{x}_2\| \left(2 - \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2\right)\right) - \Phi\left(\frac{\theta}{2} \|\boldsymbol{x}_2\| \|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|\right), \end{aligned} \tag{30}$$

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 (25) 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 (25)

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

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 (30) is lower bounded by

$$\mathbb{P}[\hat{z}_{\text{MLE}} = 1 | z = 2] \geq \mathbb{P}\left[\left(\frac{p}{r}\right)^{(K-1)/2} \Delta_{\min} \lesssim N(0, 1) \lesssim \left(\frac{p}{r}\right)^{(K-1)/2}\right] \geq C > 0, \tag{31}$$

where the existence of strictly positive constant C is based on the SNR assumption (25). Combining (26), (29) and (31) 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

$$\mathcal{S} = \mathcal{S}(\gamma) = \begin{bmatrix}1\\1\\1\end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2} \begin{bmatrix}1\\-1\\0\end{bmatrix}^{\otimes K}. \tag{32}$$

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 (6).

(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

$$\liminf_{p\to\infty} \sup_{\mathcal{X}\in\mathcal{P}_{\text{non-degree}}(\gamma)} \mathbb{P}(\exists k \in [K],\ \hat{\boldsymbol{M}}_k \neq \boldsymbol{M}_k) \geq 1/2, \tag{33}$$

under the HPC Conjecture 1. The inequality (33) 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 (32) 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 (32) with that in the proof of Han et al. (2020, Theorem 7). □

PROOF OF THEOREM 4

Proof of Theorem 4. We prove Theorem 4 under the symmetric dTBM (2) with parameters $(z, \mathcal{S}, \boldsymbol{\theta})$. We drop the subscript k in the matricizations $\boldsymbol{M}_k, \boldsymbol{X}_k, \boldsymbol{S}_k$. For simplicity, let $\hat{z}$ denote the output, $\hat{z}^{(0)}$, of Sub-algorithm 1.

First, by Lemma 4, there exists a positive constant such that $\min_{z(i)\neq z(j)} \left\|\boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:}\right\| \geq c_0\Delta_{\min}$. By the balance assumption on $\boldsymbol{\theta}$ and Lemma 7, we have

$$\min_{\pi\in\Pi} \sum_{i:\hat{z}(i)\neq\pi(z(i))} \theta(i)^2 \leq \sum_{i\in S_I} \theta(i)^2 + 4\sum_{i\in S} \theta(i)^2, \tag{34}$$

where

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

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

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

where the last inequality follows Lemma 5, the assumption that $\min_{i\in[p]}\theta(i) \geq c$, and the constraint $\min_{a\in[r]}\|\boldsymbol{S}_{a:}\| \geq c_3$ in the parameter space (3). Thus, we have

$$\sum_{i\in P}\theta(i)^2 \lesssim \sum_{i\in P}\|\boldsymbol{X}_{i:}\|^2 p^{-(K-1)}r^{K-1}. \tag{35}$$

On the other hand, note that

$$\begin{aligned}
\sum_{i\in S}\|\boldsymbol{X}_{i:}\|^2 &\leq 2\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2 + 2\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:} - \boldsymbol{X}_{i:}\|^2 && (36)\\
&\leq \frac{8}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{x}}_{\hat{z}(i)} - \boldsymbol{X}^s_{i:}\|^2 + 2\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 && (37)\\
&\leq \frac{16}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\left[\|\hat{\boldsymbol{x}}_{\hat{z}(i)} - \hat{\boldsymbol{X}}^s_{i:}\|^2 + \|\hat{\boldsymbol{X}}^s_{i:} - \boldsymbol{X}^s_{i:}\|^2\right] + 2\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 && (38)\\
&\leq \frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{X}}^s_{i:} - \boldsymbol{X}^s_{i:}\|^2 + 2\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 && (39)\\
&\leq \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2} + 2\right)\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 && (40)\\
&\lesssim \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2} + 2\right)\left(p^{K/2}r + pr^2 + r^K\right)\sigma^2, && (41)
\end{aligned}$$

where inequalities (36) and (38) follow from the triangle inequality, (37) follows from the definition of S, (39) follows from the update rule of k-means in Step 5 of Sub-algorithm 1, (40) follows from Lemma 2, and the last inequality (41) follows from Lemma 6. Also, note that

$$\sum_{i\in S_0}\|\boldsymbol{X}_{i:}\|^2 = \sum_{i\in S_0}\|\hat{\boldsymbol{X}}_{i:} - \boldsymbol{X}_{i:}\|^2 \leq \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim \left(p^{K/2}r + pr^2 + r^K\right)\sigma^2, \tag{42}$$

where the equation follows from the definition of S_0. Therefore, combining the inequalities (34), (35), (41), and (42), 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 $\min_{i\in[p]}\theta(i)\geq c$, we finally obtain the result

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

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

Useful Corollary of Theorem 4

Corollary 1 (initial misclassification error)**.** Suppose $\text{SNR}\gtrsim p^{-K/2}\log p$ and $\sigma^2=1$. Then, the misclustering loss for the initialization is upper bounded as

$$L^{(0)}=\frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}\mathbb{1}\left\{z^{(0)}(i)=b\right\}\|[\boldsymbol{S}_{\pi^{(0)}(z(i)),:}]^s-[\boldsymbol{S}_{b:}]^s\|^2\lesssim\frac{\Delta_{\min}^2}{r\log p},$$

where $\pi^{(0)}$ minimizes the initial misclustering error; i.e., $\pi^{(0)}=\arg\min_{\pi\in\Pi}\sum_{i\in[p]}\mathbb{1}\left\{z^{(0)}(i)\neq\pi\circ z(i)\right\}$.

Proof of Corollary 1. Without loss of generality, we assume $\pi^{(0)}$ is the identity mapping such that $\pi^{(0)}(a)=a$ for all $a\in[r]$. Note that $\boldsymbol{X}_{i:}^s$ have only r different values. We let $\boldsymbol{X}_a^s=\boldsymbol{X}_{i:}^s$ for all i such that $z(i)=a, a\in[r]$.

Notice that

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

Therefore, when p is large enough, we have

$$\begin{aligned}\sum_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2\|\hat{\boldsymbol{X}}_i^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2&\lesssim\sum_{i\in[p]}\left(\|\boldsymbol{X}_{i:}\|^2-\|\boldsymbol{X}_{i:}-\hat{\boldsymbol{X}}_{i:}\|^2\right)\|\hat{\boldsymbol{X}}_{i:}^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2\\&\lesssim\sum_{i\in[p]}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{X}}_{i:}^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2\\&\lesssim\eta\sum_{i\in[p]}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{X}}_{i:}^s-\boldsymbol{X}_{i:}^s\|^2\\&\lesssim\|\hat{\mathcal{X}}-\mathcal{X}\|_F^2\\&\lesssim p^{K/2}r+pr^2+r^K.\end{aligned}\tag{43}$$

Hence, we have

$$\begin{aligned}\sum_{i\in[p]}\|\hat{\boldsymbol{X}}_{i:}^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2&\lesssim\sum_{i\in[p]}\theta(i)^2\|\hat{\boldsymbol{X}}_i^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2\\&\lesssim\frac{r^{K-1}}{p^{K-1}}\sum_{i\in[p]}\|\boldsymbol{X}_{i:}\|^2\|\hat{\boldsymbol{X}}_{i:}^s-\hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2\\&\lesssim\frac{r^{K-1}}{p^{K-1}}\left(p^{K/2}r+pr^2+r^K\right),\end{aligned}\tag{44}$$

where the first inequality follows from the assumption $\min_{i\in[p]}\theta(i)\geq c$, the second inequality follows from the inequality (35), and the last inequality comes from the inequality (43).

Next, we consider the following quantity,

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

$$\lesssim \frac{r^{K-1}}{p^{K-1}}\left(p^{K/2}r + pr^2 + r^K\right), \tag{45}$$

where the first inequality follows from the assumption of $\theta(i)$ and triangle inequality, the second inequality follows from Lemma 2, and the last inequality follows from (44). In addition, with Theorem 4 and the condition SNR $\gtrsim p^{-K/2}\log p$, for all $a \in [r]$, we have

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

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

$$\begin{aligned}
\|\hat{\boldsymbol{x}}_a - \boldsymbol{X}_a^s\|^2 &= \frac{\sum_{i \in z^{-1}(a) \cap (z^{(0)})^{-1}(a)} \left\|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}\right\|^2}{|z^{-1}(a) \cap (z^{(0)})^{-1}(a)|} \\
&\lesssim \frac{r}{p}\left(\sum_{i\in[p]} \|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{X}}_{i:}^s\|^2 + \sum_{i\in[p]} \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}\|^2\right) \\
&\lesssim \frac{r^K}{p^K}\left(p^{K/2}r + pr^2 + r^K\right),
\end{aligned} \tag{46}$$

where the last inequality follows from the inequality (44).

Finally, we obtain

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

where the first inequality follows from Lemma 4, and the third inequality follows from inequalities (45) and (46). □

Useful Definitions and Lemmas for the Proof of Theorem 4

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

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

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

$$\leq \frac{2\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|}{\|\boldsymbol{v}_2\|}.$$

Therefore, Lemma 2 is proved. □

Definition 3 (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 $\boldsymbol{w} = \{\boldsymbol{w}_i \colon \boldsymbol{w}_i = [\![w_{ik}]\!] \in \mathbb{R}^{p_i}\}_{i=1}^d$ as

$$\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a}) = [a_1 \circ \boldsymbol{w}_1, \ldots, a_d \circ \boldsymbol{w}_d]^T, \quad \text{where} \quad a_i \circ \boldsymbol{w}_i = [a_i w_{i1}, \ldots, a_i w_{ip_i}]^T, \text{ for all } i \in [d]. \tag{47}$$

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

$$\min_{i\in[d]} \|\boldsymbol{w}_i\|^2 \|\boldsymbol{a}\|^2 \leq \|\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a})\|^2 \leq \max_{i\in[d]} \|\boldsymbol{w}_i\|^2 \|\boldsymbol{a}\|^2. \tag{48}$$

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 3 (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

$$\frac{\min_{i\in[d]}\|\boldsymbol{w}_i\|}{\max_{i\in[d]}\|\boldsymbol{w}_i\|} \sin(\boldsymbol{a}, \boldsymbol{b}) \overset{*}{\leq} \sin(\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a}), \text{Pad}_{\boldsymbol{w}}(\boldsymbol{b})) \overset{**}{\leq} \frac{\max_{i\in[d]}\|\boldsymbol{w}_i\|}{\min_{i\in[d]}\|\boldsymbol{w}_i\|} \sin(\boldsymbol{a}, \boldsymbol{b}). \tag{49}$$

Proof of Lemma 3. We prove the two inequalities separately with similar ideas.

First, we prove the inequality ** in (49). Decomposing $\boldsymbol{b}$ yields

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

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

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

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

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

where the second inequality follows by applying the property (48) to vectors $\boldsymbol{b}$ and $\boldsymbol{a}^\perp$.

Next, we prove inequality * in (49). With the decomposition of $\text{Pad}_{\boldsymbol{w}}(\boldsymbol{b})$ and the inverse weighted padding operator, we have

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

Therefore, we obtain

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

where the second inequality follows by applying the property (48) to vectors $\boldsymbol{b}$ and $\text{Pad}_{\boldsymbol{w}}^{-1}((\text{Pad}_{\boldsymbol{w}}(\boldsymbol{a}))^\perp)$. □

Lemma 4 (Angle gap in $\mathcal{X}$)**.** Consider the dTBM model (2). Suppose Assumption 1 holds and $\boldsymbol{\theta}$ is balanced satisfying (7). Then the angle gap in $\mathcal{X}$ is asymptotically lower bounded by the angle gap in $\mathcal{S}$; i.e., for all i, j such that $z(i) \neq z(j)$

$$\|\boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:}\| \gtrsim \|\boldsymbol{S}^s_{z_1(i):} - \boldsymbol{S}^s_{z_1(j):}\| \gtrsim \Delta_{\min}.$$

Proof of Lemma 4. 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., $\mathrm{vec}(\mathcal{S}') = \boldsymbol{S}_{z(i):}$. Define weight vectors $\boldsymbol{w}_{a_2,\cdots,a_K}$ correspond to the elements in $\mathcal{S}'_{a_2,\ldots,a_K}$ by

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

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 $\boldsymbol{w} = \{\boldsymbol{w}_{a_2,\cdots,a_K}\}_{a_k\in[r],k\in[K]/\{1\}}$. Specifically, we have $\|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 = \prod_{k=2}^K \|\boldsymbol{\theta}_{z^{-1}(a_k)}\|^2$, and by the balanced assumption (7) we have

$$\max_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 = (1+o(1)) \min_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2. \tag{50}$$

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 (47) and the balanced assumption (50), we have

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

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

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

2) **Case 2:** Suppose $\left\langle \boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):}\right\rangle > 0$. 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

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

where the second inequality follows from Lemma 3, and the last inequality follows from the balanced weight (50) and Lemma 2.

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

$$\|\boldsymbol{X}^s_{i:} - \boldsymbol{X}^s_{j:}\| \gtrsim \|\boldsymbol{S}^s_{z_1(i):} - \boldsymbol{S}^s_{z_1(j):}\| \gtrsim \Delta_{\min}.$$

□

Lemma 5 (Singular value of weighted membership matrix)**.** Under the assumption that $\min_{i\in[p]}\theta(i) \geq c$, the singular values of $\boldsymbol{\Theta M}$ are bounded as

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

Proof of Lemma 5. Note that

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

By the definition of singular values, we have

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

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

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

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

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

Therefore, we complete the proof of Lemma 5. □

Lemma 6 (Singular-value gap-free tensor estimation error bound)**.** Consider an order-K tensor $\mathcal{A} = \mathcal{X} + \mathcal{Z} \in \mathbb{R}^{p\times\cdots\times p}$, where $\mathcal{X}$ has Tucker rank $(r, ...r)$ and $\mathcal{Z}$ has independent sub-Gaussian entries with parameter σ^2. Let $\hat{\mathcal{X}}$ denote the double projection estimated tensor in Step 2 of Sub-algorithm 1 in the main paper. Then with probability at least $1 - C\exp(-cp)$, we have

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

where C, c are some positive constants.

Proof of Lemma 6. See Han et al. (2020, Proposition 1). □

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

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

for some constant $c > 0$, we have

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

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

$$S = \{i \in S_0^c : \|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\| \ge c\}.$$

Proof of Lemma 7. 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

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

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

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

Following Gao et al. (2018, Lemma 6) **[FIXME (Miaoyan): Remove citation. fill in details in a self-contained way.]**, we have

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

and $|R_3|\le|R_1|$. We only need to bound $\sum_{i\in\cup_{u\in R_3}C_u}\theta(i)^2$ to finish the proof. Note that

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

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

Proof of Theorem 5

Proof of Theorem 5. We prove Theorem 5 under the symmetric dTBM (2) 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 variance $\sigma=1$, and that the identity permutation minimizes the initial misclustering error; i.e., $\pi^{(0)}=\arg\min_{\pi\in\Pi}\sum_{i\in[p]}\mathbb{1}\left\{z^{(0)}(i)\neq\pi\circ z(i)\right\}$ and $\pi^{(0)}(a)=a$ for all $a\in[r]$.

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

Notation.

1) Projection. We use $\boldsymbol{I}_d$ to denote the identity matrix of dimension d. 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 $\boldsymbol{v}^\perp$.

2) We define normalized membership matrices

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

and the dual normalized membership matrices

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

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

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

4) We define the matricizations of tensors

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

5) We define the angle-based misclustering loss in the t-th iteration

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

and the oracle loss

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

where m is a positive universal constant defined in (61).

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

$$\|\boldsymbol{E}\boldsymbol{V}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}}\left(p^{1/2}+r^{(K-1)/2}\right),\quad \|\boldsymbol{E}\boldsymbol{V}\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}},\quad \|\boldsymbol{W}_{a:}^T\boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}} \text{ for all } a\in[r], \tag{51}$$

$$\sup_{\boldsymbol{U}_k\in\mathbb{O}_{p,r},k=2,\ldots,K}\|\boldsymbol{E}(\boldsymbol{U}_2\otimes\cdots\otimes\boldsymbol{U}_K)\|_\sigma \lesssim \left(\sqrt{r^{K-1}}+K\sqrt{pr}\right), \tag{52}$$

$$\sup_{\boldsymbol{U}_k\in\mathbb{O}_{p,r},k=2,\ldots,K}\|\boldsymbol{E}(\boldsymbol{U}_2\otimes\cdots\otimes\boldsymbol{U}_K)\|_F \lesssim \left(\sqrt{pr^{K-1}}+K\sqrt{pr}\right), \tag{53}$$

$$\xi \lesssim \exp\left(-\frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}}\right), \tag{54}$$

$$L^{(t)} \lesssim \frac{\Delta_{\min}^2}{r\log p},\quad \text{for}\quad t=0,1,\ldots,T. \tag{55}$$

Further, inequality (51) holds by replacing $\boldsymbol{V}$ to $\boldsymbol{V}^{(t)}$ and $\boldsymbol{W}_{:a}$ to $\boldsymbol{W}_{:a}^{(t),T}$ when initialization condition (55) holds.

Step 2 (Misclustering loss decomposition). Next, we derive the 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 via

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

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

$$\|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \le \|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2. \tag{56}$$

Note that the event (56) also holds for the degenerate entity i with $\|\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}\|=0$ due to the convention that $\boldsymbol{a}^s=\boldsymbol{0}$ if $\boldsymbol{a}=\boldsymbol{0}$. Arranging the terms in (56) yields the decomposition

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

where

$$F_{ib}^{(t)} = 2\left\langle \boldsymbol{E}_{i:}\boldsymbol{V}^{(t)}, \left([\tilde{\boldsymbol{S}}_{z(i):}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\right) - \left([\tilde{\boldsymbol{S}}_{b:}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\right)\right\rangle + 2\left\langle \boldsymbol{E}_{i:}\left(\boldsymbol{V}-\boldsymbol{V}^{(t)}\right), [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s\right\rangle \tag{57}$$

$$\begin{aligned} G_{ib}^{(t)} = {} & \left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2 - \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\right) \\ & - \left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 - \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\right), \end{aligned} \tag{58}$$

$$H_{ib}^{(t)} = \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 - \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 + \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2. \tag{59}$$

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

$$\mathbb{1}\left\{z^{(t+1)}(i)=b\right\} \leq \mathbb{1}\left\{z^{(t+1)}(i)=b, \left\langle \boldsymbol{E}_{j:}\boldsymbol{V}, [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s \right\rangle \leq -\frac{1}{4}\|\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\|\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2\right\}$$
$$+ \mathbb{1}\left\{z^{(t+1)}(i)=b, \frac{1}{2}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2 \leq \|\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\|^{-1}F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)}\right\}. \quad (60)$$

Note that

$$\|\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\| = \theta(i)\|\boldsymbol{S}_{i:}(\boldsymbol{\Theta}\boldsymbol{M})^{\otimes(K-1),T}\boldsymbol{W}^{(t),\otimes^{K-1}}\| \geq \theta(i)\|\boldsymbol{S}_{z(i):}\|\lambda_r^{K-1}(\boldsymbol{\Theta}\boldsymbol{M})\lambda_r^{K-1}(\boldsymbol{W}^{(t)}) \geq \theta(i)m, \quad (61)$$

where the first inequality follows from the property of eigenvalues; the last inequality follows from Lemma 5, Lemma 9, and assumption that $\min_{a\in[r]}\|\boldsymbol{S}_{z(i):}\| \geq c_3 > 0$; and $m > 0$ is a positive constant related to c_3. Plugging the lower bound of $\|\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}\|$ (61) into the inequality (60) gives

$$\mathbb{1}\left\{z^{(t+1)}(i)=b\right\} \leq A_{ib} + B_{ib}, \quad (62)$$

where

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

Taking the weighted summation of (62) over $i \in [p]$ yields

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

where ξ is the oracle loss such that

$$\xi = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]/z(i)}A_{ib}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2. \quad (63)$$

Similarly to ξ in (63), we define

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

Step 3 (Derivation of contraction inequality). In this step we derive the upper bound of ζ_{ib} and obtain the contraction inequality (19). Note that

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

where the first inequality follows from the inequality (79) in Lemma 10, and the last inequality follows from the assumption that $\min_{i\in[p]}\theta(i) \geq c > 0$. Following Han et al. (2020, Step 4, Proof of Theorem 2) and Lemma 10, we have

$$\frac{1}{p}\sum_{i\in[p]}\sum_{b\in[r]/z(i)}\mathbb{1}\left\{z^{(t+1)}(i)=b\right\}\frac{(F_{ib}^{(t)})^2}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \leq C_1L^{(t)}, \quad (65)$$

and

$$\frac{1}{p}\sum_{i\in[p]}\sum_{b\in[r]/z(i)}\mathbb{1}\left\{z^{(t+1)}(i)=b\right\}\frac{\theta(i)(G_{ib}^{(t)})^2}{\|[\boldsymbol{S}_{z(i)\cdot}]^s-[\boldsymbol{S}_{b\cdot}]^s\|^2}\lesssim\frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]/z(i)}\mathbb{1}\left\{z^{(t+1)}(i)=b\right\}(\Delta_{\min}^2+L^{(t)})$$
$$\leq C_2(L^{(t+1)}+L^{(t)}), \tag{66}$$

where the last inequality follows from the definition of $L^{(t)}$ and the constraint of $\boldsymbol{\theta}$ in parameter space (3). Choosing the constants C, C_1, C_2 in (64) (65) (66) carefully **[FIXME (Miaoyan): be specific. Add one sentence to explain]**, we have

$$\frac{1}{p}\sum_{i\in[p]}\sum_{b\in[r]/z(i)}\zeta_{ib}^{(t)}\leq C_4L^{(t+1)}+C_5L^{(t)}, \tag{67}$$

for some small **[FIXME (Miaoyan): In what sense "small"? Add one sentence.]** positive constants C_4, C_5. Plugging the inequality (67) into the decomposition (63), we obtain the contraction inequality

$$L^{(t+1)}\leq M\xi+\rho L^{(t)}, \tag{68}$$

where $M\geq 1$ and $\rho\in[0,1)$ is the contraction parameter.

Combining the inequality (54) in Condition 1 and Lemma 8, we have proven Theorem 5 with inequality (68).

Step 4 (Verification of Condition 1). Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (51), (52), and (53) describe the property of the sub-Gaussian noise tensor $\mathcal{E}$, and the readers can find the proof directly in Han et al. (2020, Step 5, Proof of Theorem 2). Here, we include only the verification of inequalities (54) and (55).

Now, we verify the oracle loss condition (54). Recall the definition of ξ,

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

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

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

where

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

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

$$P_1\lesssim\exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(j)\cdot}]^s-[\boldsymbol{S}_{b\cdot}]^s\|^2\right). \tag{69}$$

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

$$P_2\leq P_{21}+P_{22}, \tag{70}$$

where with some positive constant $t > 0$,

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

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

$$P_{21} \le \mathbb{P}\left(t\|\boldsymbol{S}_{z(i):}\| \le \|\tilde{\boldsymbol{S}}_{z(i):} - \boldsymbol{S}_{z(i):}\|\right) \le \mathbb{P}\left(c_3 t \le \|\boldsymbol{W}_{:z(i)}^T \boldsymbol{E}\boldsymbol{V}\|\right) \lesssim \exp\left(-\frac{p^K t^2}{r^K}\right), \tag{71}$$

where the first inequality follows from the basic inequality in Lemma 2, the second inequality follows from the assumption that $\min_{a\in[r]}\|\boldsymbol{S}_{z(i):}\| \ge c_3 > 0$ in (3), and the last inequality follows from the Bernstein inequality.

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

$$P_{22} \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}t^2}\|[\boldsymbol{S}_{z(j):}]^s - [\boldsymbol{S}_{b:}]^s\|^4\right). \tag{72}$$

We take $t = \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|$ in P_{21} and P_{22}, and plug the inequalities (71) and (72) into to the upper bound for P_2 in (70). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2\right). \tag{73}$$

Combining the upper bounds (69) and (73) gives

$$\mathbb{P}\left(\left\langle e_i, [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s\right\rangle \le -\frac{\theta(i)m}{4}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2\right) \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2\right). \tag{74}$$

Hence, we have

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

where the first inequality follows from the constraint that $\sum_{i\in[p]}\theta(i) = p$, and the last inequality follows from (74).

By Markov's inequality, we have

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

and thus the condition (54) 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 (55) by induction. With output $z^{(0)}$ from Sub-algorithm 2 and the assumption SNR $\gtrsim p^{-K/2}\log p$, by Corollary 1, we have

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

Therefore, the condition (55) holds for $t = 0$. Assume the condition (55) also holds for all $t \le t_0$. Then, by the decomposition (68), 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 from the condition (54) and the last inequality follows from the assumption that $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$. Thus, the condition (55) holds for $t_0 + 1$, and the condition (55) is proved by induction. □

Useful Definitions and Lemmas for the Proof of Theorem 5

Lemma 8 (Misclustering error and loss)**.** Define the misclustering error in the t-th iteration as $\ell^{(t)} = \ell(z^{(t)}, z)$. We have

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

Proof of Lemma 8. By the definition of minimal gap in Assumption 1, we have

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

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

Lemma 9 (Singular-value property of membership matrices)**.** Suppose the condition (55) holds. Then, for all $a \in [r]$, we have $|\left(z^{(t)}\right)^{-1}(a)| \asymp p/r$. Moreover, we have

$$\lambda_r(\boldsymbol{M}) \asymp \|\boldsymbol{M}\|_\sigma \asymp \sqrt{p/r}, \quad \lambda_r(\boldsymbol{W}) \asymp \|\boldsymbol{W}\|_\sigma \asymp \sqrt{r/p}. \tag{75}$$

The inequalities (75) also hold by replacing $\boldsymbol{M}$ and $\boldsymbol{W}$ to $\boldsymbol{M}^{(t)}$ and $\boldsymbol{W}^{(t)}$ respectively. Further, we have

$$\lambda_r(\boldsymbol{W}\boldsymbol{W}^T) \asymp \left\|\boldsymbol{W}\boldsymbol{W}^T\right\|_\sigma \asymp r/p, \tag{76}$$

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

Proof of Lemma 9. The proof for the inequality (75) can be found in Han et al. (2020, Proof of Lemma 4)

For inequality (76), note that for all $k \in [r]$,

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

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

Lemma 10 (Upper bound for $F_{ib}^{(t)}, G_{ib}^{(t)}$ and $H_{ib}^{(t)}$)**.** Under the Condition 1, we have

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left(F_{ib}^{(t)}\right)^2}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2}\|\boldsymbol{E}_{i:}\boldsymbol{V}\|^2 + \left(1 + \frac{rL^{(t)}}{\Delta_{\min}^2}\right)\|\boldsymbol{E}_{i:}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\|^2, \tag{77}$$

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left(G_{ib}^{(t)}\right)^2}{\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2} \lesssim \left(\Delta_{\min}^2 + L^{(t)}\right), \tag{78}$$

$$\max_{i\in[p]}\max_{b\neq z(i)}\frac{\left|H_{ib}^{(t)}\right|}{\|[\boldsymbol{S}_{z(i):}]^s-[\boldsymbol{S}_{b:}]^s\|^2}\leq\frac{1}{4}. \tag{79}$$

Proof of Lemma 10. We prove the each of the inequalities in Lemma 10 separately.

1) Upper bound for $F_{ib}^{(t)}$, i.e., inequality (77). Recall the definition of $F_{ib}^{(t)}$,

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

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left(F_{ib}^{(t)}\right)^2&\leq 8\left(\left\langle\boldsymbol{E}_{i:}\boldsymbol{V}^{(t)},\left([\tilde{\boldsymbol{S}}_{z(i):}]^s-[\boldsymbol{S}_{z(i):}^{(t)}]^s\right)-\left([\tilde{\boldsymbol{S}}_{b:}]^s-[\boldsymbol{S}_{b:}^{(t)}]^s\right)\right\rangle\right)^2\\&\quad+8\left(\left\langle\boldsymbol{E}_{i:}(\boldsymbol{V}-\boldsymbol{V}^{(t)}),[\tilde{\boldsymbol{S}}_{z(i):}]^s-[\tilde{\boldsymbol{S}}_{b:}]^s\right\rangle\right)^2\\&\lesssim\left(\|\boldsymbol{E}_{i:}\boldsymbol{V}\|^2+\|\boldsymbol{E}_{i:}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\|^2\right)\max_{a\in[r]^s}\|[\tilde{\boldsymbol{S}}_{a:}]^s-[\boldsymbol{S}_{a:}^{(t)}]^s\|\\&\quad+\|\boldsymbol{E}_{i:}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\|^2\|[\tilde{\boldsymbol{S}}_{z(i):}]^s-[\tilde{\boldsymbol{S}}_{b:}]^s\|.\end{aligned} \tag{80}$$

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

$$\begin{aligned}\|[\tilde{\boldsymbol{S}}_{a:}]^s-[\boldsymbol{S}_{a:}^{(t)}]^s\|^2&=\|[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\\&\leq 2\|[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s\|^2+2\|[\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s-[\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\\&\lesssim\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2}+\frac{rr^{2K}+pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\\&\lesssim rL^{(t)}+\frac{rr^{2K}+pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\\&\lesssim rL^{(t)},\end{aligned} \tag{81}$$

where the second inequality follows from the inequalities (93) and (94) in Lemma 11, the third inequality follows from the condition (55) in Condition 1, and the last inequality follows from the assumption that $\Delta_{\min}^2\gtrsim p^{-K/2}\log p$.

Note that

$$\begin{aligned}\|[\tilde{\boldsymbol{S}}_{z(i):}]^s-[\tilde{\boldsymbol{S}}_{b:}]^s\|^2&=\|[\tilde{\boldsymbol{S}}_{z(i):}]^s-[\boldsymbol{S}_{z(i):}]^s+[\boldsymbol{S}_{z(i):}]^s-[\boldsymbol{S}_{b:}]^s+[\boldsymbol{S}_{b:}]^s-[\tilde{\boldsymbol{S}}_{b:}]^s\|^2\\&\lesssim\|[\boldsymbol{S}_{z(i):}]^s-[\boldsymbol{S}_{b:}]^s\|^2+\max_{a\in[r]}\|[\boldsymbol{S}_{a:}]^s-[\tilde{\boldsymbol{S}}_{a:}]^s\|^2\\&\lesssim\|[\boldsymbol{S}_{z(i):}]^s-[\boldsymbol{S}_{b:}]^s\|^2+\max_{a\in[r]}\frac{1}{\|\boldsymbol{S}_{a:}\|^2}\|\boldsymbol{W}_{:a}^T\boldsymbol{E}\boldsymbol{V}\|^2\\&\lesssim\|[\boldsymbol{S}_{z(i):}]^s-[\boldsymbol{S}_{b:}]^s\|^2,\end{aligned} \tag{82}$$

where the second inequality follows from Lemma 2, and the last inequality follows from the assumptions on $\|\boldsymbol{S}_{a:}\|$ in the parameter space (3), the inequality (51) in Condition 1 and the assumption $\Delta_{\min}^2\gtrsim p^{-K/2}\log p$.

Therefore, we finish the proof of inequality (77) by plugging the inequalities (81) and (82) into the upper bound (80).

2) Upper bound for $G_{ib}^{(t)}$, i.e., inequality (78). By definition of $G_{ib}^{(t)}$, we rearrange terms and obtain

$$\begin{aligned}G_{ib}^{(t)}&=\left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2-\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\right)\\&\quad-\left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{S}_{b:}^{(t)}]^s\|^2-\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2\right)\\&=2\left\langle[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s,\left([\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{S}_{z(i):}^{(t)}]^s\right)-\left([\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{S}_{b:}^{(t)}]^s\right)\right\rangle\\&=G_1+G_2-G_3,\end{aligned} \tag{83}$$

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

$$
\begin{aligned}
|G_1|^2 &\le \left| \|[\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2 - \|[\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \right|^2 \\
&\le \max_{a\in[r]} \|[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T} \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^4 \\
&\lesssim \frac{r^4}{\Delta_{\min}^4}(L^{(t)})^4 + \frac{r^2 r^{4K} + p^2 r^{2K+4}}{p^{2K}} \frac{(L^{(t)})^2}{\Delta_{\min}^4} \\
&\le c\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right),
\end{aligned} \tag{84}
$$

where the third inequality follows from the inequality (95) in Lemma 9 and the last inequality follows from the assumption that $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$ and inequality (55) in Condition 1.

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

$$
\begin{aligned}
|G_2|^2 &\le 2\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 \|[\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2 \\
&\le \frac{2}{\|\boldsymbol{W}_{z(i):}^T \boldsymbol{X}\boldsymbol{V}^{(t)}\|^2} \max_{a\in[r]} \|\boldsymbol{W}_{:a}^T \boldsymbol{E}\boldsymbol{V}^{(t)}\|^2 \max_{a\in[r]} \|[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T} \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 \\
&\lesssim \frac{r^{2K-1} + Kpr^{K+1}}{p^K} \left(\frac{r^2}{\Delta_{\min}^2}(L^{(t)})^2 + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) \\
&\le c\Delta_{\min}^2 L^{(t)},
\end{aligned} \tag{85}
$$

where the second inequality follows from Lemma 2, the third inequality follows from the inequality (52) in Condition 1, the inequalities (95) and (114) in the proof of Lemma 11, and the last inequality follows from the assumption $\Delta_{\min}^2 \ge p^{-K/2}\log p$ and inequality (55) in Condition 1.

For G_3, note that by triangle inequality

$$
\begin{aligned}
\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2 &\le \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + 2\max_{a\in[r]} \|[\boldsymbol{W}_{:a}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^T \boldsymbol{X}\boldsymbol{V}]^s\|^2 \\
&\lesssim \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2},
\end{aligned} \tag{86}
$$

where the last inequality follows from the inequality (113) in the proof of Lemma 11. Then we have

$$
\begin{aligned}
|G_3|^2 &\le 2\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 \max_{a\in[r]} \|[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T} \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|^2 \\
&\le 2\left(\|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2 + \|[\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s\|^2 \right) \\
&\quad \times \max_{a\in[r]} \left\| [\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:a}^{(t),T} \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \right\|^2 \\
&\lesssim \left(\|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 + \frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} \right) \left(\frac{r^2 (L^{(t)})^2}{\Delta_{\min}^2} + \frac{r r^{2K} + p r^{K+2}}{p^K} \frac{L^{(t)}}{\Delta_{\min}^2} \right) + c\Delta_{\min}^2 L^{(t)} \\
&\lesssim \|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{b:}^s\|^2 (\Delta_{\min}^2 + L^{(t)}) + c\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right),
\end{aligned} \tag{87}
$$

where the third inequality follows from the same procedure to derive (84) and (85), and the last inequality follows from the assumption $\Delta_{\min}^2 \ge p^{-K/2}\log p$ and inequality (55) in Condition 1.

Plugging the inequalities (84), (85), and (87) into the upper bound (83), we finish the proof of inequality (78).

3) Upper bound for $H_{ib}^{(t)}$, i.e., the inequality (79). By definition of H_{ib}, we rearrange terms and obtain

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

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

$$|H_1| \leq \frac{4\max_{a\in[r]}\|\boldsymbol{W}_{:a}^T\boldsymbol{E}\boldsymbol{V}^{(t)}\|^2}{\|\boldsymbol{W}_{z(i):}^T\boldsymbol{X}\boldsymbol{V}^{(t)}\|^2} \leq \frac{r^{2K-1} + Kpr^{K+1}}{p^K} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2, \tag{88}$$

following the derivation of G_2 in inequality (85) and the assumption that $\Delta_{\min}^2 \geq p^{-K/2}\log p$.

For H_2, by the inequality (86), we have

$$|H_2| \lesssim 2\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2 + \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2, \tag{89}$$

where the last inequality follows from the condition (55) in Condition 1.

For H_3, by Cauchy-Schwartz inequality, we have

$$|H_3| \lesssim \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\||H_1|^{1/2} \lesssim \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{a:}]^s\|^2, \tag{90}$$

following the inequalities (86) and (88).

Therefore, combining inequalities (88), (89), and (90), we have finish the proof of inequality (79).

□

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

$$\|\boldsymbol{V} - \boldsymbol{V}^{(t)}\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}}\frac{r}{\Delta_{\min}^2}L^{(t)}, \tag{91}$$

$$\|\boldsymbol{E}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1} + pr)}{p^{K-1}}}\frac{r}{\Delta_{\min}^2}L^{(t)}, \tag{92}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}]^s - [\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s\| \lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}, \tag{93}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s - [\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\| \lesssim \sqrt{\frac{rr^{2K} + pr^{K+2}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{rL^{(t)}}{\Delta_{\min}}, \tag{94}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s-[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\|\lesssim\frac{rL^{(t)}}{\Delta_{\min}}+\sqrt{\frac{rr^{2K}+pr^{K+2}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{95}$$

In addition, the inequality (94) also holds by replacing $\boldsymbol{W}_{:b}^{(t)}$ to $\boldsymbol{W}_{:b}$.

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

1) Inequality (91). 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 (91) by applying Lemma 8 to the above inequality.

2) Inequality (92). By Han et al. (2020, Proof of Lemma 5), we have

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

Also, we complete the proof of inequality (91) by applying Lemma 8 to the above inequality.

3) Inequality (93). We upper bound the desired quantity by triangle inequality,

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

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,

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

where

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

We first consider I_{11}. 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 9, we have $\sum_{i\in[p]}\mathbb{1}\left\{z^{(t)}(i)=b\right\}\gtrsim p/r$. Then, we have

$$\sum_{a\neq b,a,b\in[r]}D_{ab}\lesssim\frac{r}{p}\sum_{i:\ z^{(t)}(i)\neq z(i)}\theta(i)\lesssim\frac{L^{(t)}}{\Delta_{\min}^2}\lesssim\frac{1}{\log p}, \tag{96}$$

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

$$D_{bb} = \frac{\sum_{i\in[p]} \theta(i)\mathbb{1}\left\{z(i) = z^{(t)}(i) = b\right\}}{\sum_{i\in[p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \geq \frac{c(\sum_{i\in[p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} - p\ell^{(t)})}{\sum_{i\in[p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \gtrsim 1 - \frac{1}{\log p}, \tag{97}$$

under the inequality (55) 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 roughly 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 (3) and the bounds of $\boldsymbol{D}$ in (96) and (97).

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

$$\begin{aligned}
\|[D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}]\sin\alpha\| &= \|(\boldsymbol{I}_d - \text{Proj}(\boldsymbol{S}_{b:})) \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}\| \\
&\leq \sum_{a\neq b, a\in[r]} D_{ab} \|(\boldsymbol{I}_d - \text{Proj}(\boldsymbol{S}_{b:}))\boldsymbol{S}_{a:}\| \\
&= \sum_{a\neq b, a\in[r]} D_{ab} \|\boldsymbol{S}_{a:}\sin(\boldsymbol{S}_{b:}, \boldsymbol{S}_{a:})\| \\
&\leq \sum_{a\neq b, a\in[r]} D_{ab} \|\boldsymbol{S}_{a:}\| \|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\|,
\end{aligned} \tag{98}$$

where the first inequality follows from the triangle inequality, and the last inequality follows from Lemma 2. Note that with bounds (96) and (97), when p is large enough, we have

$$\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\| = \|D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b, a\in[r]} D_{ab}\boldsymbol{S}_{a:}\| \geq D_{bb}\|\boldsymbol{S}_{b:}\| - \sum_{a\neq b, a\in[r]} D_{ab}\|\boldsymbol{S}_{a:}\| \geq C_1, \tag{99}$$

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

$$\begin{aligned}
I_{11} &\leq \sqrt{2}\sin\alpha \\
&= \frac{\|[D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b,a\in[r]} D_{ab}\boldsymbol{S}_{a:}]\sin\alpha\|}{\|D_{bb}\boldsymbol{S}_{b:} + \sum_{a\neq b,a\in[r]} D_{ab}\boldsymbol{S}_{a:}\|} \\
&\leq \frac{1}{C_1} \sum_{a\neq b, a\in[r]} D_{ab}\|\boldsymbol{S}_{a:}\| \|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\| \\
&\lesssim \frac{r}{p} \sum_{i\in[p]} \theta(i) \sum_{b\in[r]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} \|\boldsymbol{S}_{b:}^s - \boldsymbol{S}_{a:}^s\| \\
&\leq \frac{rL^{(t)}}{\Delta_{\min}},
\end{aligned} \tag{100}$$

where the second inequality follows from (98) and (99), and the last two inequalities follow by the definition of D_a and $L^{(t)}$, and the constraint of $\|\boldsymbol{S}_{b:}\|$ in parameter space (3).

We now 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

$$\|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}\| \lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{101}$$

Notice that

$$\|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}\| \leq \|\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T}\|\,\|\boldsymbol{X}\boldsymbol{V}\|_F \lesssim \frac{r^{3/2}L^{(t)}}{\sqrt{p}\Delta_{\min}^2}\|\boldsymbol{S}\|\|\boldsymbol{\Theta}\boldsymbol{M}\|_\sigma \lesssim \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}, \tag{102}$$

where the second inequality follows from Han et al. (2020, Inequality (121), Proof of Lemma 5) and the last inequality follows from Lemma 5 and (55) 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 (99). Therefore, we have

$$\begin{aligned} I_{12} &\lesssim \|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}\| + \|(\boldsymbol{W}_{:b}^T - \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}\|\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}} + \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}\sqrt{\frac{r^{2K}}{p^K}} \\ &\lesssim \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}, \end{aligned} \tag{103}$$

where second inequality follows from the inequalities (101), (102), and (51) in Condition 1.

Hence, combining inequalities (100) and (103) yields

$$I_1 \lesssim \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\sqrt{L^{(t)}}}{\Delta_{\min}}. \tag{104}$$

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 (99). By triangle inequality and (51) in Condition 1, we have

$$I_2 \leq \frac{\|\boldsymbol{W}_{:b}^T \boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^T \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{105}$$

and

$$I_3 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{106}$$

Therefore, combining the inequalities (104), (105), and (106), we finish the proof of inequality (93).

4) Inequality (94). Here we only show the proof of inequality (94) 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

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

where

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

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

For J_1, by triangle inequality, we have

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

where

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

We first consider J_{11}. Define the matrix $\boldsymbol{V}^k \coloneqq \boldsymbol{W}^{\otimes(k-1)} \otimes \boldsymbol{W}^{(t),\otimes(K-k)}$ for $k = 2, \dots, K-1$, and denote $\boldsymbol{V}^1 = \boldsymbol{V}^{(t)}, \boldsymbol{V}^K = \boldsymbol{V}$. Also, 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 $k = 1, \dots, K-1$. 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} in inequality (100), we bound J_{11}^k by 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

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

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

$$\begin{aligned}
\left\langle \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^k, \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{k+1} \right\rangle &= \langle \text{Mat}_{k+1}(\mathcal{T}_k), \text{Mat}_{k+1}(\mathcal{T}_{k+1}) \rangle \\
&= \left\langle \boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^k, \boldsymbol{S}\boldsymbol{Z}^k \right\rangle \\
&= \sum_{b\in[r]} \left(D_{bb}\|\boldsymbol{S}_{b:}\boldsymbol{Z}^k\|^2 + \sum_{a\neq b, a\in[r]} D_{ab} \left\langle \boldsymbol{S}_{a:}\boldsymbol{Z}^k, \boldsymbol{S}_{b:}\boldsymbol{Z}^k \right\rangle \right) \\
&\gtrsim (1 - \log p^{-1}) \min_{a\in[r]} \|\boldsymbol{S}_{a:}\boldsymbol{Z}^k\|^2 - \log p^{-1} \max_{a\in[r]} \|\boldsymbol{S}_{a:}\boldsymbol{Z}^k\|^2, \qquad (107)
\end{aligned}$$

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 definitions, and the last inequality follows from the bounds of $\boldsymbol{D}$ in (96) and (97).

Note that

$$\|\boldsymbol{D}\|_\sigma \le \|\boldsymbol{D}\|_F \le \sqrt{\sum_{b\in[r]} D_{bb}^2 + (\sum_{a\neq b, a, b\in[r]} D_{ab})^2} \lesssim \sqrt{r + \log^2 p^{-1}} \lesssim 1, \qquad (108)$$

where the second inequality follows from inequality (96), and the fact that for all $b \in [r]$,

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

Also, we have

$$\lambda_r(\boldsymbol{D}) \geq \lambda_r(\boldsymbol{W}^{(t)})\lambda_r(\boldsymbol{\Theta}\boldsymbol{M}) \gtrsim 1, \tag{109}$$

following the Lemma 5 and Lemma 9. Then, for all $k \in [K]$, we have

$$1 \lesssim \|\boldsymbol{D}_{:b}\|\, \lambda_r(\boldsymbol{D})^{K-k-1} \leq \lambda_{r^{K-2}}(\boldsymbol{Z}^k) \leq \|\boldsymbol{Z}^k\|_\sigma \leq \|\boldsymbol{D}_{:b}\|\, \|\boldsymbol{D}\|_\sigma^{K-k-1} \lesssim 1. \tag{110}$$

Thus, we have bounds

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

Hence, when p is large enough, the inner product (107) is positive, which implies $\beta_k \in [0, \pi/2)$ and thus $2\sin\frac{\beta_k}{2} \leq \sqrt{2}\sin\beta_k$.

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

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

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 $\tilde{\boldsymbol{D}}_{ab} = \frac{\sum_{i\in[p]}\theta(i)\mathbb{1}\{z^{(t)}=b, z(i)=a\}}{\sum_{i\in[p]}\theta(i)\mathbb{1}\{z^{(t)}=b\}}$.

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

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

Note that we have $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \log^{-1}(p)$ by Condition 1 and Lemma 8. With the locally linear stability assumption, we have

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

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

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

where the last inequality follows from Lemma 9, the constraint $\min_{i\in[p]}\theta(i) \geq c > 0$, $\|\boldsymbol{p}(z^{(t)})\| \lesssim p$ and $\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{(t),-1}(a)}\|_1 \gtrsim p$. By the geometry property of trigonometric function, we have

$$\sin\beta_{k1} = \min_{c\in\mathbb{R}} \frac{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\otimes k-1} \otimes (\boldsymbol{D} - c\tilde{\boldsymbol{D}}) \otimes \boldsymbol{D}^{\otimes K-k-1}\|}{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\otimes k-1} \otimes \boldsymbol{D}^{\otimes K-k}\|} \tag{111}$$

$$
\begin{aligned}
&\leq \frac{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\|\|\boldsymbol{D}-c_0\tilde{\boldsymbol{D}}\|_\sigma\|\boldsymbol{D}\|_\sigma^{K-k-1}}{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\|\lambda_r^{K-k}(\boldsymbol{D})}\\
&\lesssim \|\boldsymbol{D}-c_0\tilde{\boldsymbol{D}}\|_F\\
&\lesssim \frac{L^{(t)}}{\Delta_{\min}},
\end{aligned}
$$

where the second inequality follows from the singular property of $\boldsymbol{D}$ in (108), (109) and the constraint of $\boldsymbol{S}$ in (3). To bound $\sin\beta_{k2}$, let $\boldsymbol{C}=\mathrm{diag}(\{\|\boldsymbol{S}_{a:}\|\}_{a\in[r]})$. We have

$$
\begin{aligned}
\sin\beta_{k2} &\lesssim \frac{\left\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\otimes k-1}\otimes(\boldsymbol{I}_r-\tilde{\boldsymbol{D}})\otimes\boldsymbol{D}^{\otimes K-k-1}\right\|}{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\otimes k}\otimes\boldsymbol{D}^{\otimes K-k-1}\|} \qquad (112)\\
&\lesssim \frac{\|(\boldsymbol{I}_r-\tilde{\boldsymbol{D}}^T)\boldsymbol{S}\boldsymbol{Z}^k\|_F}{\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\|\lambda_r^{K-k-1}(\boldsymbol{D})}\\
&\lesssim \|(\boldsymbol{I}_r-\tilde{\boldsymbol{D}}^T)\boldsymbol{S}\boldsymbol{C}^{-1}\|_F\|\boldsymbol{C}\boldsymbol{Z}^k\|_\sigma\\
&\lesssim \frac{r}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}\|\boldsymbol{S}_{b:}^s-\boldsymbol{S}_{z(i):}^s\|\\
&\lesssim \frac{L^{(t)}}{\Delta_{\min}},
\end{aligned}
$$

where the third inequality follows from the singular property of $\boldsymbol{D}$ and the boundedness of $\boldsymbol{S}$, and the fourth inequality follows from the definition of $\tilde{\boldsymbol{D}}$, boundedness of $\boldsymbol{S}$, the lower bound of $\boldsymbol{\theta}$, and the singular property of $\boldsymbol{Z}^k$ in inequality (110).

Combining (111) and (112) yields

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

Finally, by triangle inequality, we obtain

$$
J_{11} \leq \sum_{k=1}^{K-1} J_{11}^k \lesssim \sum_{k=1}^{K-1}\sin\beta_k \lesssim (K-1)\frac{rL^{(t)}}{\Delta_{\min}}. \qquad (113)
$$

We now 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

$$
\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\| = \|\boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^1\| \geq \lambda_r(\boldsymbol{D})\,\|\boldsymbol{S}\|\,\lambda_{r^{K-2}}(\boldsymbol{Z}^1) \gtrsim 1, \qquad (114)
$$

where the inequality follows from the bounds (109) and (110).

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

$$
\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}(\boldsymbol{V}-\boldsymbol{V}^{(t)})\| \lesssim \sqrt{\frac{r^{2K+1}+pr^{2+K}}{p^K}}\frac{(K-1)\sqrt{L^{(t)}}}{\Delta_{\min}}. \qquad (115)
$$

Notice that

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

$$\lesssim \frac{\sqrt{rL^{(t)}}}{\Delta_{\min}}, \tag{116}$$

where the first inequality follows from the tensor algebra in inequality (107), the second inequality follows from the fact that $\boldsymbol{I} = \boldsymbol{W}^T \boldsymbol{\Theta} \boldsymbol{M}$, and the last inequality follows from Han et al. (2020, Proof of Lemma 5). It follows from (116) and Lemma 9 that

$$\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\| \leq \|\boldsymbol{W}_{:b}^{(t),T}\| \sum_{k=1}^{K-1} \|\boldsymbol{X}(\boldsymbol{V}^k - \boldsymbol{V}^{k+1})\|_F \lesssim \frac{\sqrt{rL^{(t)}}}{\sqrt{p}\Delta_{\min}}. \tag{117}$$

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

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

where the second inequality follows from inequalities (115), (117), and the inequality (51) 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 (99) and (114), respectively. By triangle inequality and inequality (51) in Condition 1, we have

$$J_2 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{118}$$

and

$$J_3 \leq \frac{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}^{(t)}\|}{\|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{X}\boldsymbol{V}^{(t)}\|} \lesssim \|\boldsymbol{W}_{:b}^{(t),T} \boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{119}$$

Therefore, combining the inequalities (113), (118), and (119), we finish the proof of inequality (94).

5) Inequality (95). By triangle inequality, we upper bound the desired quantity

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

following the inequalities (93) and (94). Therefore, we finish the proof of inequality (95).

□