Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu and Miaoyan Wang

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

The classical tensor block model suffers from drawbacks to model real world data in spite of the popularity. The key underlying assumption of block model is that all nodes in the same community are exchangeable; i.e., the nodes have no individual-specific parameters apart from the community-specific parameters. However, the exchangeability assumption is often non-realistic. Each node may contribute to the data variation by its own multiplicative effect. We call the unequal node-specific effects the *degree heterogeneity*. Such degree heterogeneity appears commonly in social networks. Ignoring the degree heterogeneity may seriously mislead the clustering results. For example, the regular block model fails to model the member affiliation in the Karate Club network [**?**] without addressing degree heterogeneity.

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity [**?**]. 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

The work of Miaoyan Wang was supported in part by NSF CAREER DMS-2141865, DMS-1915978, DMS-2023239, EF-2133740, and funding from the Wisconsin Alumni Research foundation. (*Corresponding author: Miaoyan Wang.*)

Jiaxin Hu and Miaoyan Wang are with the Department of Statistics, University of Wisconsin-Madison, Madison, WI, 53706 USA. (e-mail: jhu267@wisc.edu, miaoyan.wang@wisc.edu)

This paper was presented in part at 25th International Conference on Artificial Intelligence and Statistics (AISTATS).

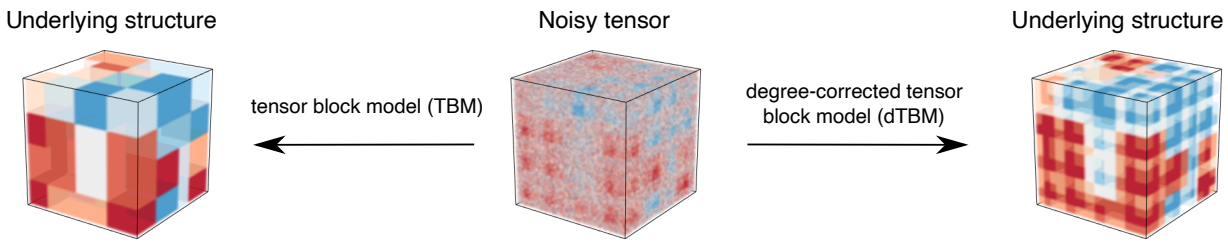

Fig. 1: Examples for order-3 tensor block model (TBM) with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

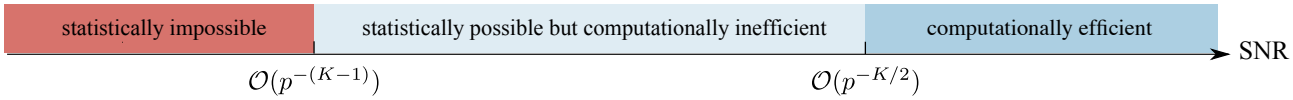

Fig. 2: SNR thresholds for statistical and computational limits in order-K dTBM with dimension $(p, ..., p)$ and $K \geq 2$. The SNR gap between statistical possibility and computational efficiency exists only for tensors with $K \geq 3$.

with respect to three different statistical and computational behaviors. We characterize, for the first time, the critical signal-to-noise (SNR) thresholds in dTBMs, revealing the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering. Specific SNR thresholds and algorithm behaviors are depicted in Figure 2.

- We provide an angle-based algorithm that achieves exact clustering *in polynomial time* under mild conditions. Simulation and data studies demonstrate that our algorithm outperforms existing higher-order clustering algorithms.

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

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

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

- *Tensor block model.* The (non-degree) tensor block model (TBM) is a higher-order extension of SBM, and its statistical-computational properties are investigated in recent literatures [**? ? ?**]. Some works [**?**] study the TBM with sparse observations, while, others [**? ?**] and our work focus on the dense regime. Extending results from non-degree to degree-corrected model is highly challenging. Our dTBM parameter space is equipped with angle-based similarity and nuisance degree parameters. The extra complexity makes the Cartesian coordinates based analysis [**?**] non-applicable to our setting. Towards this goal, we have developed a new polar coordinates based analysis to control the model complexity. We ~~also develop~~ have also developed a new angle-based iteration algorithm to achieve optimal clustering rates *without the need of estimating nuisance degree parameters*.

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

- *Global-to-local algorithm strategy.* Our methods generalize the recent global-to-local strategy for matrix learning [**? ? ?**] to tensors [**? ? ?**]. 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 [**?**], 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

	Gao et al. (2018)[?]	Ahn et al. (2018)[?]	Han et al. (2020)[?]	Ghoshdastidar et al. (2019)[?]	Ke et al. (2019)[?]	**Ours**
Allow tensors of arbitrary order	×	√	√	√	√	√
Allow degree heterogeneity	√	×	×	√	√	√
Singular-value gap-free clustering	√	√	√	×	×	√
Misclustering rate (for order K^*)	-	$p^{-(K-1)}\alpha^{-1}$**	$\exp(-p^{K/2})$	p^{-1}	p^{-2}	$\exp(-p^{K/2})$
Consider sparse observation	×	√	×	×	×	×

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 $r = \mathcal{O}(1)$. **The parameter $\alpha = f(p) > 0$ denotes the sparsity level which is some function of dimension p.

$\|\boldsymbol{a}\| = \sqrt{\sum_i a^2(i)}$ to denote the ℓ_2-norm, $\|\boldsymbol{a}\|_1 = \sum_i |a_i|$ to denote the ℓ_1 norm of $\boldsymbol{a}$. For two vector $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension, we denote the angle between $\boldsymbol{a}, \boldsymbol{b}$ by

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

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

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~~ K $(p_1, \ldots, p_K)$-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 $\mathcal{X} \in \mathbb{R}^{p \times \cdots \times p}$, we omit the subscript and use $\text{Mat}(\mathcal{X}) \in \mathbb{R}^{p \times p^{K-1}}$ to denote the unfolding. For two sequences $\{a_p\}, \{b_p\}$, we denote $a_p \lesssim b_p$ or $a_p = \mathcal{O}(b_p)$ if $\lim_{p \to \infty} a_p / b_p \leq c$~~for some constant $c \geq 0$~~, $a_p \gtrsim b_p$ or $a_p = \Omega(b_p)$ if $\lim_{p \to \infty} a_p / b_p \geq c$, for some constant $c \geq 0$; $a_p = o(b_p)$ if $\lim_{p \to \infty} a_p / b_p = 0$, and ~~$a_p = \Omega(b_p)$~~ $a_p \asymp b_p$ if both $b_p \lesssim a_p$ and $a_p \lesssim b_p$. Throughout the paper, we use the terms "community" and "clusters" exchangeably.

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

II. Model Formulation and Motivations

A. Degree-corrected tensor block model

Suppose that we have an order-K data tensor $\mathcal{Y} \in \mathbb{R}^{p \times \cdots \times p}$. Assume ~~there exist $r \geq 2$~~ that there exist $r \geq 1$ 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 [**? ?**],

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

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

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

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

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

B. Motivating examples

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

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

b) Community detection in hypergraphs: The hypergraph network is a powerful tool to represent the complex entity relations with higher-order interactions [**?**]. 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 that there exist r disjoint communities among p nodes, and the connection probabilities depend on the community assignments and node-specific parameters. Then, the equation (1) models $\mathbb{E}\mathcal{Y}$ with unknown degree heterogeneity $\boldsymbol{\theta}$ 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 [**?**]. 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 that there exist r disjoint communities among p nodes and r_l disjoint communities among the L layers. The multi-layer network community detection is modeled by the general asymmetric ~~asymmetric~~ dTBM model (1)

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

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

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

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

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

where $\{(\boldsymbol{\theta}_k, \boldsymbol{M}_k)\}_{k=1}^3$ represents the degree 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 (1). 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 $(z, \mathcal{S}, \boldsymbol{\theta})$,

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

where $c_i > 0$'s are universal constants. We briefly describe the rationale of the constraints in (2). First, the entrywise positivity constraint on $\boldsymbol{\theta} \in \mathbb{R}^p_+$ 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 (1); see Example 1 below. Second, recall that the quantity $|z^{-1}(a)|$ denotes the number of nodes in the a-th community. The constants c_1, c_2 in the $|z^{-1}(a)|$ ~~bound~~ bounds assume the roughly balanced size across r communities. ~~Third, the constant c_3 requires that all slides in $\mathcal{S}$ have non-degenerate norm. Particularly, the lower bound c_3 excludes the no purely zero slide case to avoid trivial non-identifiability of model ; see Example 2 below. The upper bound c_4 is a technical constraint to avoid the slides with unbounded norm as dimension grows; in practice, the constraint $\max_{a:} \|\mathrm{Mat}(\mathcal{S})_{a:}\| \le c_4$ would likely never be active with a large $c_4 \ge \|\mathcal{Y}\|_F$.~~ Third, the constant c_3 requires that all slides in $\mathcal{S}$ have non-d
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. Our constraints in $\mathcal{P}$ are mild compared with previous literature; see Table II for comparison.

Example 1 (Positivity of degree parameters)**.** Here we provide an example to show the positivity ~~constraints~~ constraint on $\boldsymbol{\theta}$

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

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

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

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

Now, instead of original decomposition, we encode $\mathcal{X}$ 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 $\boldsymbol{\Theta}' = \mathrm{diag}(\boldsymbol{\theta}') = \mathrm{diag}(1, 1, 1, 1)$, the core tensor $\mathcal{S}' \in \mathbb{R}^{2\times2\times2}$ has following mode-3 slices, and the membership matrix $\boldsymbol{M}' \in \{0, 1\}^{4\times2}$ defines the clustering $z' : [4] \to [2]$; i.e.,

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

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

Example 2 (Non-identifiability with purely zero core slice)**.** Consider an order-2 dTBM with core tensor $\boldsymbol{S} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}$ degree matrices $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \mathrm{diag}(1, 1, 1, 1)$, and mean tensor

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

~~and $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \mathrm{diag}(1, 1, 1, 1)$,~~ Replacing $\boldsymbol{\Theta}_1$ by $\boldsymbol{\Theta}'_1 = (3/2, 1/2, 1, 1)$ leads to the same mean tensor $\mathcal{X}$.

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

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

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

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

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

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

The identifiability guarantee for the dTBM is stronger than classical Tucker model. In the Tucker model, the factor matrix $\boldsymbol{M}$ 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 Critical Values for Higher-Order Tensors

A. Assumptions

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

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

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

With quantification (4), we consider the following parameter space,

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

The 1-block dTBM does not belong to the space $\mathcal{P}(\gamma)$ when $\gamma < 0$~~by~~, due to the convention in Assumption 1. Our goal is

to characterize the clustering accuracy with respect to γ under the space $\mathcal{P}(\gamma)$.

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

$$\min_{a\in[r]}\|\boldsymbol{\theta}_{z^{-1}(a)}\| = (1+o(1))\max_{a\in[r]}\|\boldsymbol{\theta}_{z^{-1}(a)}\|. \tag{5}$$

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

Lemma 1 (Angle gaps in $\mathcal{X}$ and $\mathcal{S}$)**.** Consider the dTBM model (1) under the parameter space $\mathcal{P}$ in (2) ~~with $r\geq 2$. Suppose $\boldsymbol{\theta}$ is balanced satisfying and $\min_{i\in[p]}\theta(i)\geq c$ from some constant $c>0$.Then, as $p\to\infty$~~with $r\geq 2$. Suppose $\boldsymbol{\theta}$ is balanced satisfying (5) and $\min_{i\in[p]}\theta(i)\geq c$ from some constant $c>0$. Then, as $p\to\infty$, for all i,j such that $z(i)\neq z(j)$, we have

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

where $\boldsymbol{X}=\mathrm{Mat}(\mathcal{X})$ and $\boldsymbol{S}=\mathrm{Mat}(\mathcal{S})$.

In practice, an estimation algorithm has access to a noisy version of $\mathcal{X}$ but not $\mathcal{S}$. Our goal is to establish the algorithm performance with respect to the signal $\Delta^2_{\min}$ in the core tensor. By Lemma 1, the mapping from the core tensor $\boldsymbol{S}_{z(i):}$ to the mean tensor $\boldsymbol{X}_{z(i):}$ preserves the angle information $\Delta^2_{\min}$ under balanced degree heterogeneity (5). Therefore, the balanced degree assumption helps to exclude the cases in which the degree heterogeneity distorts the algorithm guarantees.

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

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

$$\boldsymbol{S}=\begin{pmatrix}1 & a\\ 1 & -a\end{pmatrix}, \tag{6}$$

and $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}_{z^{-1}(1)}\|^2 = p^m\|\boldsymbol{\theta}_{z^{-1}(2)}\|^2$, where ~~$m\in[-1,1]$~~$m\in[-1,1]$ is a scalar parameter controlling the skewness of degrees. Let $\Delta^2_{\boldsymbol{X}}$ denote the minimal angle gap of the mean tensor, defined by

$$\Delta^2_{\boldsymbol{X}} := \min_{i,j\in[p],z(i)\neq z(j)}\left\|\frac{\boldsymbol{X}_{i:}}{\|\boldsymbol{X}_{i:}\|}-\frac{\boldsymbol{X}_{j:}}{\|\boldsymbol{X}_{j:}\|}\right\|, \tag{7}$$

where ~~$\boldsymbol{X}=\mathrm{Mat}_1(\mathcal{X})$~~ $\boldsymbol{X}=\mathrm{Mat}(\mathcal{X})$. Take $a=p^{-1/4}$ in the model setup (6). We have

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

~~Based on our theory in later Sections~~Based on the Theorem 2 in Section III, the dTBM is impossible to solve when $\Delta^2_{\boldsymbol{X}}\lesssim p^{-1}$ even though $\Delta^2_{\min}\asymp p^{-1/2}$; that is, the dTBM estimation depends on the relative magnitude of m vs. $1/2$. In such a setting, the proposed signal notion $\Delta^2_{\min}$ alone fails to fully characterize dTBM.

Remark 2 (Flexibility in balanced degree assumption)**.** One important note is that our balance assumption (5) does not preclude the mild degree heterogeneity. In fact, within each of the clusters, we allow the highest degree at the order $\mathcal{O}(p)$, whereas the lowest degree at the order $\Omega(1)$. This range is more relaxed than previous work [**?**] that restricts the highest degree in the sub-linear regime $o(p)$ and the lowest degree at the order $\Omega(1)$.

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

Last, let $\hat{z}$ and z be the estimated and true clustering functions in the family (2). Define the misclustering error by

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

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~~ infimum over all permutations accounts for the ambiguity in cluster label permutation.

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

B. Statistical critical value

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

$$(\hat{z}_{\mathrm{MLE}},\hat{\mathcal{S}}_{\mathrm{MLE}},\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})=\arg\min_{(z,\boldsymbol{\theta},\mathcal{S})\in\mathcal{P}}\|\mathcal{Y}-\mathcal{X}(z,\mathcal{S},\boldsymbol{\theta})\|_F^2. \tag{8}$$

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

- **Impossibility.** ~~Assume $p\to\infty$ and $2\leq r\lesssim p^{1/3}$.~~Assume $p\to\infty$ and $2\leq r\lesssim p^{1/3}$. Let $\mathcal{P}_{\mathcal{S}}(\gamma):=\{\mathcal{S}:c_3\leq\|\mathrm{Mat}(\mathcal{S})_{a:}\|\leq c_4, a\in[r]\}\cap\{\mathcal{S}:$

$\Delta^2_{\min} = p^\gamma\}$ denote the space for valid $\mathcal{S}$ satisfying SNR condition (4), and $\mathcal{P}_{z,\boldsymbol{\theta}} := \{\boldsymbol{\theta} \in \mathbb{R}^p_+,\ \frac{c_1 p}{r} \le |z^{-1}(a)| \le \frac{c_2 p}{r}, \left\|\boldsymbol{\theta}_{z^{-1}(a)}\right\|_1 = |z^{-1}(a)|, a \in [r]\}$ denote the space for valid $(z, \boldsymbol{\theta})$, where c_1, c_2, c_3, c_4 are the constants in parameter space (2). If the signal exponent satisfies $\gamma < -(K-1)$, then, for any true core tensor $\mathcal{S} \in \mathcal{P}_{\mathcal{S}}(\gamma)$, no estimator $\hat{z}_{\text{stat}}$ achieves exact recovery in expectation; that is, when $\gamma < -(K-1)$, we have

$$\liminf_{p\to\infty} \inf_{\mathcal{S}\in\mathcal{P}_{\mathcal{S}}(\gamma)} \inf_{\hat{z}_{\text{stat}}} \sup_{(z,\boldsymbol{\theta})\in\mathcal{P}_{z,\boldsymbol{\theta}}} \mathbb{E}\left[p\ell(\hat{z}_{\text{stat}}, z)\right] \ge 1. \tag{9}$$

Further, we define the parameter space $\mathcal{P}'(\gamma') := \mathcal{P} \cap \{\Delta^2_{\boldsymbol{X}} = p^{\gamma'}\}$, where $\Delta^2_{\boldsymbol{X}}$ is the mean tensor minimal gap in (7). When $\gamma' < -(K-1)$, we have

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

• **MLE achievability.** Suppose that the signal exponent satisfies $\gamma > -(K-1) + c_0$ for an arbitrary constant ~~$c_0 > 0$~~$c_0 \ge 0$. Furthermore, assume ~~$\boldsymbol{\theta}$~~that $\boldsymbol{\theta}$ is balanced and $\min_{i\in[p]} \theta(i) \ge c$ from some constant $c > 0$. Then, ~~when $p \to \infty$, fixed $r \ge 1$,~~ when $p \to \infty$, for fixed $r \ge 1$, the MLE in (8) achieves exact recovery in high probability; that is,

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

with probability going to 1.

The proofs for the two parts in Theorem 2 are in the Appendix ~~H~~B, Section **??** and Section **??**, respectively. The first part of Theorem 2 demonstrates impossibility of exact recovery whenever the core tensor $\mathcal{S}$ satisfies SNR condition (4) with exponent $\gamma < -(K-1)$. The proof is information-theoretical, and therefore the results apply to all statistical estimators, including but not limited to MLE and trace maximization [**?**]. The minimax bound (9) indicates the worst case impossibility for a particular core tensor $\mathcal{S}$ with signal exponent $\gamma < -(K-1)$; i.e., under the assumptions of Theorem 2, when $\gamma < -(K-1)$, we have

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

Such worst case impossibility is ~~frequently~~ studied in related works [**? ?**] while our lower bound (9) provides a stronger impossibility statement for arbitrary core ~~tensor with weak signal.~~

tensors with weak signals. The second part of Theorem 2 shows the exact recovery of MLE when ~~the $\gamma > -(K-1) + c$~~ $\gamma > -(K-1) + c_0$ for an arbitrary constant ~~$c > 0$~~$c_0 \ge 0$. Combining the impossibility and achievability results, we conclude that the boundary $\gamma_{\text{stat}} := -(K-1)$ is the critical value for statistical performance of dTBM with respect to our SNR.

C. Computational critical value

~~In this section, we derive the computational critical value of dTBMs.~~ The computational critical value means the minimal SNR required for ~~exactly~~ exact recovery with *polynomial-time* computational cost. An important ingredient to establish the computational limits is the *hypergraphic planted clique (HPC) conjecture* [**? ?**]. 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 with $K \ge 2$. Suppose the sequence $\{\kappa_p\}$ such that $\limsup_{p\to\infty} \log \kappa_p / \log \sqrt{p} \le (1-\tau)$ for any $\tau > 0$. Then, for every sequence of polynomial-time test $\{\varphi_p\} : H \mapsto \{0, 1\}$ we have

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

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

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

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

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

• **Polynomial-time algorithm achievability.** Suppose ~~the parameter space~~ that we have fixed $r \ge 1, K \ge 2$, and the signal exponent satisfies $\gamma > -K/2 + c_0$ for an arbitrary constant ~~$c_0 > 0$~~$c_0 \ge 0$. Furthermore, assume ~~fixed $r \ge 1, K \ge 2$,the degree~~ that the degree $\boldsymbol{\theta}$ is balanced, lower bounded in that $\min_{i\in[p]} \theta_i \ge c$ for some constant $c > 0$, and satisfies the ~~local~~ locally linear stability in Definition 2 ~~in the neighborhood $\mathcal{N}(z, \varepsilon)$ for all $\varepsilon \le E_0$~~

~~and some $E_0 \geq \check{C}\log^{-1} p$ with some positive constant $\check{C}$~~ in the neighborhood $\mathcal{N}(z,\varepsilon)$ for all $\varepsilon \leq E_0$ and some $E_0 \geq \log^{-1}$. Then, ~~as $p \to \infty$~~ as $p \to \infty$, there exists a polynomial-time algorithm $\hat{z}_{\text{ploy}}$ that achieves exact recovery in high probability; that is,

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

with probability going to 1.

The proofs for the two parts in Theorem 3 are in the Appendix ~~H~~B, Section **??** and Section **??**, respectively. The first part of Theorem 3 indicates the impossibility of exact recovery by polynomial-time algorithms when $\gamma < -K/2$, and the second part shows the existence of such algorithm when ~~$\gamma > -K/2 + c_0$ for an arbitrary constant c_0 and~~ $\gamma \geq -K/2 + c_0$ for an arbitrary constant $c_0 > 0$ under extra technical assumptions. In Section IV, we will present an efficient polynomial-time algorithm in this setting. Therefore, we conclude that $\gamma_{\text{comp}} := -K/2$ is the critical value for computational performance of dTBM with respect to our SNR.

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

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

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

IV. Polynomial-time algorithm under mild SNR

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

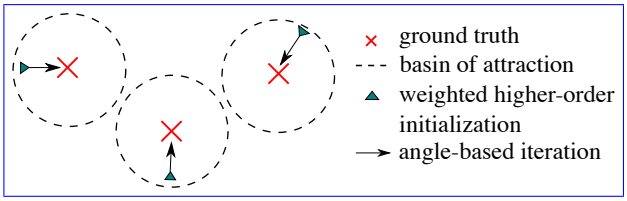

Fig. 3: Illustration of our global-to-local algorithm.

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

$$L(\hat{z}, z) = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}\mathbb{1}\{\hat{z}(i) = b\} \cdot \left\|\left[\boldsymbol{S}_{z(i):}\right]^s - \left[\boldsymbol{S}_{b:}\right]^s\right\|^2, \quad (10)$$

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 $\boldsymbol{a}$. The following lemma indicates the close relationship between the loss $L(\hat{z}, z)$ and error $\ell(\hat{z}, z)$. The loss $L(\hat{z}, z)$ serves as an important intermediate quantity to control the misclustering error.

Lemma 2 (Relationship between misclustering error and loss)**.** ~~Consider the dTBM under the parameter space $\mathcal{P}$.~~ Consider the dTBM under the parameter space $\mathcal{P}$. Suppose $\min_{i\in[p]}\theta(i) > c$ for some constant $c > 0$. We have $\ell(\hat{z}, z)\Delta^2_{\min} \leq L(\hat{z}, z)$.

A. Weighted higher-order initialization

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

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

This implies that, all node i belonging to the a-th community (i.e., $z(i) = a$) share the same normalized mean vector $\theta(i)^{-1}\boldsymbol{X}_{i:}$, 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 the denoising step and the clustering step. The denoising step (lines 1-2 in Sub-algorithm 1) estimates $\mathcal{X}$ from $\mathcal{Y}$ by a double projection spectral method. The first projection performs HOSVD [**?**] via $\boldsymbol{U}_{\text{pre},k} = \text{SVD}_r\left(\text{Mat}_k(\mathcal{Y})\right), k \in [3]$, where $\text{SVD}_r(\cdot)$ returns the top-r left singular vectors. The second projection performs

HOSVD on the projected $\mathcal{Y}$ onto the multilinear Kronecker space $\boldsymbol{U}_{\text{pre},k} \otimes \boldsymbol{U}_{\text{pre},k}$; i.e.,

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

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

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

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

The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted k-means clustering. We write $\hat{\boldsymbol{X}} = \text{Mat}_1(\hat{\mathcal{X}})$, 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 [**?**], we apply the clustering on the unfolded tensor $\hat{\boldsymbol{X}}$ rather than on the factors ~~$\hat{\boldsymbol{U}}_1$~~$\hat{\boldsymbol{U}}_k$. This strategy relaxes the singular-value gap condition [**? ?**]. We assign degenerate rows with purely zero entries to an arbitrarily random cluster; these nodes are negligible in high-dimensions because of the lower bound on $\|\text{Mat}(\mathcal{S})_{a:}\|$ in (2). The final result gives the initial cluster assignment $z^{(0)}$. Full procedures for clustering are provided in Sub-algorithm 1.

We now establish the misclustering error rate of initialization.

Theorem 4 (Error for weighted higher-order initialization)**.** Consider the general sub-Gaussian dTBM with ~~fixed $r \geq 1$, $K \geq 2$,~~ fixed $r \geq 1$, $K \geq 2$, i.i.d. noise under the parameter space $\mathcal{P}$, and Assumption 1. Assume $\min_{i\in[p]}\theta(i) \geq c$ for some constant $c > 0$~~and let~~. Let $\Delta_{\boldsymbol{X}}$ denote the minimal gap in mean tensor defined in (7)~~. Let~~, and let $z_k^{(0)}$ denote the output of Sub-algorithm 1. With probability going to 1, ~~as $p \to \infty$~~as $p \to \infty$, we have

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

Further, assume that $\boldsymbol{\theta}$ is balanced as (5). We have

$$\ell(z_k^{(0)}, z) \lesssim \frac{r^K p^{-K/2}}{\text{SNR}} \text{ and } L(z_k^{(0)}, z) \lesssim \sigma^2 r^K p^{-K/2}, \quad (11)$$

~~with probability going to 1 as $p \to \infty$.~~with probability going to 1 as $p \to \infty$.

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

Remark 6 (Failure of conventional tensor HOSVD)**.** If we use conventional HOSVD for tensor denoising; that is, we use $\boldsymbol{U}_{\text{pre},k}$ in place of ~~$\hat{\boldsymbol{U}}_k$~~ $\hat{\boldsymbol{U}}_k$ in line 2, then the misclustering rate becomes $\mathcal{O}(p^{-1})$ for all $K \geq 2$. This rate is substantially worse than our current rate (11).

Remark 7 (Singular-value gap-free clustering)**.** Note that our clustering directly applies to the estimated mean tensor $\hat{\mathcal{X}}$ rather than the leading tensor factors $\hat{\boldsymbol{U}}_k$. Applying clustering to the tensor factors suffers from the non-identifiability issue due to the infinitely many orthogonal rotations when the number of blocks $r \geq 3$ in the absence of singular-value gaps. Such ambiguity causes the trouble for effective clustering [**?**]. 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.,

$$z(i) = \arg\max_{a\in[r]} \cos(\boldsymbol{x}_i, \ \boldsymbol{s}_a), \quad \text{for all } i \in [p]. \quad (12)$$

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

Our Sub-algorithm 2 shares the same spirit as in the 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 the t-th iteration, where ~~$t = 1, \ldots$~~ $t = 1, 2, \ldots$ For core tensor, we consider the following update strategy

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

Intuitively, $\mathcal{S}^{(t)}$ becomes closer to the true core $\mathcal{S}$ as ~~$z^{(t)}$~~ $z_k^{(t)}$ 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 $z^{(t)}$~~the reduced tensor $\mathcal{Y}_1^{\text{d}} \in \mathbb{R}^{p\times r\times r}$ on the first mode with given $z_k^{(t)}$, where

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

Similarly, we also obtain $\mathcal{Y}_2^{\text{d}}, \mathcal{Y}_3^{\text{d}}$. We use $\boldsymbol{Y}_k^d$ and $\boldsymbol{S}_k^{(t)}$ to denote the $\text{Mat}_k(\mathcal{Y}^{\text{d}})$ and $\text{Mat}_k(\mathcal{S}^{(t)})$. The rows $\boldsymbol{Y}_{k,i:}^d$ and $\boldsymbol{S}_{k,a:}^{(t)}$ correspond to the $\boldsymbol{x}_i$ and $\boldsymbol{s}_a$ in the one-dimensional

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

Sub-algorithm 1: Weighted higher-order initialization

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

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

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

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

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

for $k \in [K]$ **do**
  Let $\hat{\boldsymbol{X}} = \text{Mat}_k(\hat{\mathcal{X}})$ and $S_0 = \{i \in [p] : \|\hat{\boldsymbol{X}}_{i:}\| = 0\}$. Set $\hat{z}(i)$ randomly in $[r]$ for $i \in S_0$.
  For all $i \in S_0^c$, compute normalized rows $\hat{\boldsymbol{X}}_{i:}^s := \|\hat{\boldsymbol{X}}_{i:}\|^{-1}\hat{\boldsymbol{X}}_{i:}$.
  Solve the clustering $\hat{z}_k\colon [p] \to [r]$ and centroids ~~$(\hat{\boldsymbol{x}}_j)_{j\in[r_k]}$~~ $\{\hat{\boldsymbol{x}}_j\}_{j\in[r]}$ using weighted k-means, such that

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

end for

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

Sub-algorithm 2: Angle-based iteration

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

for $t = 0$ to $T-1$ **do**
  Update the block tensor $\mathcal{S}^{(t)}$ via $\mathcal{S}^{(t)}(a_1, ..., a_K) = \text{Ave}\{\mathcal{Y}(i_1, \ldots, i_K) : z_k^{(t)}(i_k) = a_k, k \in [K]\}$.
  for $k \in [K]$ **do**
    Calculate ~~reduced tensor $\mathcal{Y}_k^{\text{d}} \in \mathbb{R}^{r\times r\times p\times r\times\cdots\times r}$~~ the reduced tensor $\mathcal{Y}_k^{\text{d}} \in \mathbb{R}^{r\times\cdots\times r\times p\times r\times\cdots\times r}$ via

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

    Let $\boldsymbol{Y}_k^{\text{d}} = \text{Mat}_k(\mathcal{Y}^{\text{d}})$ and $J_0 = \{i \in [p] : \|\boldsymbol{Y}_{i:}^{\text{d}}\| = 0\}$. Set $z_k^{(t+1)}(i)$ randomly in $[r]$ for $i \in J_0$.
    Let ~~$\boldsymbol{S}^{(t)} = \text{Mat}(\mathcal{S}^{(t)})$~~ $\boldsymbol{S}_k^{(t)} = \text{Mat}_k(\mathcal{S}^{(t)})$. For all $i \in J_0^c$, update the cluster assignment by

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

  end for
end for

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

clustering (12). Then, we obtain the updated assignment by

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

provided that $\boldsymbol{S}_{k,a:}^{(t)}$ is a non-zero vector. Otherwise, if $\boldsymbol{S}_{k,a:}^{(t)}$ is a zero vector, then we make the convention to assign $z_k^{(t+1)}(i)$ randomly in ~~$[p]$~~$[r]$. 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 $\mathcal{N}(z, \epsilon) = \{\bar{z}\colon \ell(\bar{z}, z) \le \epsilon\}$. Let $\bar{z}\colon [p] \to [r]$ be a clustering function. We define two vectors associated with $\bar{z}$,

$$\boldsymbol{p}(\bar{z}) = (|\bar{z}^{-1}(1)|, \ldots, |\bar{z}^{-1}(r)|)^T,$$

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

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

$$\sin(\boldsymbol{p}(\bar{z}),\ \boldsymbol{p}_{\boldsymbol{\theta}}(\bar{z})) \lesssim \varepsilon\Delta_{\min}, \quad \text{for all } \bar{z} \in \mathcal{N}(z, \varepsilon). \tag{13}$$

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 (2). The condition (13) controls the impact of node degree to the $\boldsymbol{p}_\theta(\cdot)$ with respect to the ~~misclassification~~ misclustering rate ε and angle gap. Intuitively, the condition (13) controls the skewness of degree so that the angle between raw cluster size and degree-weighted cluster size is well controlled. The stability assumption is proposed for technical convenience, and we relax this condition in numerical studies; see Section VI.

Theorem 5 (Error for angle-based iteration)**.** Consider the general sub-Gaussian dTBM with ~~fixed $r \geq 1$, $K \geq 2$,~~ fixed $r \geq 1$, $K \geq 2$, independent noise under the parameter space $\mathcal{P}$, and Assumption 1. ~~Assume the local linear stability of degree holds in the neighborhood $\mathcal{N}(z, \varepsilon)$ for all $\varepsilon \leq E_0$ and some $E_0 \geq \check{C} \log^{-1} p$ with some positive constant $\check{C}$.~~Assume that the locally linear stability of degree holds in the neighborhood $\mathcal{N}(z, \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \leq E_0^{K/2}$ for all $E_0 \geq \log^{-1} p$. Let $\{z_k^{(0)}\}_{k=1}^K$ be the initialization for ~~Sub-Algorithm~~Sub-algorithm 2 and $z_k^{(t)}$ be the t-th iteration output on the k-th mode. Suppose $\min_{i\in[p]} \theta(i) \geq c$ for some ~~constant~~constant $c > 0$, the SNR $\geq \tilde{C} p^{-(K-1)} \log p$ for some sufficiently large constant $\tilde{C}$, and the initialization satisfies

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

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

$$\ell(z, \hat{z}_k^{(t+1)}) \lesssim \underbrace{\text{SNR}^{-1} \exp\left(-\frac{p^{K-1}\text{SNR}}{r^{K-1}}\right)}_{\text{statistical error}} + \underbrace{\rho^t \ell(z, z_k^{(0)})}_{\text{computational error}}. \quad (14)$$

From the conclusion (14), 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 $t \to \infty$.

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

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

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

Therefore, our combined algorithm is *computationally efficient* as long as 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~~ critical value in Theorem 3. Therefore, our algorithm is optimal regarding the signal requirement and lies in the sharpest *computationally efficient* regime in Figure 2.

C. Extension to Bernoulli observations

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

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

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

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim p^{K/2} \quad (15)$$

where $\mathcal{X} = \mathbb{E}\mathcal{Y}$ and $\hat{\mathcal{X}}$ is defined in Step 2 of Sub-algorithm 1. Unfortunately, the inequality (15) 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* [**?**], $\text{Mat}_{sq}(\cdot)$, of Bernoulli tensor $\mathcal{Y} \in \{0,1\}^{p_1 \times \cdots \times p_K}$. Specifically, the square matricization $\text{Mat}_{sq}(\mathcal{Y}) \in \{0,1\}^{p^{\lfloor K/2 \rfloor} \times p^{\lceil K/2 \rceil}}$ has entries $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2) = \mathcal{Y}(i_1, \ldots, i_K)$, where

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

The matrix $\text{Mat}_{sq}(\mathcal{Y})$ is asymmetric. We ~~are able to~~ interpret $\text{Mat}_{sq}(\mathcal{Y})$ as the adjacency matrix for a bipartite network with connections between two groups of ~~node~~nodes. The two groups of nodes in the bipartite network have $p_1 \cdots p_{\lfloor K/2 \rfloor}$ and $p_{\lceil K/2 \rceil} \cdots p_K$ nodes, respectively. The entry $[\text{Mat}_{sq}(\mathcal{Y})](j_1, j_2)$ refers to the presence of connection between the nodes indexed by combinations $(i_1, \ldots, i_{\lfloor K/2 \rfloor})$ and $(i_{\lceil K/2 \rceil}, \ldots, i_K)$. We summarize the procedure in Sub-algorithm 3.

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

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

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

$$\hat{\mathcal{X}}' = \mathop{\arg\min}_{\text{rank}(\text{Mat}_{sq}(\mathcal{X})) \leq r^{\lceil K/2 \rceil}} \|\text{Mat}_{sq}(\mathcal{X}) - \text{Mat}_{sq}(\mathcal{Y})\|_F^2. \quad (16)$$

2: Implement lines 3-5 of Sub-algorithm 1 with $\hat{\mathcal{X}}$ replaced by $\hat{\mathcal{X}}'$ in (16).

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

Proposition 1 (Error for Bernoulli initialization)**.** Consider the Bernoulli dTBM in the parameter space $\mathcal{P}$ ~~with fixed $r \geq 1, K \geq 2$and~~ with fixed $r \geq 1, K \geq 2$. Assume that Assumption 1 holds. ~~Assume that~~ , $\boldsymbol{\theta}$ is balanced, and $\min_{i\in[p]} \theta(i) \geq c$ for some constant $c > 0$. Let ~~$z^{(0)}$~~ $z_k^{(0)}$ denote the output of Sub-algorithm 3. With probability going to 1 ~~as $p \to \infty$~~as $p \to \infty$, we have

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

Remark 8 (Comparison with Gaussian model)**.** The Bernoulli bound $\mathcal{O}(p^{-\lfloor K/2 \rfloor})$ in Proposition 1 is relatively looser than

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

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

Hence, for fair comparison, we compare the best performance of our algorithm and [**?**] under the strongest signal setting of each model. Since both methods contain an iteration procedure, we set the iteration number to infinity to avoid the computational error. Considering the largest angle-based SNR $\asymp 1$ in Theorem 5, our Bernoulli clustering achieves exponential error rate of order $\exp(-p^{(K-1)})$; considering the largest singular gap $\Delta_{\text{singular}} \asymp 1$ in Theorem 1 of [**?**], the spectral clustering has a polynomial error rate of order p^{-2}. Our algorithm still shows a better theoretical accuracy than the competitive work for Bernoulli observations.

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

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

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

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

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

D. Practical issues

Computational complexity. Our two-stage algorithm has a computational cost polynomial in tensor dimension p. Specifically, the complexity of Sub-algorithm 1 is $\mathcal{O}(Kp^{K+1} + Krp^K)$, where the first term is contributed by the double projection and the calculation of $\hat{\mathcal{X}}$, and the second term comes from normalization and the k-means. The cost of each update in Sub-algorithm 2 is $\mathcal{O}(p^K + pr^K)$, where p^K comes from the calculation of $\mathcal{S}^{(t)}$ and ~~$\mathcal{Y}^{\text{d}}$~~$\mathcal{Y}_k^{\text{d}}$, and pr^K comes from the normalization of ~~$\mathcal{Y}^{\text{d}}$~~$\mathcal{Y}_k^{\text{d}}$, the calculation of $\mathcal{S}^{(t)}$, and the cluster assignment update in Step ~~10.~~13.

Hyper-parameter selection. In our theoretical analysis, we have assumed the true cluster number r is given to our algorithm. In practice, the cluster number r is often unknown, and we now propose a method to choose r from data. We impose the Bayesian information criterion (BIC) and choose the cluster number that minimizes BIC; i.e., under the symmetric Gaussian dTBM (1),

$$\hat{r} = \underset{r \in \mathbb{Z}_+}{\arg\min} \left(p^K \log(\|\hat{\mathcal{X}} - \mathcal{Y}\|_F^2) + p_e(r) K \log p \right), \quad (17)$$

with $\hat{\mathcal{X}} = \hat{\mathcal{S}}(r) \times_1 \hat{\boldsymbol{\Theta}}(r)\hat{\boldsymbol{M}}(r) \times_2 \cdots \times_K \hat{\boldsymbol{\Theta}}(r)\hat{\boldsymbol{M}}(r)$, where the triplet ~~$(\hat{z}(r), \hat{\boldsymbol{S}}(r), \hat{\boldsymbol{\theta}}(r))$~~ $(\hat{z}(r), \hat{\mathcal{S}}(r), \hat{\boldsymbol{\theta}}(r))$ 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 (17) 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.,

$$\hat{\boldsymbol{\theta}}(r) = (\hat{\theta}(1, r), \ldots, \hat{\theta}(p, r))^T,$$

with $\hat{\theta}(i, r) = \|\hat{\boldsymbol{Y}}^{\text{d}}(r)_{i:}\| / \sum_{j:\hat{z}(j,r)=\hat{z}(i,r)} \|\hat{\boldsymbol{Y}}^{\text{d}}(r)_{j:}\|$, $\hat{\boldsymbol{Y}}^{\text{d}}(r) = \text{Mat}(\hat{\mathcal{Y}}^{\text{d}}(r))$, $\hat{\mathcal{Y}}^{\text{d}}(r)(i, a_2, \ldots, a_K) = \text{Ave}\{\mathcal{Y}(i, i_2, \ldots, i_K) : \hat{z}(i_k, r) = a_k, k \neq 1\}$, 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 VI-A.

V. Comparison with Non-degree Tensor Block Model

We discuss the connections and ~~difference~~ differences between dTBM and TBM [**?**] from three aspects: signal notions, theoretical results, and algorithms. Without loss of generality, let $\sigma^2 = 1$.

• *Signal notion.* The signal levels in both TBM [**?**] and our dTBM are functions of the core tensor $\mathcal{S}$. We emphasize that

the signal notions are different between the two models. In particular, the Euclidean-based signal notion in TBM [**?**] fails to accurately describe the phase transition in our dTBM due to the possible heterogeneity in degree $\boldsymbol{\theta}$. To compare, we denote our angle-based signal notion in (4) and the Euclidean-based SNR in [**?**] as Δ^2_{ang} and Δ^2_{Euc}, respectively:

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

By Lemma **??** in the Appendix ~~H~~B, we have

$$\Delta^2_{\text{ang}} \max_{a\in[r]} \|\boldsymbol{S}_{a:}\|^2 \leq \Delta^2_{\text{Euc}}.$$

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

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

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

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

The core matrix $\boldsymbol{S}$ lies in the parameter spaces of TBM and our dTBM. Here, the constraint $\gamma \leq -1$ is added to ensure the bounded condition of $\boldsymbol{S}$ in our parameter space in (2). The angle-based and Euclidean-based signal levels of $\boldsymbol{S}$ are

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

We conclude that TBM with $\boldsymbol{S}$ achieves exact recovery with a polynomial-time algorithm; see [**?** , Theorem 4]. By contrast, the dTBM with the same $\boldsymbol{S}$ and input $r = 2$ violets the identifiability condition, and thus fails to be solved by all estimators; see our Theorem 1.

• *Theoretical results.* In both works, we study the phase transition of TBM and dTBM with respect to the Euclidean and angle-based SNRs. We briefly summarize the results in [**?**] and compare with ours.

Statistical critical value:

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

Computational critical value:

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

The above comparison reveals four major differences.

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

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

Third, from the perspective of proofs, we develop new dTBM-specific techniques to handle the extra degree heterogeneity. In our Theorem 2, we construct a special non-trivial degree heterogeneity to establish the lower bound for arbitrary core tensor with small angle gap, while, TBM [**?**] considers the constructions without degree parameter. In our Theorem 3, we construct a rank-2 tensor to relate HPC conjecture to Δ^2_{ang}, while TBM [**?**] constructs a rank-1 tensor to relate HPC conjecture to Δ^2_{Euc}. The asymptotic non-equivalence between Δ^2_{ang} and Δ^2_{Euc} renders our proof technically more involved.

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

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

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

• *Algorithms.* Both [**?**] and our work propose the two-step algorithm, which combines warm initialization and iterative refinement to achieve exact recovery. This local-to-global

strategy is not new in clustering literature [**?** **?**]. The highlight of our algorithm is the angle-based update in lines 10-14, Sub-algorithm 2, which is specifically designed for dTBM to avoid the estimation of $\boldsymbol{\theta}$. This angle-based update brings new proof challenges. We develop polar-coordinate based techniques to establish the error rate for the proposed algorithm. ~~"~~

VI. Numerical studies

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

We generate order-3 tensors with *assortative* [**?**] 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 (4). Hence, we easily adjust SNR via varying α. ~~Note that the~~ 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 (2). Also, we set $\sigma^2 = 1$ for Gaussian data without further specification.

A. Verification of theoretical results

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

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

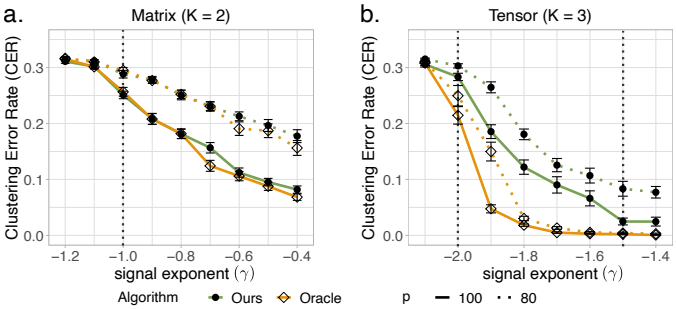

Fig. 4: SNR phase transitions for clustering in dTBM with $p = \{80, 100\}, r = 5$ under (a) matrix case with $\gamma \in [-1.2, -0.4]$ and (b) tensor case with $\gamma \in [-2.1, -1.4]$.

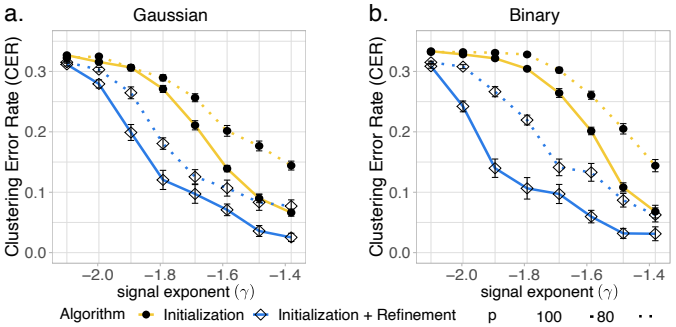

Fig. 5: CER versus signal exponent (γ) for initialization only and for combined algorithm. We set $p = \{80, 100\}, r = 5, \gamma \in [-2.1, -1.4]$ under (a) Gaussian models and (b) Bernoulli models.

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

B. Comparison with other methods

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

- **HOSVD**: HOSVD on data tensor and k-means on the rows of the factor matrix;
- **HOSVD+**: HOSVD on data tensor and k-means on the ℓ_2-normalized rows of the factor matrix;
- **HLloyd** [**?**]: High-order clustering algorithm developed for non-degree tensor block models;

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

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

• **SCORE** [**?**]: Tensor-SCORE for clustering developed for sparse binary tensors.

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

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

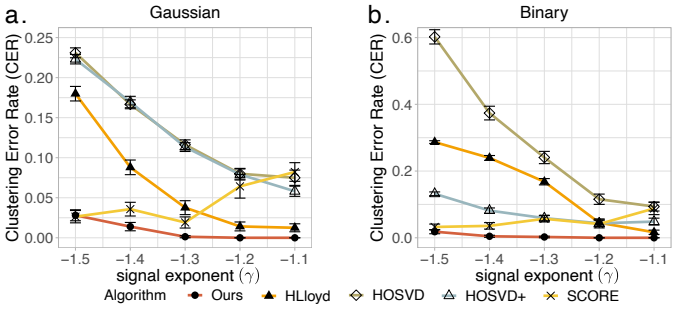

Fig. 6: CER versus signal exponent (denoted γ) for different methods. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under (a) Gaussian and (b) Bernoulli models.

The only exception in Figure 6 is the slightly better performance of **HLloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLloyd** disappears with higher degree heterogeneity. We perform extra simulations to verify the impact of degree effects. We use the same setting as in the first experiment in the Section VI-B, except that we now generate the degree heterogeneity $\boldsymbol{\theta}$ from Pareto distribution prior to normalization. The density function of Pareto distribution is $f(x|a, b) = ab^a x^{-(a+1)} \mathbb{1}\{x \geq b\}$, where 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 7 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.

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

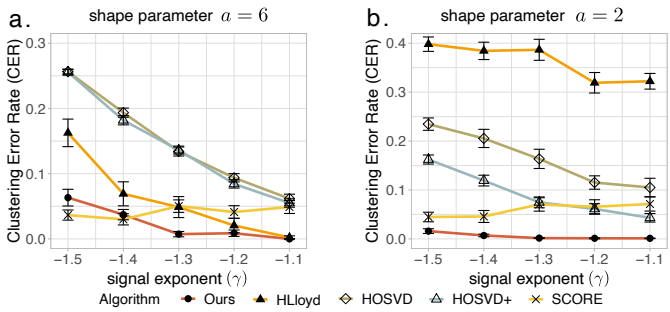

Fig. 7: CER comparison versus signal exponent (denoted γ) under (a) low (shape parameter $a = 6$) (b) high (shape parameter $a = 2$) degree heterogeneity. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under Gaussian model.

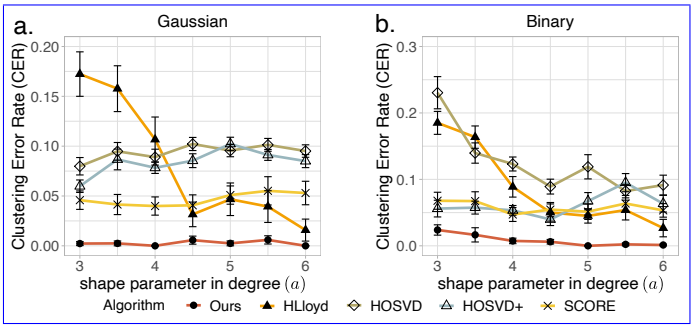

Fig. 8: CER versus shape parameter in degree (denoted $a \in [3, 6]$) for different methods. We set $p = 100, r = 5, \gamma = -1.2$ under (a) Gaussian and (b) Bernoulli models.

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 8 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.

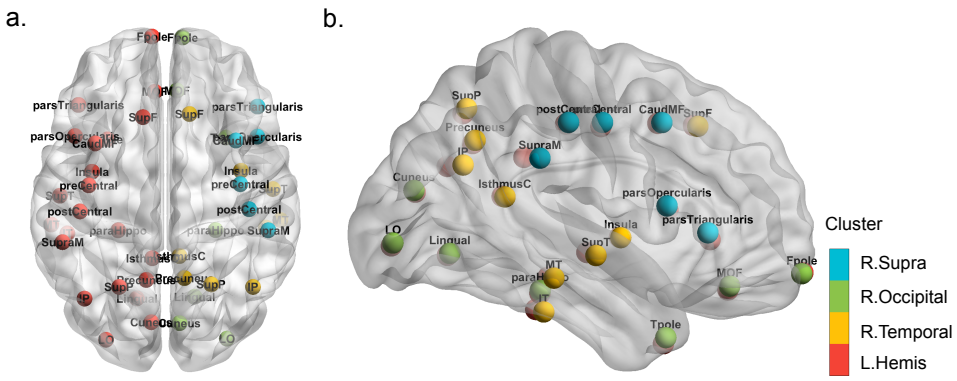

Fig. 9: Illustration of brain node clustering results for HCP data with (a) top and (b) side views.

This experiment, again, highlights the benefit of addressing degree heterogeneity in higher-order clustering.

VII. Real data applications

A. Human brain connectome data analysis

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

We apply our general asymmetric algorithm to the HCP data with the numbers of clusters on three modes $r_1 = r_2 = 4$ and $r_3 = 3$. The selection of r_1 and r_2 follows the human brain anatomy and the symmetry in the brain network, and the r_3 is specified following previous analysis [**?**]. 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 [**?**]. 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 [**?**]. Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the data for the first half of 2006-2007 year. The dataset records the co-sponsorship of 116 legislators from top 5 parties and 802 bill proposals. We reconstruct legislation network as an order-3 binary tensor $\mathcal{Y} \in \{0,1\}^{116\times 116\times 116}$, where $\mathcal{Y}_{ijk} = 1$ if the legislators (i,j,k) have sponsored the same bill, and $\mathcal{Y}_{ijk} = 0$ otherwise. The true party affiliations of legislators are provided and serve as the ground truth. We apply various higher-order clustering methods to $\mathcal{Y}$ with $r = 5$. Table IV shows that our **dTBM** achieves the best performance compared to others. The second best method is the two-stage algorithm **HLloyd**, followed by the spectral methods **SCORE** and **HOSVD+**. This result is consistent with our simulations under strong signal and moderate degree heterogeneity. The comparison suggests that our method **dTBM** is more appealing in real-world applications.

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

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

VIII. Proof Sketches

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

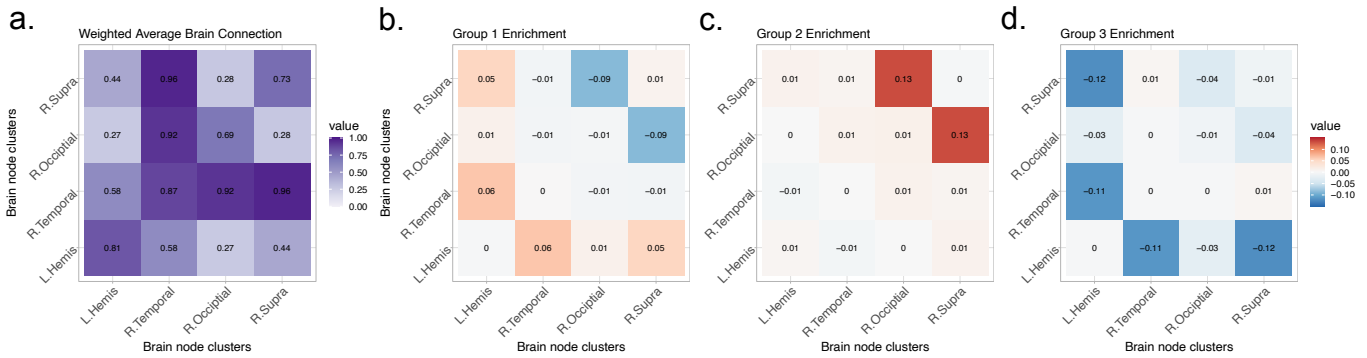

Fig. 10: Mode 3 slices of estimated core tensor $\hat{\mathcal{S}}$. (a) Average estimated slice weighted by the group size; (b)-(d) Group-specified enrichment, i.e., the difference between each slice of $\hat{\mathcal{S}}$ and the averaged slice.

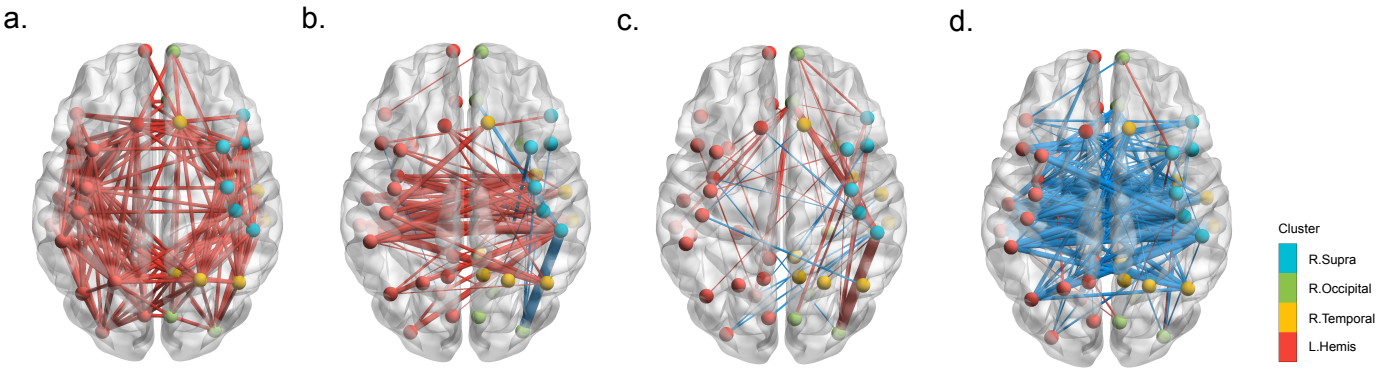

Fig. 11: Observed brain connections in the population and each group of individuals. (a) Average brain network; (b)-(d) Group-specified brain network enrichments in Groups 1-3. Red edges represent the positive enrichment and blue edges represent the negative enrichment.

A. Proof sketch of Theorem 2 (Impossibility) and Theorem 3 (Impossibility)

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

$$\inf_{\hat{z}_{\text{stats}}} \sup_{(z,\boldsymbol{\theta})\in\mathcal{P}_{z,\boldsymbol{\theta}}} \mathbb{E}[p\ell(\hat{z}_{\text{stat}}, z)] \geq \inf_{\hat{z}_{\text{stats}}} \mathbb{E}[p\ell(\hat{z}_{\text{stat}}, z^*_{\text{stats}})|(z^*_{\text{stats}}, \mathcal{S}^*, \boldsymbol{\theta}^*_{\text{stats}})] \geq 1; \quad (18)$$

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

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

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

A notable detail in the proof of statistical impossibility is the arbitrariness of $\mathcal{S}^*$. The first infimum over $\mathcal{P}_{\mathcal{S}}(\gamma)$ in the minimax rate (9) requires that the lower bound (18) holds for any $\mathcal{S}^* \in \mathcal{P}_{\mathcal{S}}(\gamma)$. The arbitrary choice of $\mathcal{S}^*$ brings extra difficulties in the parameter construction, and consequently a non-trivial $\boldsymbol{\theta}^*_{\text{stats}} \neq \mathbf{1}$ is chosen to address the arbitrariness. Previous TBM construction in the proof of [**?** , Theorem 6] with $\boldsymbol{\theta}^*_{\text{stats}} = \mathbf{1}$ is no longer applicable in our case. Meanwhile, our construction $(z^*_{\text{comp}}, \mathcal{S}^*_{\text{comp}}, \boldsymbol{\theta}^*_{\text{comp}})$ leads to a rank-2 mean tensor to relate the HPC Conjecture while TBM [**?** , Theorem 7] constructs a rank-1 mean tensor. Hence, we emphasize that dTBM-specific techniques are required to obtain our impossibility results, though the proof idea is common for minimax lower bound analysis.

B. Proof sketch of Theorem 4

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

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

$$\lesssim \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^K}\|\hat{\mathcal{X}}-\mathcal{X}\|_F^2$$
$$\lesssim \frac{r^K p^{-K/2}}{\mathrm{SNR}}, \tag{19}$$

where $\mathcal{X} = \mathbb{E}\mathcal{Y}$ is the true mean. The first inequality in (19) 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~~ The second inequality relies on the key Lemma 1, which indicates

$$\min_{z(i)\neq z(j)} \|[\boldsymbol{X}_{i:}]^s - [\boldsymbol{X}_{j:}]^s\| \gtrsim \Delta_{\min}, \tag{20}$$

where $\boldsymbol{X} = \mathrm{Mat}(\mathcal{X})$. The most challenging part in the proof of Theorem 4 lies in the derivation of inequality (20) (or the proof of Lemma 1), in which the proof of [**?**] is no longer applicable due to different angle gap assumption in our dTBM. ~~We~~ To address the angle gap notion, we develop the extra padding technique in Lemma **??** and balance assumption (5)~~to derive~~. Last, we finish the proof of Theorem 4 by showing the third inequality of (19) using [**?** , Proposition 1].

C. Proof sketch of Theorem 5

The proof of Theorem 5 is inspired by the proof idea of [**?** , Theorem 2]. We develop extra polar-coordinate based techniques with angle gap characterization to address the nuisance degree heterogeneity. Recall the intermediate quantity, misclustering loss, defined in (10)

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

~~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~~ misclustering error of interest via the inequality $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2}$ in Lemma 2. Therefore, it suffices to control $L^{(t)}$. Further, we introduce the oracle estimators for core tensor under the true cluster assignment via

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

where $\boldsymbol{W} = \boldsymbol{M}\left(\mathrm{diag}(\boldsymbol{1}_p^T\boldsymbol{M})\right)^{-1}$ 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

$$\mathbb{1}\left\{z^{(t+1)}(i)=b\right\} = \mathbb{1}\left\{\|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \leq \|[\boldsymbol{Y}_{i:}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2\right\}$$
$$\leq A_{ib} + B_{ib}, \tag{21}$$

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

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

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

The terms $F_{ib}^{(t)}, G_{ib}^{(t)}, H_{ib}^{(t)}$ are controlled by $z^{(t)}, \mathcal{S}^{(t)}$; see the detailed definitions in (**??**), (**??**), (**??**). 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 (21) decomposes the misclustering loss in the $(t+1)$-th iteration into the oracle loss and the loss in t-th iteration. This decomposition leads to the separation of statistical error and computational error in the final upper bound of Theorem 5.

Specifically, we prove the contraction inequality

$$L^{(t+1)} \leq M\xi + \rho L^{(t)},$$
$$\text{with } \xi = \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}A_{ib}\left\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\right\|^2, \tag{22}$$

where M is a positive constant, $\rho \in (0,1)$ 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 (22) 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 [**?**] 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

$$\xi \lesssim \mathrm{SNR}^{-1}\exp\left(-\frac{p^{K-1}\mathrm{SNR}}{r^{K-1}}\right). \tag{23}$$

Combining the decomposition (22) and the oracle bound (23), we finish the proof of Theorem 5.

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