Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

Jiaxin Hu
University of Wisconsin – Madison

Miaoyan Wang
University of Wisconsin – Madison

Abstract

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

1 INTRODUCTION

Multiway arrays have been widely collected in various fields including social networks (Anandkumar et al., 2014), neuroscience (Wang et al., 2017), and computer science (Koniusz and Cherian, 2016). Tensors effectively represent the multiway data and serve as the foundation in higher-order data anal-

ysis. One data example is from multi-tissue multi-individual gene expression study (Wang et al., 2019; Hore et al., 2016), where the data tensor consists of expression measurements indexed by (gene, individual, tissue) triplets. Another example is *hypergraph* network (Ghoshdastidar and Dukkipati, 2017; Ghoshdastidar et al., 2017; Ahn et al., 2019; Ke et al., 2019) in social science. A K-uniform hypergraph can be naturally represented as an order-K tensor, where each entry indicates the presence of K-way hyperedge among nodes (a.k.a. entities). In both examples, identifying the similarity among tensor entities is important for scientific discovery.

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

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*.
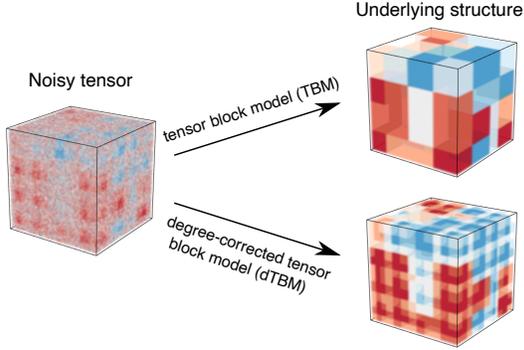
Figure 1: Examples for order-3 TBM with and without degree correction. Both TBM and dTBM have four communities on each mode, while dTBM allows a richer structure with degree heterogeneity.

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

The *degree-corrected tensor block model* (dTBM) has been proposed recently to account for the degree heterogeneity (Ke et al., 2019). The dTBM combines a higher-order checkerboard structure with degree parameter $\theta = (\theta(1), \ldots, \theta(p))^T$ 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 4. The spherical clustering avoids the estimation of nuisance degree heterogeneity. The usage of angle similarity brings new challenges to the theoretical results, and we develop new polar-coordinate based techniques in the proofs.

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

• We develop a general dTBM and establish the identifiability for the uniqueness of clustering using the notion of angle separability.

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

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

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

Related Work. Our work is closely related to but also distinct from several lines of existing research. Table 1 summarizes the most relevant models.

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

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

Global-to-local algorithm strategy. Our methods gen-

statistically impossible | statistically possible but computationally inefficient | computationally efficient → SNR

$$\mathcal{O}(p^{-(K-1)}) \qquad\qquad \mathcal{O}(p^{-K/2})$$

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

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

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

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

Notation. We use lower-case letters (e.g., a, b) for scalars, lower-case boldface letters (e.g., $\boldsymbol{a}, $) 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 $1\{\cdot\}$ 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 subvector by restricting the indices in the set $I \subset [p]$. We use $\|\boldsymbol{a}\| = \sqrt{\sum_i a^2(i)}$ to denote the ℓ_2-norm, $\|\boldsymbol{a}\|_1 = \sum_i |a_i|$ to denote the ℓ_1 norm of $\boldsymbol{a}$. For two vector $\boldsymbol{a}, \boldsymbol{b}$ of the same dimension, we denote the angle between $\boldsymbol{a}, \boldsymbol{b}$ by $\cos(\boldsymbol{a}, \boldsymbol{b}) = \langle \boldsymbol{a}, \boldsymbol{b}\rangle / \|\boldsymbol{a}\| \|\boldsymbol{b}\|$, 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 $(p_1, \ldots, p_K)$-dimensional tensor $\mathcal{X}$

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

where the entries of $\mathcal{X}$ are defined by $\mathcal{X}(i_1, \ldots, i_K) = \sum_{(j_1, \ldots, j_K)} \mathcal{S}(j_1, \ldots, j_K) \boldsymbol{M}_1(i_1, j_1) \cdots \boldsymbol{M}_K(i_K, j_K)$. For a matrix $\boldsymbol{Y}$, we use $\boldsymbol{Y}_{i:}$ (respectively, $\boldsymbol{Y}_{:i}$) to denote the i-th row (respectively, i-th column) of the matrix. Similarly, for an order-3 tensor, we use $\mathcal{Y}_{::i}$ to denote the i-th matrix slide of the tensor. We use $\text{Ave}(\cdot)$ to denote the operation of taking averages across elements and $\text{Mat}_k(\cdot)$ to denote the unfolding operation that reshapes the tensor along mode k into a matrix. For a symmetric tensor $\mathcal{Y} \in \mathbb{R}^{p \times \cdots \times p}$, we omit the subscript and use $\text{Mat}(\mathcal{Y}) \in \mathbb{R}^{p \times p^{K-1}}$ to denote the unfolding. For two sequences $\{a_p\}, \{b_p\}$, we denote $a_p \lesssim b_p$ or $a_p = O(b_p)$ if $\lim_{p \to \infty} a_p/b_p \leq c$ for some constant $c \geq 0$, $a_p = o(b_p)$ if $\lim_{p \to \infty} a_p/b_p = 0$, and $a_p = \Omega(b_p)$ if both $b_p \lesssim a_p$ and $a_p \lesssim b_p$. Throughout the paper, we use the terms "community" and "clusters" exchangeably.

2 MODEL FORMULATION

2.1 Degree-corrected Tensor Block Model

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

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

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

$$\mathbb{E}\,\mathcal{Y} = \mathcal{S} \times_1 \Theta M \times_2 \cdots \times_K \Theta M, \qquad (1)$$

where $\Theta = \mathrm{diag}(\theta(1), ..., \theta(p)) \in \mathbb{R}^{p \times p}$ is a diagonal matrix, $M \in \{0,1\}^{p \times r}$ is the membership matrix associated with community assignment z such that $M(i,j) = 1\{z(i) = j\}$. By definition, each row of M has one copy of 1's and 0's elsewhere. Note that the discrete nature of M renders our model (1) more challenging than Tucker decomposition. We call a tensor $\mathcal{Y}$ an r-block tensor with degree θ if $\mathcal{Y}$ admits (1). We are particularly interested in high-dimensional regime where p grows whereas $r = O(1)$. The extension to general asymmetrical dTBM is obtained via replacing (M, Θ) in (1) by mode-specific parameters (M_k, Θ_k) for every mode $k \in [K]$. Here, we give two special cases of dTBM.

Example 1 (Gaussian TBM)**.** Let $\theta(i) = 1$ for all $i \in [p]$ and $\mathcal{E}$ be a noise tensor with i.i.d. $N(0, \sigma^2)$ entries. Our dTBM reduces to a non-degree Gaussian TBM (Wang and Zeng, 2019; Han et al., 2020), which is widely used in previous clustering algorithms (Wang and Zeng, 2019; Chi et al., 2020).

Example 2 (Binary dTBM)**.** Consider a K-uniform hypergraph $H = (V, E)$, where $V = [p]$ collects the nodes with r disjoint communities and E collects all the K-way hyperedges. Let $\mathcal{Y} \in \{0,1\}^{p \times \cdots \times p}$ denote the adjacency tensor, where the entries encode the presence or absence of hyperdeges among p nodes. Specifically, let $\mathcal{Y}(i_1, ..., i_K) = 1$ if $(i_1, ..., i_K) \in E$, otherwise, $\mathcal{Y}(i_1, ..., i_K) = 0$, for all $(i_1, ..., i_K) \in [p]^K$. The equation (1) models $\mathbb{E}\,\mathcal{Y}$ with degree heterogeneity and subgaussianity parameter $\sigma^2 = 1/4$.

2.2 Identifiability under Angle Gap

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 $(z, \mathcal{S}, \theta)$,

$$\mathcal{P} = \left\{ (z, \mathcal{S}, \theta) \colon \theta \in \mathbb{R}^p_+, \text{for } a \in [r], \frac{c_1 p}{r} \leq |z^{-1}(a)| \leq \frac{c_2 p}{r}, \right.$$

$$\left. c_3 \leq \|\mathrm{Mat}(\mathcal{S})_{a:}\| \leq c_4, \|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)| \right\}, \qquad (2)$$

where $c_i > 0$'s are universal constants. We briefly describe the rationale of the constraints in (2). First,

the entrywise positivity constraint on $\theta \in \mathbb{R}^p_+$ is imposed to avoid sign ambiguity between entries in $\theta_{z^{-1}(a)}$ and $\mathcal{S}$. This constraint allows the trigonometric cos to describe the angle similarity in the Assumption 1 below and Sub-algorithm 2 in Section 4. 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 Supplement A.1. Second, recall that the quantity $|z^{-1}(a)|$ denotes the number of nodes in a-th community. The constants c_1, c_2 in the $|z^{-1}(a)|$ bound assume the roughly balanced size across r communities. Third, the constants c_3, c_4 in the magnitude of $\mathrm{Mat}(\mathcal{S})_{a:}$ requires no purely zero slide in $\mathcal{S}$, so the core tensor $\mathcal{S}$ is not trivially reduced to a lower rank. Lastly, the ℓ_1 normalization $\|\theta_{z^{-1}(a)}\|_1 = |z^{-1}(a)|$ is imposed to avoid the scalar ambiguity between $\theta_{z^{-1}(a)}$ and $\mathcal{S}$. This constraint, again, incurs no restriction to model flexibility but makes our presentation cleaner. See Supplement A.2 for the parameter space comparison with previous work (Gao et al., 2018; Han et al., 2020; Ke et al., 2019).

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 (Wang and Zeng, 2019). Therefore, one can not apply the classical identifiability conditions for low-rank tensors to dTBM. Here, we introduce a key separation condition on the core tensor.

Assumption 1 (Angle gap)**.** Let $S = \mathrm{Mat}(\mathcal{S})$. Assume the minimal gap between normalized rows of S is bounded away from zero; i.e., for $r \geq 2$,

$$\Delta_{\min} := \min_{a \neq b \in [r]} \left\| \frac{S_{a:}}{\|S_{a:}\|} - \frac{S_{b:}}{\|S_{b:}\|} \right\| > 0. \qquad (3)$$

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

Theorem 1 (Model identifiability)**.** Consider the dTBM with $r \geq 2$. The parameterization (1) is

unique in $\mathcal{P}$ up to cluster label permutations, if and only if Assumption 1 holds.

The identifiability guarantee for the dTBM is more appealing than classical Tucker model. In the Tucker model, the factor matrix $\mathbf{M}$ is identifiable only up to orthogonal rotations. In contrast, our model does not suffer from rotational invariance. This property benefits the interpretation of dTBM in practice.

3 THEORETICAL LIMITS

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

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

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

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

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

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

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

In Sections 3.1 and 3.2, we provide the lower bounds of $\ell(\hat{z}, z)$ for general Gaussian dTBMs (1) without symmetric assumptions. For general (asymmetric) Gaussian dTBMs, we assume Gaussian noise $\mathcal{E}(i_1, \ldots, i_K) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, and we extend the parameter space (2) to allow K clustering functions $(z_k)_{k \in [K]}$, one for each mode. For notational simplicity, we still use z and $P(\gamma)$ for this general (asymmetric) model. All lower bounds should be interpreted as the worst-case results across K modes.

3.1 Statistical Critical Values

The statistical limit means the minimal SNR required for solving dTBMs with *unlimited computational cost.* Our following result shows the minimax lower bound of SNR for exact recovery in dTBM.

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

$$\sup_{(z,S,\varepsilon) \in P(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{\mathsf{stat}}, z)\right] \gtrsim 1.$$

Theorem 2 demonstrates the impossibility of exact recovery of the assignment when $\gamma < (K-1)$ in the high-dimensional regime $p \to \infty$ for fixed r. The proof is information-theoretical, and therefore the results apply to all statistical estimators, including but not limited to, maximum likelihood estimation (MLE) (Wang and Zeng, 2019) and trace maximization (Ghoshdastidar and Dukkipati, 2017). As we will show in Section 4, the SNR threshold $(K-1)$ is also a minimax upper bound, because MLE achieves exact recovery when $\gamma > (K-1)$. Hence, the boundary $\gamma_{\mathsf{stat}} := (K-1)$ is the critical value for statistical performance of dTBM.

3.2 Computational Critical Values

The computational limit means the minimal SNR required for exactly recovery with *polynomial-time* computational cost. An important ingredient to establish the computational limits is the *hypergraphic planted clique (HPC) conjecture* (Zhang and Xia, 2018; Brennan and Bresler, 2020). 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 Supplement C.4. Under the HPC conjecture, we establish the SNR lower bound that is necessary for any *polynomial-time* estimator to achieve exact clustering.

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

$$\liminf_{p \to \infty} \sup_{(z,S,\varepsilon) \in P(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{\mathsf{comp}}, z)\right] \gtrsim 1.$$

Theorem 3 indicates the impossibility of exact recovery by polynomial-time algorithms when $\gamma < K/2$. Therefore, $\gamma_{\mathsf{comp}} := K/2$ is the critical value for computational performance of dTBM. In Section 4, we will show the condition $\gamma > K/2$ suffices for our proposed polynomial-time estimator. Thus, $\gamma_{\mathsf{comp}} := K/2$ is the critical value for computational performance of dTBM.

Remark 1 (Statistical-computational gaps)**.** Now, we have established the phase transition of exact clustering under order-K dTBM by combing Theorems 2 and 3. Figure 2 summarizes our results of critical SNRs when $K \geq 2$. Particularly, dTBM reduces to matrix degree-corrected model when $K = 2$, and the statistical and computational bounds show the same critical value. When $K = 1$, dTBM reduces to the degree-corrected sub-Gaussian mixture model (GMM). Earlier work (Lu and Zhou, 2016) implies that polynomial-time algorithms are able to achieve the statistical minimax lower bound in GMM. Hence, the statistical-to-computational gap emerges only for higher-order tensors with $K \geq 3$, which reveals the intrinsic distinctions among (vector) one-dimensional clustering, (matrix) biclustering, and (tensor) higher-order clustering. We also find that the extra complexity from θ does not render the estimation of z qualitatively harder; see comparison of our phase transition with non-degree TBM (Han et al., 2020).

4 ALGORITHM

In this section, we present an efficient polynomial-time clustering algorithm under mild SNR. The procedure takes a global-to-local approach. See Figure 3 for illustration. The global step finds the basin of attraction with polynomial miclustering error, whereas the local iterations improve the initial clustering to exact recovery. Both steps are critical to obtain a satisfactory algorithm output.
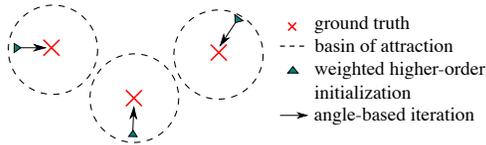

Figure 3: Illustration of our global-to-local algorithm.

4.1 Initialization

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

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

This implies that, all node i belonging to a-th community (i.e., $z(i) = a$) share the same normalized mean vector $\theta(i)^{-1} \mathbf{X}_{i:}$, and vice versa. Intuitively, one can apply k-means clustering to the vectors $\{\theta(i)^{-1} \mathbf{X}_{i:}\}_{i \in [p]}$, which leads to main idea of our Sub-algorithm 1. Specifically, our initialization consists of denoising step and clustering step. The denoising step (lines 1-2 in Sub-algorithm 1) estimates

X from Y by a double projection spectral method. The double projection improves usual matrix spectral methods in order to alleviate the noise effects for $K \geq 3$ (Han et al., 2020). The clustering step (lines 3-5 in Sub-algorithm 1) performs the weighted k-means clustering. The choice of weights is to bound the k-means objective function by the Frobenius-norm accuracy of $\hat{X}$. Unlike existing clustering algorithm (Ke et al., 2019), we apply the clustering on the unfolded tensor $\hat{\mathbf{X}}$ rather than on the factors $\hat{\mathbf{U}}$. This strategy relaxes the singular-value gap condition (Gao et al., 2018; Han et al., 2020). Full procedures are provided in Sub-algorithm 1.

We now establish the misclustering error rate of initialization. We call θ is balanced if the relative extent of heterogeneity is comparable across clusters in that

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

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

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

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

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

Remark 3 (Failure of conventional tensor HOSVD)**.** If we use conventional HOSVD for tensor denoising; that is, we use $\mathbf{U}_{\mathrm{pre}}$ in place of $\hat{\mathbf{U}}$ in line 2, then the misclustering rate becomes $O(p^{-1})$ for all $K \geq 2$. This rate is substantially worse than our rate (7).

4.2 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

---

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.

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

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

3: Let $\hat{\boldsymbol{X}} = \mathrm{Mat}(\hat{\mathcal{X}})$ and $S_0 = \{i \in [p]: \hat{\boldsymbol{X}}_{i:} = 0\}$. Set $\hat{z}(i)$ randomly in $[r]$ for $i \in S_0$.

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

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

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

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

Sub-algorithm 2: Angle-based iteration

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

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

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

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

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

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

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

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

11: **end for**

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

---

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 $\boldsymbol{x}_i = \theta(i)\boldsymbol{s}_{z(i)} + \epsilon_i, i \in [p]$, where $\boldsymbol{s}_i$'s are known cluster centroids, $\theta(i)$'s are unknown positive degrees, and $z: [p] \to [r]$ 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.,

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

The classical Euclidean-distance based clustering (Han et al., 2020) fails to recover z in the presence of degree heterogeneity, even under noiseless case. In contrast, the angle-based k-means clustering achieves accurate recovery without explicit estimation of θ. Our Sub-algorithm 2 shares the same spirit as angle-based k-means, except that we use estimated centroids $\boldsymbol{s}_a^{(t)}$ in place of $\boldsymbol{s}_a$ based on estimated assignment in previous iterations. See Sub-algorithm 2 for full procedures.

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

Definition 1 (Locally linear stability)**.** Define the ε-neighborhood of z by $N(z, \varepsilon) = \{\bar{z}: \ell(\bar{z}, z) \leq \varepsilon\}$. Let $\bar{z}: [p] \to [r]$ be a clustering function. The degree is ε-locally linearly stable if and only if

$$\sin(\boldsymbol{p}(\bar{z}), \boldsymbol{p}_\theta(\bar{z})) \lesssim \varepsilon\Delta_{\mathsf{min}}, \quad \text{for all } \bar{z} \in N(z, \varepsilon), \quad (8)$$

where $\boldsymbol{p}(\bar{z}) = (|\bar{z}^{-1}(1)|, \ldots, |\bar{z}^{-1}(r)|)^T$ and $\boldsymbol{p}_\theta(\bar{z}) = (\sum_{z^{-1}(1)} \theta_1, \ldots, \sum_{z^{-1}(r)} \theta_1)^T$.

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 (8) controls the impact of node degree to the $\boldsymbol{p}_\theta(\cdot)$ with respect to the misclassification rate ε and angle gap.

Theorem 5 (Error for angle-based iteration)**.** Consider the setup as in Theorem 4. Suppose $r = \mathcal{O}(1)$ and SNR $\geq \tilde{C}p^{-K/2}\log p$ for some sufficiently large constant $\tilde{C}$. Assume the local linear stability of degree holds in the neighborhood $N(z, \varepsilon)$ for all $\varepsilon \leq E_0$ and some $E_0 \geq \check{C}\log^{-1}p$ with some positive constant $\check{C}$. Let $z^{(t)}$ denote the t-th iteration output in Sub-algorithm 2 with initialization $z^{(0)}$ from Sub-algorithm 1. With probability going to 1, there exists a contraction parameter $\rho \in (0, 1)$ such that

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

The iteration error is decomposed into two parts: statistical error and computational error. The statistical error is unavoidable with noisy data regardless t, whereas the computational error decays in an exponential rate as the number of iterations $t \to \infty$. Theorem 5 implies that, with probability going to 1, our estimate $z^{(T)}$ achieves exact recovery within polynomial iterations; more precisely,

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

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

5 NUMERICAL STUDIES

We evaluate the performance of our algorithm[1] in this section. We report average errors and standard deviations across 30 replications in each experiment. Clustering accuracy is assessed by clustering error rate (CER, i.e., one minus rand index). Note that CER between $(\hat{z}, z)$ is equivalent to misclustering error $\ell(\hat{z}, z)$ up to constant multiplications (Meilă, 2012), and lower CERs indicate better performances.

We generate order-3 tensors with *assortative* (Gao et al., 2018) core tensors to control SNR; i.e., we set $\mathcal{S}_{aaa} = s_1$ for $a \in [r]$ and others be s_2, where $s_1 > s_2 > 0$. Let $\alpha = s_1/s_2$. We set α close to 1 such that $1 - \alpha = o(p)$. In particular, we have $\alpha = 1 + \Omega(p^{\gamma/2})$ with $\gamma < 0$ by Assumption 1 and definition (4). Hence, we easily adjust SNR via varying α. 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 θ 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 θ to satisfy (2). We set $\sigma^2 = 1$ for Gaussian data.

5.1 Verification of Theoretical Results

The first experiment verifies statistical-computational gap described in Section 3. Consider the Gaussian
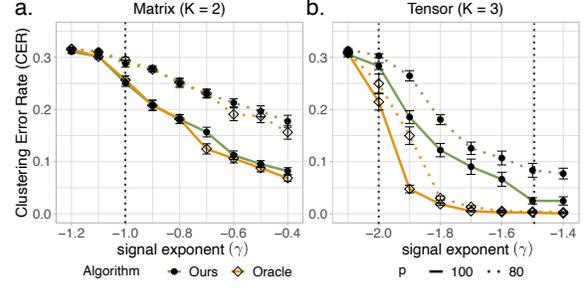

Figure 4: SNR phase transitions for clustering in dTBM with $p = \{80, 100\}, r = 5$ under (a) matrix case with $\gamma \in [-1.2, -0.4]$ and (b) tensor case with $\gamma \in [-2.1, -1.4]$.
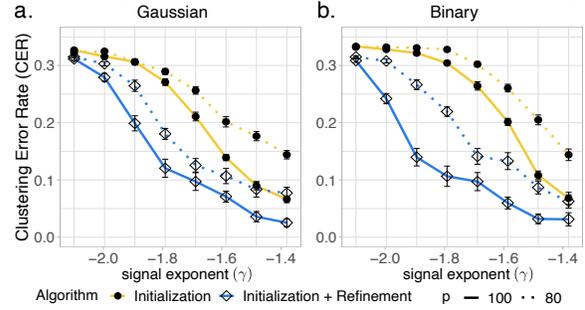

Figure 5: CER versus signal exponent (γ) for initialization only and for combined algorithm. We set $p = \{80, 100\}, r = 5, \gamma \in [-2.1, -1.4]$ under (a) Gaussian models and (b) Bernoulli models.

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.

5.2 Comparison with Other Methods

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

• **HOSVD**: HOSVD on data tensor and k-means on the rows of the factor matrix;

• **HOSVD+**: HOSVD on data tensor and k-means on the ℓ_2-normalized rows of the factor matrix;

• **HLloyd** (Han et al., 2020): High-order clustering algorithm developed for non-degree TBM;

• **SCORE** (Ke et al., 2019): Tensor-SCORE for clustering developed for binary tensors.

Among the four alternative algorithms, the **SCORE** is the closest method to ours. We set the tuning parameters of **SCORE** as in previous literature (Ke et al., 2019). The methods **SCORE** and **HOSVD+** are designed for dTBM (1), 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 under Gaussian and Bernoulli models with $p = 100, r = 5$. We call our algorithm as **dTBM** in comparison.

We investigate the effects of signal to clustering performance by varying $\gamma \in [-1.5, -1.1]$. Figure 6 shows the consistent outperformance of our method **dTBM** among all algorithms. The sub-optimality of **SCORE** and **HOSVD+** indicates the necessity of local iterations on the clustering. Furthermore, Figure 6 shows the inadequacy of non-degree algorithms in the presence of mild degree heterogeneity. The only exception is the slightly better performance of **HLloyd** over **HOSVD+** under Gaussian model. However, we find the advantage of **HLloyd** disappears with higher degree heterogeneity; see Supplement B. The experiment demonstrates the benefits of addressing heterogeneity in higher-order clustering tasks.

The last experiment investigates the effects of degree heterogeneity to clustering performance. We use the same setting as in the first experiment in the Section 5.2, except that we fix the signal exponent $\gamma = -1.2$ and generate the degree heterogeneity 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 the shape parameter $a \in [3, 6]$ and choose b such that $\mathbb{E}X = a(a-1)^{-1}b = 1$ for X following Pareto(a, b). Note that a smaller a leads to a larger variance in θ and hence a larger degree heterogeneity. Figure 7 demonstrates the stability of degree-corrected algorithms (**dTBM**, **SCORE**, **HOSVD+**) over the entire range of degree heterogeneity under consideration. In contrast, non-degree algorithms (**HLloyd**, **HOSVD**) show poor performance with

large heterogeneity, especially in Bernoulli cases. This experiment, again, highlights the benefit of addressing degree heterogeneity in clustering.
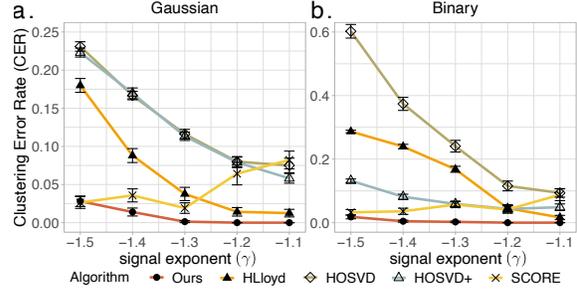

Figure 6: CER versus signal exponent (denoted γ) for different methods. We set $p = 100, r = 5, \gamma \in [-1.5, -1.1]$ under (a) Gaussian and (b) Bernoulli models.
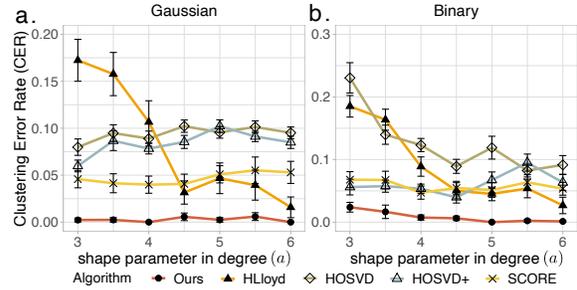

Figure 7: CER versus shape parameter in degree (denoted $a \in [3, 6]$) for different methods with $p = 100, r = 5, \gamma = -1.2$ under (a) Gaussian and (b) Bernoulli models.

5.3 Peru Legislation Data Analysis

We apply our method to the legislation networks in the Congress of the Republic of Peru (Lee et al., 2017). Because of the frequent political power shifts in the Peruvian Congress during 2006-2011, we choose to focus on the data for the first half of 2006-2007 year. The dataset records the co-sponsorship of 116 legislators from top 5 parties and 802 bill proposals. We reconstruct legislation network as an order-3 binary tensor $\mathcal{Y} \in \{0, 1\}^{116 \times 116 \times 116}$, where $\mathcal{Y}_{ijk} = 1$ if the legislators (i, j, k) have sponsored the same bill, and $\mathcal{Y}_{ijk} = 0$ otherwise. 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 2 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+**. The result is consistent with simulations under moderate heterogeneity.

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

Table 2: Clustering errors (measured by CER) for various methods in the analysis of Peru Legislation dataset.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplementary Material:
Multiway Spherical Clustering via Degree-Corrected Tensor Block Models

A   PARAMETER SPACE

Here, we provide the parameter space comparison with previous works and extra explanation for the positvity constraint of degree parameters.

A.1   Positivity of Degree Parameters

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

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

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

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

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

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

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

A.2   Parameter Space Comparison with Previous Work

Table 3 indicates that our parameter space is flexible compared to existing work. The balanced community assumption is mild in literature. In fact, our simulation requires only the positive community size (i.e., $\min_{a \in [r]} |z^{-1}(a)| \geq 1$) but no balanced community assumptions; see Section 5. The outperformance of dTBM shows the robustness of our algorithm to the imbalanced communities.

B   ADDITIONAL EXPERIMENTS

We provide extra simulation results for the comparison with other higher-order clustering algorithms. We use the same setting as in the first experiment in the Section 5.2, except that we now generate the degree heterogeneity θ from Pareto distribution with shape parameter a prior to normalization. We consider the Gaussian model under low ($a = 6$) and high ($a = 2$) degree heterogeneity. Figure 8 shows that the errors for non-degree algorithms (**HLloyd**, **HOSVD**) increases with degree heterogeneity. In addition, the advantage of **HLloyd** over **HOSVD+** disappears with higher degree heterogeneity. This experiment supports the conclusion we obtained in Section 5.2.

Assumptions in parameter space	Gao et al. (2018)	Han et al. (2020)	Ke et al. (2019)	Ours
Balanced community size	p	p	p	p
Balanced degree parameters	p		p	p
Flexible in-group connections		$\bar{p}$		p
Gap among cluster centers	In-between cluster difference	Euclidean gap	Eigen gap	Angle gap

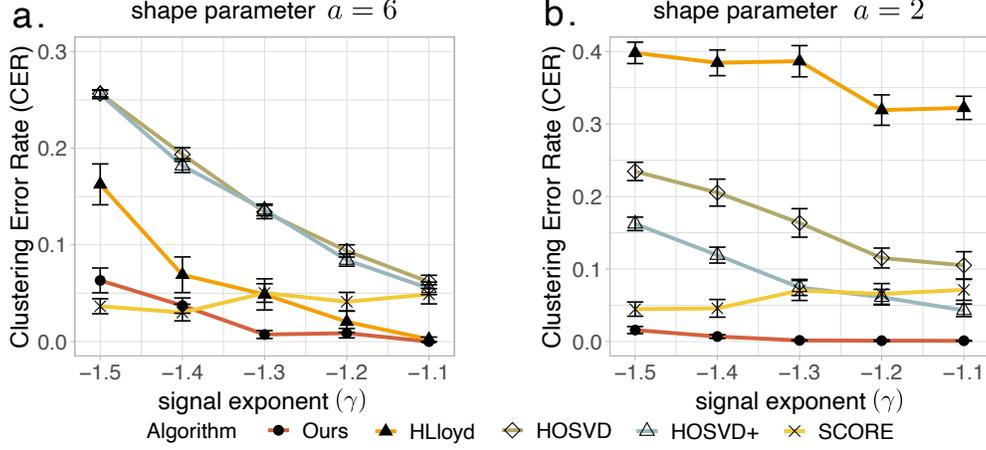Table 3: Comparison of parameter space of previous works with our work.

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

C PROOFS

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

C.1 Notation

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

Preliminaries.

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

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

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

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

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

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

Model extension to generalized dTBM.

The general order-K $(p_1, \ldots, p_K)$-dimensional dTBM model with r_k communities and degree heterogeneity

13

$_k = \{\theta_k(i)\} \in \mathbb{R}_+^{p_k}$ is represented by

$$Y = X + E, \quad \text{where} \quad X = S \times_1 \Theta_1 M_1 \times_2 \cdots \times_K \Theta_K M_K, \tag{9}$$

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

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

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

$$
\left. _k \in \mathbb{R}_+^p, \frac{c_1 p_k}{r_k} \le |z_k^{-1}(a)| \le \frac{c_2 p_k}{r_k}, c_3 \le \|S_{k,a:}\| \le c_4, |_{k,z_k^{-1}(a)}| = |z_k^{-1}(a)|, a \in [r_k], k \in [K] \right\}.
$$

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

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

We also consider the generalized Assumption 1 on angle gap.

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

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

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

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

C.2   Proof of Theorem 1

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

$$
X = S \times_1 \Theta_1 M_1 \times_2 \cdots \times_K \Theta_K M_K^\ell = S^\ell \times_1 \Theta_1^\ell M_1^\ell \times_2 \cdots \times_K \Theta_K^\ell M_K^\ell, \tag{10}
$$

where $(\{z_k\}, S, \{_k\}) \in P(\{r_k\})$ and $(\{z_k^\ell\}, S^\ell, \{_k^\ell\}) \in P(\{r_k^\ell\})$ are two sets of parameters. We prove the sufficient and necessary conditions separately.

$(\Leftarrow)$ For the necessity, it suffices to construct two distinct parameters up to cluster label permutation, if the model (9) violates Assumption 2. Without loss of generality, we assume $\left\| S_{1,1:}^s - S_{1,2:}^s \right\| = 0$.

If $S_{1,1:}$ is a zero vector, construct $_1^\ell$ such that $_{1,z_1^{-1}(1)}^\ell \neq _{1,z_1^{-1}(1)}$. Let $\{z_k^\ell\} = \{z_k\}$, $S^\ell = S$, and $_k^\ell = _k$ for all $k = 2, \ldots, K$. Then the triplet $(\{z_k^\ell\}, S^\ell, \{_k^\ell\})$ is distinct from $(\{z_k\}, S, \{_k\})$ up to label permutation. Similar conclusion holds when $S_{1,2:}$ is a zero vector.

If neither $S_{1,1:}$ nor $S_{1,2:}$ is a zero vector, there exists a positive constant c such that $S_{1,1:} = c S_{1,2:}$. Thus, there exists a core tensor $S_0 \in \mathbb{R}^{r_1 - 1 \times \cdots \times r_K}$ such that

$$
S = S_0 \times_1 CR, \quad \text{where} \quad C = \mathrm{diag}(1, c, 1, \ldots, 1) \in \mathbb{R}^{r_1 \times r_1}, \quad R = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{1}_{r_1 - 2} \end{pmatrix} \in \mathbb{R}^{r_1 \times (r_1 - 1)}.
$$

14

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

$$\boldsymbol{M}_1^\ell = \boldsymbol{M}_1 \boldsymbol{R}, \quad S^\ell = S_0 \times_1 \boldsymbol{D}, \quad \theta_1^\ell(i) = \begin{cases} \frac{1}{1+c}\theta_1(i) & i \in z_1^{-1}(1), \\ \frac{c}{1+c}\theta_1(i) & i \in z_1^{-1}(2), \\ \theta_1(i) & \text{otherwise,} \end{cases}$$

and $\boldsymbol{M}_k^\ell = \boldsymbol{M}_k$, $\ell_k = \ell_k$ for all $k = 2, \ldots, K$. Then we have constructed a triplet $(\mathcal{f}z_k^\ell \mathcal{g}, S^\ell, \mathcal{f} \ell_k \mathcal{g})$ that is distinct from $(\mathcal{f}z_k \mathcal{g}, S, \mathcal{f} \ell_k \mathcal{g})$ up to label permutation.

$(\Leftarrow)$ For the sufficiency, it suffices to show that all possible triplets $(\mathcal{f}z_k^\ell \mathcal{g}, S^\ell, \mathcal{f} \ell_k \mathcal{g})$ are identical to $(\mathcal{f}z_k \mathcal{g}, S, \mathcal{f} \ell_k \mathcal{g})$ up to label permutation if the model (9) satisfies Assumption (2). We show the uniqueness of the three parameters, $\mathcal{f}\boldsymbol{M}_k\mathcal{g}, \mathcal{f}S\mathcal{g}, \mathcal{f} \ell_k \mathcal{g}$ separately.

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

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

Next, we show the uniqueness of ℓ_k for all $k \in [K]$ provided that $z_k = z_k^\ell$. Similarly, consider $k = 1$ only, and omit the procedure for $k \geq 2$.

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

$$\frac{\boldsymbol{X}_{1, z_1(i):}}{\boldsymbol{X}_{1, z_1(j):}} = \frac{\boldsymbol{X}_{1, z_1(i):}^\ell}{\boldsymbol{X}_{1, z_1(j):}^\ell}, \quad \text{which implies} \quad \frac{\theta_1(j)}{\theta_1(i)} = \frac{\theta_1^\ell(j)}{\theta_1^\ell(i)}. \tag{11}$$

Let $\theta_1^\ell(j) = c\theta_1(j)$ for some positive constant c. By equation (11), we have $\theta_1^\ell(i) = c\theta_1(i)$ for all $i \in z_1^{-1}(a)$. By the constraint $(\mathcal{f}z_k \mathcal{g}, S^\ell, \mathcal{f} \ell_k \mathcal{g}) \in \mathcal{P}(\mathcal{f}r_k \mathcal{g})$, we have

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

which implies $c = 1$. Hence, we have proven $\ell_1 = \ell_1^\ell$ provided that $z_1 = z_1^\ell$.

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

Let $\boldsymbol{D}_k = \left[(\boldsymbol{\Theta}_k^\ell \boldsymbol{M}_k^\ell)^T (\boldsymbol{\Theta}_k^\ell \boldsymbol{M}_k^\ell) \right]^{-1} (\boldsymbol{\Theta}_k^\ell \boldsymbol{M}_k^\ell)^T, k \in [K]$. By the parameterization (10), we have

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

Therefore, we finish the proof of Theorem 1. □

Useful Lemma for the Proof of Theorem 1

Lemma 1 (Motivation of angle-based clustering). Consider the signal tensor $\mathcal{X}$ in the generalized dTBM (9) with $(\mathcal{f}z_k \mathcal{g}, S, \mathcal{f} \ell_k \mathcal{g}) \in \mathcal{P}(\mathcal{f}r_k \mathcal{g})$ and $r_k \geq 2$. Then, for any $k \in [K]$ and index pair $(i, j) \in [p_k]^2$, we have

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

15

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

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

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

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

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

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

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

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

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

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

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

C.3 Proof of Theorem 2

Proof of Theorem 2. We will prove a more general conclusion than the main paper by allowing growing r_k's. Consider the generalized dTBM (9) in the special case that $p_k = p$ and $r_k = r$ for all $k \in [K]$. Specifically, we will show that, under the assumptions $K \geq 1, r \lesssim p^{1/3}$ and SNR condition

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

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

$$\sup_{(\{z_k\}, S, \{\theta_k\}) \in \mathcal{P}(\gamma)} \mathbb{E}\left[p\ell(\hat{z}_{k,\text{stat}}, z_k)\right] \gtrsim 1. \tag{12}$$

Since the inequality (12) is a minimax lower bound, it suffices to show the inequality holds for a particular $(\{z_k\}, S, \{\theta_k\}) \in \mathcal{P}(\gamma)$. Specifically, we consider the estimation problem based on a particular parameter point $(\{z_k\}, S, \{\theta_k\})$ with the following three properties:

$$\text{(i) } \theta_k(i) = 1 \text{ for all } i \in [p]; \quad \text{(ii) } \Delta_{\min} \asymp \left(\frac{p}{r}\right)^{-\frac{K-1}{2}}\sigma; \quad \text{(iii) } |z_k^{-1}(a)| = \frac{p}{r} \in \mathbb{Z}_+ \text{ for all } a \in [r], \tag{13}$$

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

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

16

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

$$\inf_{\hat{z}_k} \sup_{z_k} \mathsf{E}\ell(\hat{z}_k, z_k) \gtrsim \frac{1}{r^3 |T_k^c|} \sum_{i \in T_k^c} \inf_{\hat{z}_k} \{\mathsf{P}[\hat{z}_k(i) = 2 | z_k(i) = 1] + \mathsf{P}[\hat{z}_k(i) = 1 | z_k(i) = 2]\}, \tag{14}$$

where $\hat{z}_k$ and z_k on the left hand side denote the generic clustering functions in $P(\gamma)$, z_k on the right hand side denotes a particular parameter satisfying properties (i)-(vi), and the infimum on the right hand side is taken over the restricted family of $\hat{z}$ satisfying (iv)-(vi). Here, the factor $r^3 = r \cdot r^2$ in (14) comes from two sources: $r^2 \asymp \binom{r}{2}$ comes from the multiple testing burden for all pairwise comparisons among r clusters; and another r comes from the number of elements $|T_k^c| \asymp p/r$ to be clustered.

Next, we need to find the lower bound of the rightmost side in (14). For simplicity, we show the bound for the mode-1 case $k = 1$ only. We drop the subscripts 1 in $z_1, T_1, \mathbf{S}_1, \cdot_1$ and omit the repeated procedures for the cases of $k = 2, \ldots, K$.

We consider the hypothesis test based on model (9). First, we reparameterize the model under the construction (13)

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

where $\mathbf{x}_a$'s are centroids in $\mathsf{R}^{p^{K-1}}$. Without loss of generality, we consider the lower bound for the summand in (14) for $i = 1$. The analysis for other $i \in T^c$ are similar. For notational simplicity, we suppress the subscript i and write $\mathbf{y}, \theta, z$ in place of $\mathbf{y}_1, \theta_1$ and $z(1)$, respectively. The equivalent vector problem for assessing the summand in (14) is

$$\mathbf{y} = \theta \mathbf{x}_z + \mathbf{e}, \tag{15}$$

where $\theta \in \mathsf{R}_+$ and $z \in \{1, 2\}$ are unknown parameters, $\mathbf{x}_1, \mathbf{x}_2 \in \mathsf{R}^{p^{K-1}}$ are given centroids, and $\mathbf{e} \in \mathsf{R}^{p^{K-1}}$ consists of i.i.d. $N(0, \sigma^2)$ entries. Then, we consider the hypothesis testing under the model (15):

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

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

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

and the MLE's of θ and z are

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

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

$$\hat{z}_{\mathsf{MLE}} = \begin{cases} 1 & \text{if } \cos(\mathbf{y}, \mathbf{x}_1) \geq \cos(\mathbf{y}, \mathbf{x}_2) \text{ and } \langle \mathbf{y}, \mathbf{x}_1 \rangle > 0, \\ 2 & \text{if } \cos(\mathbf{y}, \mathbf{x}_1) < \cos(\mathbf{y}, \mathbf{x}_2) \text{ and } \langle \mathbf{y}, \mathbf{x}_2 \rangle > 0, \\ 1 \text{ or } 2 \text{ with equal probability} & \text{otherwise.} \end{cases} \tag{16}$$

The Neyman-Pearson Lemma implies

$$\inf_{\hat{z}} \{\mathsf{P}[\hat{z} = 2 | z = 1] + \mathsf{P}[\hat{z} = 1 | z = 2]\} = \mathsf{P}[\hat{z}_{\mathsf{MLE}} = 1 | z = 2] + \mathsf{P}[\hat{z}_{\mathsf{MLE}} = 2 | z = 1]. \tag{17}$$

By symmetric, it suffices to bound $\mathsf{P}[\hat{z}_{\mathsf{MLE}} = 1 | z = 2]$. Using (16), we obtain

$$\mathsf{P}[\hat{z}_{\mathsf{MLE}} = 1 | z = 2] = \mathsf{P}\left[\cos(\theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_1) \geq \cos(\theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_2) \text{ and } \langle \theta \mathbf{x}_2 + \mathbf{e}, \mathbf{x}_1 \rangle > 0\right]$$

$$\overset{(*)}{\geq} \mathsf{P}\left[\left\langle \mathbf{e}, \frac{\mathbf{x}_1^s}{\|\mathbf{x}_1^s\|} - \frac{\mathbf{x}_2^s}{\|\mathbf{x}_2^s\|} \right\rangle \geq \frac{\theta}{2} \|\mathbf{x}_2\| \|\mathbf{x}_1^s - \mathbf{x}_2^s\|\right]$$

$$P\left[\langle \boldsymbol{e}, \boldsymbol{x}_1^s \rangle \geq -\frac{\theta}{2}\|\boldsymbol{x}_2\|(2-\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2)\right]$$

$$\overset{(\cdot)}{\geq} \Phi\left(\frac{\theta}{2}\|\boldsymbol{x}_2\|(2-\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|^2)\right) \cdot \Phi\left(\frac{\theta}{2}\|\boldsymbol{x}_2\|\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\|\right), \tag{18}$$

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

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

By construction (13) of $(\{z_k\}, S, \{\phi_k\})$ with three properties and lower bound $\min_{a \in [r]}\|\boldsymbol{S}_{a:}\| \geq c_3$ in the definition of $P(\gamma)$, we have $\theta = 1$, $\|\boldsymbol{x}_2\| \geq \|\boldsymbol{S}_{2:}\|\min_{a \in [r]}\left\|\sum_{z^{-1}(a)}\right\| \geq K^{-1} \& (\frac{p}{r})^{(K-1)/2}$. Also, note that under the construction (13)

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

which implies $\|\boldsymbol{x}_1^s - \boldsymbol{x}_2^s\| = \|\boldsymbol{S}_{1:}^s - \boldsymbol{S}_{2:}^s\| = \Delta_{\min} \geq 1$. Therefore, the equation (18) is lower bounded by

$$P[\hat{z}_{\mathsf{MLE}} = 1 | z = 2] \geq P\left[\left(\frac{p}{r}\right)^{(K-1)/2}\Delta_{\min} \cdot N(0,1) \geq \cdot\left(\frac{p}{r}\right)^{(K-1)/2}\right] \geq C > 0, \tag{19}$$

where the existence of strictly positive constant C is based on the SNR assumption (13). Combining (14), (17) and (19) yields

$$\inf_{\hat{z}_1}\sup_{(\{z_k\}, S, \{\phi_k\}) \in P(\gamma)} \mathsf{E}\ell(\hat{z}_1, z_1) \gtrsim C > 0,$$

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

$$\inf_{\hat{z}_k}\sup_{(\{z_k\}, S, \{\phi_k\}) \in P(\gamma)} \mathsf{E}\left[p\ell(\hat{z}_k, z_k)\right] \gtrsim 1.$$

$\square$

C.4 Proof of Theorem 3

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

$$S = S(\gamma) = \begin{bmatrix}1\\1\\1\end{bmatrix}^{\otimes K} + \sigma p^{-\gamma/2}\begin{bmatrix}1\\1\\0\end{bmatrix}^{\otimes K}. \tag{20}$$

Then, we consider the signal tensor family

$$P_{\mathsf{shifted}}(\gamma) = \{X: X = S \times_1 M_1 \times_2 \cdots \times_K M_K, \; M_k \in \{0,1\}^{p \times 3} \text{ is a membership matrix that}$$
$$\text{satisfies } |M_k(:,i)| \asymp p \text{ for all } i \in [3] \text{ and } k \in [K]\}.$$

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

(i) For every $\gamma \in \mathbb{R}$, $P_{\mathsf{shifted}}(\gamma) \subseteq P(\gamma)$, where $P(\gamma)$ is the degree-corrected cluster tensor family (5).

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

18

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

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

$$\liminf_{p \to \infty} \sup_{X \in P_{\text{non-degree}}(\gamma)} \mathsf{P}(\exists k \in [K], \ \hat{\boldsymbol{M}}_k \neq \boldsymbol{M}_k) \geq 1/2, \tag{21}$$

under the HPC Conjecture 1. The inequality (21) implies

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

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

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

We complete the proof by verifying the properties (i)-(ii). For (i), we verify that the angle gap for the core tensor S in (20) is on the order of $\sigma p^{-\gamma/2}$. Specifically, write $\mathbf{1} = (1, 1, 1)$ and $\boldsymbol{e} = (1, -1, 0)$. We have

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

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

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

Therefore, we have shown that $P_{\text{shifted}}(\gamma) = P(\gamma)$. Finally, the property (ii) follows directly by comparing the definition of S in (20) with that in the proof of Han et al. (2020, Theorem 7). $\square$

Useful Definition and Conjecture for Theorem 3

Definition 2 (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 $H_k(p, 1/2)$ denote the Erdős-Rényi K-hypergraph where the edge $(i_1, \ldots, i_K)$ belongs to E with probability $1/2$. Further, we let $H_K(p, 1/2, \kappa)$ denote the hyhpergraph with planted cliques of size κ. Specifically, we generate a hypergraph from $H_k(p, 1/2)$, 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

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

Conjecture 1 (HPC conjecture)**.** Consider the HPC detection problem in Definition 2. Suppose the sequence $\{\kappa_p\}$ such that $\limsup_{p \to \infty} \log \kappa_p / \log \sqrt{p} \leq (1 - \tau)$. Then, for every sequence of polynomial-time test $\{\varphi_p\} : H \mapsto \{0, 1\}$ we have

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

C.5 Proof of Theorem 4

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

19

First, by Lemma 4, there exists a positive constant such that $\min_{z(i)\neq z(j)}\left\|\boldsymbol{X}_{i:}^{s}-\boldsymbol{X}_{j:}^{s}\right\| \geq c_{0}\Delta_{\min}$. By the balance assumption on π and Lemma 7, we have

$$\min_{\pi_2}\sum_{i:\hat{z}(i)\neq\pi(z(i))}\theta(i)^2 \leq \sum_{i\in S_I}\theta(i)^2 + 4\sum_{i\in S}\theta(i)^2, \tag{22}$$

where

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

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

$$\sum_{i\in P}\|\boldsymbol{X}_{i:}\|^2 = \sum_{i\in P}\|\theta(i)\boldsymbol{S}_{z(i):}(\boldsymbol{\Theta}\boldsymbol{M})^{T,(K-1)}\|^2$$

$$\geq \sum_{i\in P}\theta(i)^2 \min_{a\in[r]}\|\boldsymbol{S}_{a:}\|^2\lambda_r^{2(K-1)}(\boldsymbol{\Theta}\boldsymbol{M})$$

$$\geq \& \sum_{i\in P}\theta(i)^2 p^{K-1}r^{(K-1)},$$

where the last inequality follows Lemma 5, the assumption that $\min_{i\in[p]}\theta(i) \geq c$, and the constraint $\min_{a\in[r]}\|\boldsymbol{S}_{a:}\| \geq c_3$ in the parameter space (2). Thus, we have

$$\sum_{i\in P}\theta(i)^2 . \sum_{i\in P}\|\boldsymbol{X}_{i:}\|^2 p^{-(K-1)}r^{-K-1}. \tag{23}$$

On the other hand, note that

$$\sum_{i\in S}\|\boldsymbol{X}_{i:}\|^2 \leq 2\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2 + 2\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}-\boldsymbol{X}_{i:}\|^2 \tag{24}$$

$$\leq \frac{8}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{x}}_{\hat{z}(i)}-\boldsymbol{X}_{i:}^s\|^2 + 2\|\hat{\boldsymbol{X}}-\boldsymbol{X}\|_F^2 \tag{25}$$

$$\leq \frac{16}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\left[\|\hat{\boldsymbol{x}}_{\hat{z}(i)}-\hat{\boldsymbol{X}}_{i:}^s\|^2 + \|\hat{\boldsymbol{X}}_{i:}^s-\boldsymbol{X}_{i:}^s\|^2\right] + 2\|\hat{\boldsymbol{X}}-\boldsymbol{X}\|_F^2 \tag{26}$$

$$\leq \frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}\sum_{i\in S}\|\hat{\boldsymbol{X}}_{i:}\|^2\|\hat{\boldsymbol{X}}_{i:}^s-\boldsymbol{X}_{i:}^s\|^2 + 2\|\hat{\boldsymbol{X}}-\boldsymbol{X}\|_F^2 \tag{27}$$

$$\leq \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}+2\right)\|\hat{\boldsymbol{X}}-\boldsymbol{X}\|_F^2 \tag{28}$$

$$. \left(\frac{16(1+\eta)}{c_0^2\Delta_{\min}^2}+2\right)\left(p^{K/2}r+pr^2+r^K\right)\sigma^2, \tag{29}$$

where inequalities (24) and (26) follow from the triangle inequality, (25) follows from the definition of S, (27) follows from the update rule of k-means in Step 5 of Sub-algorithm 1, (28) follows from Lemma 2, and the last inequality (29) follows from Lemma 6. Also, note that

$$\sum_{i\in S_0}\|\boldsymbol{X}_{i:}\|^2 = \sum_{i\in S_0}\|\hat{\boldsymbol{X}}_{i:}-\boldsymbol{X}_{i:}\|^2 \leq \|\hat{\boldsymbol{X}}-\boldsymbol{X}\|_F^2 . \left(p^{K/2}r+pr^2+r^K\right)\sigma^2, \tag{30}$$

where the equation follows from the definition of S_0. Therefore, combining the inequalities (22), (23), (29), and (30), we have

$$\min_{\pi_2}\sum_{i:\hat{z}(i)\neq\pi(z(i))}\theta(i)^2 . \left(\sum_{i\in S}\|\boldsymbol{X}_{i:}\|^2 + \sum_{i\in S_0}\|\boldsymbol{X}_{i:}\|^2\right)p^{-(K-1)}r^{-K-1}$$

$$. \frac{\sigma^2 r^{K-1}}{\Delta_{\min}^2 p^{K-1}}\left(p^{K/2}r+pr^2+r^K\right).$$

20

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

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

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

Useful Corollary of Theorem 4

Corollary 1 (Initial misclustering loss). Under the setup of Theorem 4, assume $\text{SNR} \geq \tilde{C} p^{K/2} \log p$ for some positive universal constant $\tilde{C}$ and $\sigma^2 = 1$. With probability going to 1, the misclustering loss for the initialization is upper bounded as

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

where $\pi^{(0)}$ minimizes the initial misclustering error; i.e., $\pi^{(0)} = \arg\min_{\pi \in \mathcal{Z}} \sum_{i \in [p]} \mathbb{1}\left\{z^{(0)}(i) \neq \pi \circ z(i)\right\}$, and $\bar{C} > 1$ is a positive universal constant.

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

Notice that

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

Therefore, when p is large enough, we have

$$\sum_{i \in [p]} \|\boldsymbol{X}_{i:}\|^2 \|\hat{\boldsymbol{X}}_i^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2 \lesssim \sum_{i \in [p]} \left(\|\boldsymbol{X}_{i:}\|^2 - \|\boldsymbol{X}_{i:} - \hat{\boldsymbol{X}}_{i:}\|^2\right) \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2$$

$$\lesssim \sum_{i \in [p]} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2$$

$$\lesssim \eta \sum_{i \in [p]} \|\hat{\boldsymbol{X}}_{i:}\|^2 \|\hat{\boldsymbol{X}}_{i:}^s - \boldsymbol{X}_{i:}^s\|^2$$

$$\lesssim \|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F^2$$

$$\lesssim p^{K/2}r + pr^2 + r^K. \tag{31}$$

Hence, we have

$$\sum_{i \in [p]} \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2 \lesssim \sum_{i \in [p]} \theta(i)^2 \|\hat{\boldsymbol{X}}_i^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2$$

$$\lesssim \frac{r^{K-1}}{p^{K-1}} \sum_{i \in [p]} \|\boldsymbol{X}_{i:}\|^2 \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2$$

$$\lesssim \frac{r^{K-1}}{p^{K-1}} \left(p^{K/2}r + pr^2 + r^K\right), \tag{32}$$

where the first inequality follows from the assumption $\min_{i \in [p]} \theta(i) \geq c$, the second inequality follows from the inequality (23), and the last inequality comes from the inequality (31).

Next, we consider the following quantity,

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

21

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

$$\cdot \frac{r^{K-1}}{p^{K-1}} \left(p^{K/2} r + pr^2 + r^K \right), \tag{33}$$

where the first inequality follows from the assumption of $\theta(i)$ and triangle inequality, the second inequality follows from Lemma 2, and the last inequality follows from (32). In addition, with Theorem 4 and the condition SNR $\gtrsim p^{-K/2} \log p$, for all $a \in [r]$, we have

$$|z^{-1}(a) \setminus (z^{(0)})^{-1}(a)| \leq |z^{-1}(a)| \cdot p\ell(z^{(0)}, z) \lesssim \frac{p}{r} \cdot \frac{p}{\log p} \lesssim \frac{p}{r},$$

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

$$\|\hat{\boldsymbol{x}}_a - \boldsymbol{X}_a^s\|^2 = \frac{\sum_{i \in z^{-1}(a) \setminus (z^{(0)})^{-1}(a)} \left\| \boldsymbol{X}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s \right\|^2}{|z^{-1}(a) \setminus (z^{(0)})^{-1}(a)|}$$

$$\cdot \frac{r}{p} \left(\sum_{i \in [p]} \|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{X}}_{i:}^s\|^2 + \sum_{i \in [p]} \|\hat{\boldsymbol{X}}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2 \right)$$

$$\cdot \frac{r^K}{p^K} \left(p^{K/2} r + pr^2 + r^K \right), \tag{34}$$

where the last inequality follows from the inequality (32).

Finally, we obtain

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

$$\cdot \frac{1}{p} \sum_{i \in [p], z^{(0)}(i) \neq z(i)} \theta(i) \|\boldsymbol{X}_{i:}^s - \boldsymbol{X}_{z^{(0)}(i)}^s\|^2$$

$$\cdot \frac{1}{p} \sum_{i \in [p], z^{(0)}(i) \neq z(i)} \theta(i) \left(\|\boldsymbol{X}_{i:}^s - \hat{\boldsymbol{x}}_{z^{(0)}(i)}^s\|^2 + \|\hat{\boldsymbol{x}}_{z^{(0)}(i)}^s - \boldsymbol{X}_{z^{(0)}(i)}^s\|^2 \right)$$

$$\bar{C} \frac{r^K}{p^K} \left(p^{K/2} r + pr^2 + r^K \right),$$

$$\frac{\bar{C} \Delta_{\min}^2}{\tilde{C} r \log p}$$

where the first inequality follows from Lemma 4, the third inequality follows from inequalities (33) and (34), and the last inequality follows from the assumption that SNR $\geq \tilde{C} p^{-K/2} \log p$. □

Useful Definitions and Lemmas for the Proof of Theorem 4

Lemma 2 (Basic inequality). For any two nonzero vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ of same dimension, we have

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

Proof of Lemma 2. For the first inequality, let $\alpha \in [0, \pi]$ denote the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. We have

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

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

22

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

$$\|\mathbf{v}_1^s - \mathbf{v}_2^s\| = \left\| \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} - \frac{\mathbf{v}_2}{\|\mathbf{v}_1\|} + \frac{\mathbf{v}_2}{\|\mathbf{v}_1\|} - \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} \right\|$$

$$\le \frac{\|\mathbf{v}_1 - \mathbf{v}_2\|}{\|\mathbf{v}_1\|} + \frac{\|\mathbf{v}_2\| - \|\mathbf{v}_1\|}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \|\mathbf{v}_2\|$$

$$\le \frac{2 \|\mathbf{v}_1 - \mathbf{v}_2\|}{\|\mathbf{v}_2\|}.$$

Therefore, Lemma 2 is proved. $\qquad\qquad\square$

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

$$\mathrm{Pad}_W(\mathbf{a}) = [a_1 \mathbf{w}_1, \ldots, a_d \mathbf{w}_d]^T, \quad \text{where} \quad a_i \mathbf{w}_i = [a_i w_{i1}, \ldots, a_i w_{ip_i}]^T, \text{ for all } i \in [d]. \quad (35)$$

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

$$\min_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2 \le \|\mathrm{Pad}_W(\mathbf{a})\|^2 \le \max_{i \in [d]} \|\mathbf{w}_i\|^2 \|\mathbf{a}\|^2. \quad (36)$$

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

$$\mathrm{Pad}_W^{-1}(\mathrm{Pad}_W(\mathbf{a})) = \mathbf{a}.$$

Lemma 3 (Angle for weighted padding vectors)**.** Suppose we have two non-zero vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Given the weight collection $\mathbf{w}$, we have

$$\frac{\min_{i \in [d]} \|\mathbf{w}_i\|}{\max_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}) \overset{*}{\le} \sin(\mathrm{Pad}_W(\mathbf{a}), \mathrm{Pad}_W(\mathbf{b})) \overset{**}{\le} \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}). \quad (37)$$

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

First, we prove the inequality ** in (37). Decomposing $\mathbf{b}$ yields

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

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

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

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

$$\sin(\mathrm{Pad}_W(\mathbf{a}), \mathrm{Pad}_W(\mathbf{b})) \le \frac{\|\mathbf{b}\| \|\mathrm{Pad}_W(\mathbf{a}^\perp)\|}{\|\mathbf{a}^\perp\| \|\mathrm{Pad}_W(\mathbf{b})\|} \sin(\mathbf{a}, \mathbf{b})$$

$$\le \frac{\max_{i \in [d]} \|\mathbf{w}_i\|}{\min_{i \in [d]} \|\mathbf{w}_i\|} \sin(\mathbf{a}, \mathbf{b}),$$

where the second inequality follows by applying the property (36) to vectors $\mathbf{b}$ and $\mathbf{a}^\perp$.

Next, we prove inequality * in (37). With the decomposition of $\mathrm{Pad}_W(\mathbf{b})$ and the inverse weighted padding operator, we have

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

23

Therefore, we obtain

$$\sin(\boldsymbol{a}, \boldsymbol{b}) \leq \frac{\|\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{b})\| \|\mathrm{Pad}_{\mathcal{W}}^{-1}((\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{a}))^\perp)\|}{\|(\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{a}))^\perp\| \|\boldsymbol{b}\|} \sin(\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{a}), \mathrm{Pad}_{\mathcal{W}}(\boldsymbol{b}))$$

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

where the second inequality follows by applying the property (36) to vectors $\boldsymbol{b}$ and $\mathrm{Pad}_{\mathcal{W}}^{-1}((\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{a}))^\perp)$. $\quad\square$

Lemma 4 (Angle gap in X)**.** Consider the dTBM model (1) under the parameter space P in (2). Suppose Assumption 1 holds and is balanced satisfying (6). Then the angle gap in X is asymptotically lower bounded by the angle gap in S; i.e., for all i, j such that $z(i) \neq z(j)$

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

Proof of Lemma 4. Note that the vector $\boldsymbol{S}_{z(i):}$ can be folded to a tensor $\mathcal{S}^\theta = \llbracket S_{a_2,\ldots,a_K}^\theta \rrbracket \in \mathbb{R}^{r^{K-1}}$; i.e., $\mathrm{vec}(\mathcal{S}^\theta) = \boldsymbol{S}_{z(i):}$. Define weight vectors $\boldsymbol{w}_{a_2,\ldots,a_K}$ correspond to the elements in $S_{a_2,\ldots,a_K}^\theta$ by

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

for all $a_k \in [r], k = 2, \ldots, K$, where $\otimes$ denotes the Kronecker product. Therefore, we have $\boldsymbol{X}_{i:} = \theta(i)\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{S}_{z(i):})$ where $\mathcal{W} = \{\boldsymbol{w}_{a_2,\ldots,a_K}\}_{a_k \in [r], k \in [K]/\{1\}}$. Specifically, we have $\|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 = \prod_{k=2}^K \|\boldsymbol{\theta}_{z^{-1}(a_k)}\|^2$, and by the balanced assumption (6) we have

$$\max_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 = (1 + o(1)) \min_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2. \tag{38}$$

Consider the inner product of $\boldsymbol{X}_{i:}$ and $\boldsymbol{X}_{j:}$ for $z(i) \neq z(j)$. By the definition of weighted padding operator (35) and the balanced assumption (38), we have

$$\langle \boldsymbol{X}_{i:}, \boldsymbol{X}_{j:} \rangle = \theta(i)\theta(j) \langle \mathrm{Pad}_{\mathcal{W}}(\boldsymbol{S}_{z(i):}), \mathrm{Pad}_{\mathcal{W}}(\boldsymbol{S}_{z(j):}) \rangle$$

$$= \theta(i)\theta(j) \min_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|^2 \langle \boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):} \rangle (1 + o(1)).$$

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

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

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

$$\|\boldsymbol{S}_{z(i):}^s - \boldsymbol{S}_{z(j):}^s\| = \sqrt{1 - \cos(\boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):})}$$

$$\leq \sin(\boldsymbol{S}_{z(i):}, \boldsymbol{S}_{z(j):})$$

$$\leq \frac{\max_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|}{\min_{(a_2,\ldots,a_K)} \|\boldsymbol{w}_{a_2,\ldots,a_K}\|} \sin(\mathrm{Pad}_{\mathcal{W}}(\boldsymbol{S}_{z(i):}), \mathrm{Pad}_{\mathcal{W}}(\boldsymbol{S}_{z(j):}))$$

$$\leq (1 + o(1))\|\boldsymbol{X}_{i:}^s - \boldsymbol{X}_{j:}^s\|,$$

where the second inequality follows from Lemma 3, and the last inequality follows from the balanced weight (38) and Lemma 2.

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

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

$\quad\square$

24

Lemma 5 (Singular value of weighted membership matrix). Under the parameter space (2) and assumption that $\min_{i \in [p]} \theta(i) \geq c$, the singular values of ΘM are bounded as

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

Proof of Lemma 5. Note that

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

By the definition of singular values, we have

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

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

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

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

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

Therefore, we complete the proof of Lemma 5. $\qquad\square$

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

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

where C, c are some positive constants.

Proof of Lemma 6. See Han et al. (2020, Proposition 1). $\qquad\square$

Lemma 7 (Upper bound of misclustering error). Let $z : [p] \to [r]$ be a cluster assignment such that $|z^{-1}(a)| \asymp p/r$ for all $a \in [r]$. Let node i correspond to a vector $x_i = \theta(i) v_{z(i)} \in \mathbb{R}^d$, where $\{v_a\}_{a=1}^r$ are the cluster centers and $\Theta = \mathrm{diag}(\theta(i)) \in \mathbb{R}_+^p$ is the positive degree heterogeneity. Assume that Θ satisfies the balanced assumption (6) such that $\frac{\max_{a\in[r]}\|\theta_{z^{-1}(a)}\|^2}{\min_{a\in[r]}\|\theta_{z^{-1}(a)}\|^2} = 1 + o(1)$. Consider an arbitrary estimate $\hat{z}$ with $\hat{x}_i = \hat{v}_{\hat{z}(i)}$ for all $i \in S$. Then, if

$$\min_{a \neq b \in [r]} \|v_a - v_b\| \geq 2c, \tag{39}$$

for some constant $c > 0$, we have

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

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

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

25

Proof of Lemma 7. For each cluster $u \in [r]$, we use C_u to collect the subset of points for which the estimated and true positions $\hat{\boldsymbol{x}}_i, \boldsymbol{x}_i$ are within distance c. Specifically, define

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

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

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

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

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

which contradicts to the assumption (39). Hence, the estimates $\hat{z}(i) \neq \hat{z}(j)$ for the nodes $i \in C_u$ and $j \in C_v$ with $u \neq v$. By the definition of R_2, the nodes in $\bigcup_{u \in R_2} C_u$ have the same assignment with z and $\hat{z}$. Then, we have

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

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

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

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

C.6 Proof of Theorem 5

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

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

Notation.

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

26

2. We define normalized membership matrices

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

and the dual normalized membership matrices

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

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

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

4. We define the matricizations of tensors

$$S = \operatorname{Mat}(S), \quad S^{(t)} = \operatorname{Mat}(S^{(t)}), \quad \tilde{S} = \operatorname{Mat}(\tilde{S})$$
$$Y = \operatorname{Mat}(Y), \quad X = \operatorname{Mat}(X), \quad E = \operatorname{Mat}(E).$$

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

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

and the oracle loss

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

where m is a positive universal constant defined in (47).

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

$$\|EV\|_\sigma \lesssim \sqrt{\frac{r^{K-1}}{p^{K-1}}}\left(p^{1/2} + r^{(K-1)/2}\right), \quad \|EV\|_F \lesssim \sqrt{\frac{r^{2(K-1)}}{p^{K-2}}}, \quad \|W_{a:}^T EV\| \lesssim \frac{r^K}{p^{K/2}} \text{ for all } a\in[r], \quad (40)$$

$$\sup_{U_k\in O_{p,r},k=2,\dots,K}\|E(U_2\otimes\cdots\otimes U_K)\|_\sigma \lesssim \left({}^p\!\sqrt{r^{K-1}} + K^p\sqrt{pr}\right), \quad (41)$$

$$\sup_{U_k\in O_{p,r},k=2,\dots,K}\|E(U_2\otimes\cdots\otimes U_K)\|_F \lesssim \left(\sqrt{pr^{K-1}} + K^p\sqrt{pr}\right), \quad (42)$$

$$\xi \lesssim \exp\left(-M\frac{\Delta_{\min}^2 p^{K-1}}{r^{K-1}}\right), \quad (43)$$

$$L^{(t)} \lesssim \frac{\bar{C}}{\tilde{C}}\frac{\Delta_{\min}^2}{r\log p}, \quad \text{for} \quad t=0,1,\dots,T, \quad (44)$$

where M is a positive universal constant in inequality (59), $\bar{C}, \tilde{C}$ are universal constants in Corollary 1 and assumption $\mathrm{SNR} \gtrsim \tilde{C}p^{-K/2}\log p$, respectively. Further, inequality (40) holds by replacing V to $V^{(t)}$ and $W_{:a}$ to $W_{:a}^{(t),T}$ when initialization condition (44) holds.

27

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

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

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

$$\|[\boldsymbol{Y}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 \leq \|[\boldsymbol{Y}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2. \tag{45}$$

Note that the event (45) also holds for the degenerate entity i with $\|\boldsymbol{Y}_{i:} \boldsymbol{V}^{(t)}\| = 0$ due to the convention that $\boldsymbol{a}^s = \boldsymbol{0}$ if $\boldsymbol{a} = \boldsymbol{0}$. Arranging the terms in (45) yields the decomposition

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

where

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

$$G_{ib}^{(t)} = \left(\|[\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{z(i):}^{(t)}]^s\|^2 - \|[\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y} \boldsymbol{V}^{(t)}]^s\|^2 \right)$$

$$\left(\|[\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{S}_{b:}^{(t)}]^s\|^2 - \|[\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^T \boldsymbol{Y} \boldsymbol{V}^{(t)}]^s\|^2 \right),$$

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

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

$$\mathbf{1}\left\{ z^{(t+1)}(i) = b \right\} \leq \mathbf{1}\left\{ z^{(t+1)}(i) = b, \left\langle \boldsymbol{E}_{j:} \boldsymbol{V}, [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s \right\rangle \geq \frac{1}{4} \|\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}\| \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2 \right\}$$

$$+ \mathbf{1}\left\{ z^{(t+1)}(i) = b, \frac{1}{2} \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|^2 \leq \|\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}\|^{-1} F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\}. \tag{46}$$

Note that

$$\|\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}\| = \theta(i) \|\boldsymbol{S}_{i:} (\boldsymbol{\Theta} \boldsymbol{M})^{(K-1),T} \boldsymbol{W}^{(t), \times K - 1}\| \geq \theta(i) \|\boldsymbol{S}_{z(i):}\| \lambda_r^{K-1}(\boldsymbol{\Theta} \boldsymbol{M}) \lambda_r^{K-1}(\boldsymbol{W}^{(t)}) \geq \theta(i) m, \tag{47}$$

where the first inequality follows from the property of eigenvalues; the last inequality follows from Lemma 5, Lemma 9, and assumption that $\min_{a \in [r]} \|\boldsymbol{S}_{z(i):}\| \geq c_3 > 0$; and $m > 0$ is a positive constant related to c_3. Plugging the lower bound of $\|\boldsymbol{X}_{i:} \boldsymbol{V}^{(t)}\|$ (47) into the inequality (46) gives

$$\mathbf{1}\left\{ z^{(t+1)}(i) = b \right\} \leq A_{ib} + B_{ib}, \tag{48}$$

where

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

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

Taking the weighted summation of (48) over $i \in [p]$ yields

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

28

where ξ is the oracle loss such that

$$\xi = \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} A_{ib} k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2. \tag{49}$$

Similarly to ξ in (49), we define

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

Step 3 (Derivation of contraction inequality). In this step we derive the upper bound of ζ_{ib} and obtain the contraction inequality. Choose the constant $\tilde{C}$ in the condition SNR $\quad \tilde{C} p^{-K/2} \log p$ that satisfies the condition of Lemma 10, inequalities (73) and (77). Note that

$$\zeta_{ib}^{(t)} = \theta(i) k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 1 \left\{ z^{(t+1)}(i) = b, \frac{1}{2} k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 \quad (\theta(i)m)^{-1} F_{ib}^{(t)} + G_{ib}^{(t)} + H_{ib}^{(t)} \right\}$$

$$\theta(i) k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 1 \left\{ z^{(t+1)}(i) = b, \frac{1}{4} k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 \quad (\theta(i)m)^{-1} F_{ib}^{(t)} + G_{ib}^{(t)} \right\}$$

$$641 \left\{ z^{(t+1)}(i) = b \right\} \left(\frac{(F_{ib}^{(t)})^2}{cm^2 k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} + \frac{\theta(i)(G_{ib}^{(t)})^2}{k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \right)$$

where the first inequality follows from the inequality (64) in Lemma 10, and the last inequality follows from the assumption that $\min_{i \in [p]} \theta(i) \quad c > 0$. Following Han et al. (2020, Step 4, Proof of Theorem 2) and Lemma 10, we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} 1 \left\{ z^{(t+1)}(i) = b \right\} \frac{(F_{ib}^{(t)})^2}{cm^2 k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \quad \frac{C_0 \bar{C}}{cm^2 \tilde{C}^2} L^{(t)},$$

for a positive universal constant C and

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} 1 \left\{ z^{(t+1)}(i) = b \right\} \frac{\theta(i)(G_{ib}^{(t)})^2}{k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \quad \frac{1}{512} \frac{1}{p} \sum_{i \in [p]} \theta(i) \sum_{b \in [r]/z(i)} 1 \left\{ z^{(t+1)}(i) = b \right\} (\Delta_{\min}^2 + L^{(t)})$$

$$\frac{1}{512} (L^{(t+1)} + L^{(t)}),$$

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

$$\frac{C_0 \bar{C}}{cm^2 \tilde{C}^2} \quad \frac{1}{512}, \tag{50}$$

we have

$$\frac{1}{p} \sum_{i \in [p]} \sum_{b \in [r]/z(i)} \zeta_{ib}^{(t)} \quad \frac{1}{8} L^{(t+1)} + \frac{1}{4} L^{(t)}. \tag{51}$$

Plugging the inequality (51) into the decomposition (49), we obtain the contraction inequality

$$L^{(t+1)} \quad \frac{3}{2} \xi + \frac{1}{2} L^{(t)}, \tag{52}$$

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

Therefore, with $\tilde{C}$ satisfying inequalities (50), (73) and (77), we obtain the conclusion in Theorem 5 via inequality (52) combining the inequality (43) in Condition 1 and Lemma 8.

Step 4 (Verification of Condition 1). Last, we verify the Condition 1 under high probability to finish the proof. Note that the inequalities (40), (41), and (42) describe the property of the sub-Gaussian noise tensor

29

E, and the readers can find the proof directly in Han et al. (2020, Step 5, Proof of Theorem 2). Here, we include only the verification of inequalities (43) and (44).

Now, we verify the oracle loss condition (43). Recall the definition of ξ,

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

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

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

where

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

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

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

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

$$P_1 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}} \| [\boldsymbol{S}_{z(j):}]^s - [\boldsymbol{S}_{b:}]^s \|^2 \right). \tag{53}$$

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

$$P_2 \leq P_{21} + P_{22}, \tag{54}$$

where with some positive constant $t > 0$,

$$P_{21} = \mathsf{P}\left(t \leq \| [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\boldsymbol{S}_{z(i):}]^s \| \right),$$

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

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

$$P_{21} \leq \mathsf{P}\left(t \| \boldsymbol{S}_{z(i):} \| \leq \| \tilde{\boldsymbol{S}}_{z(i):} - \boldsymbol{S}_{z(i):} \| \right) \leq \mathsf{P}\left(c_3 t \leq \| \boldsymbol{W}^T_{:z(i)} \boldsymbol{E} \boldsymbol{V} \| \right) \lesssim \exp\left(-\frac{p^K t^2}{r^K} \right), \tag{55}$$

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

30

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

$$P_{22} \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}t^2}\|[\boldsymbol{S}_{z(j):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^4\right). \tag{56}$$

We take $t = \|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa$ in P_{21} and P_{22}, and plug the inequalities (55) and (56) into to the upper bound for P_2 in (54). We obtain that

$$P_2 \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\right). \tag{57}$$

Combining the upper bounds (53) and (57) gives

$$P\left(\left\langle e_i, [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s \right\rangle \geq \frac{\theta(i)m}{4}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\right) \lesssim \exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\right). \tag{58}$$

Hence, we have

$$
\begin{aligned}
\mathsf{E}\xi &= \frac{1}{p}\sum_{i\in[p]}\theta(i)\sum_{b\in[r]}P\left\{\left\langle \boldsymbol{E}_{i:}\boldsymbol{V}, [\tilde{\boldsymbol{S}}_{z(i):}]^s - [\tilde{\boldsymbol{S}}_{b:}]^s \right\rangle \geq \frac{\theta(i)m}{4}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\right\}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\\
&\lesssim \frac{1}{p}\sum_{i\in[p]}\theta(i)\max_{i\in[p],b\in[r]}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\exp\left(-\frac{p^{K-1}}{r^{K-1}}\|[\boldsymbol{S}_{z(i):}]^s - [\boldsymbol{S}_{b:}]^s\|_\kappa^2\right)\\
&\leq \exp\left(-M\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right),
\end{aligned}
\tag{59}
$$

where M is a positive constant, the first inequality follows from the constraint that $\sum_{i\in[p]}\theta(i) = p$, and the last inequality follows from (58).

By Markov's inequality, we have

$$P\left(\xi \geq \mathsf{E}\xi + \exp\left(-M\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right)\right) \leq 1 - C\exp\left(-M\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right),$$

and thus the condition (43) holds with probability at least $1 - C\exp\left(-M\frac{p^{K-1}}{r^{K-1}}\Delta_{\min}^2\right)$ for some constant $C > 0$.

Finally, we verify the bounded loss condition (44) by induction. With output $z^{(0)}$ from Sub-algorithm 2 and the assumption SNR $\geq \tilde{C}p^{-K/2}\log p$, by Corollary 1, we have

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

Therefore, the condition (44) holds for $t = 0$. Assume the condition (44) also holds for all $t \leq t_0$. Then, by the decomposition (52), we have

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

where the second inequality follows from the condition (43) and the last inequality follows from the assumption that $\Delta_{\min}^2 \gtrsim p^{-K/2}\log p$. Thus, the condition (44) holds for $t_0 + 1$, and the condition (44) is proved by induction. $\qquad\square$

31

Useful Definitions and Lemmas for the Proof of Theorem 5

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

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

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

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

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

Lemma 9 (Singular-value property of membership matrices). Under the setup of Theorem 5, suppose the condition (44) holds. Then, for all $a \in [r]$, we have $j(z^{(t)})^{-1}(a)j \asymp p/r$. Moreover, we have

$$\lambda_r(\boldsymbol{M}) \quad k\boldsymbol{M}k_\sigma \asymp \sqrt{p/r}, \quad \lambda_r(\boldsymbol{W}) \quad k\boldsymbol{W}k_\sigma \asymp \sqrt{r/p}. \tag{60}$$

The inequalities (60) also hold by replacing $\boldsymbol{M}$ and $\boldsymbol{W}$ to $\boldsymbol{M}^{(t)}$ and $\boldsymbol{W}^{(t)}$ respectively. Further, we have

$$\lambda_r(\boldsymbol{W}\boldsymbol{W}^T) \quad \left\| \boldsymbol{W}\boldsymbol{W}^T \right\|_\sigma \asymp r/p, \tag{61}$$

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

Proof of Lemma 9. The proof for the inequality (60) can be found in Han et al. (2020, Proof of Lemma 4) For inequality (61), note that for all $k \in [r]$,

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

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

Lemma 10 (Upper bound for $F_{ib}^{(t)}, G_{ib}^{(t)}$ and $H_{ib}^{(t)}$). Under the Condition 1 and the setup of Theorem 5, assume the the constant $\tilde{C}$ in the condition SNR $\geq \tilde{C} p^{-K/2} \log p$ is large enough to satisfy the inequalities (73) and (77). We have

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(F_{ib}^{(t)}\right)^2}{k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \lesssim \frac{rL^{(t)}}{\Delta_{\min}^2} k\boldsymbol{E}_{i:}\boldsymbol{V}k^2 + \left(1 + \frac{rL^{(t)}}{\Delta_{\min}^2}\right) k\boldsymbol{E}_{i:}(\boldsymbol{V} \quad \boldsymbol{V}^{(t)})k^2, \tag{62}$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{\left(G_{ib}^{(t)}\right)^2}{k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \leq \frac{1}{512}\left(\Delta_{\min}^2 + L^{(t)}\right), \tag{63}$$

$$\max_{i \in [p]} \max_{b \neq z(i)} \frac{\left|H_{ib}^{(t)}\right|}{k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2} \leq \frac{1}{4}. \tag{64}$$

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

32

1. Upper bound for $F_{ib}^{(t)}$, i.e., inequality (62). Recall the definition of $F_{ib}^{(t)}$,

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

By Cauchy-Schwartz inequality, we have

$$\left(F_{ib}^{(t)}\right)^2 \quad 8\left(\left\langle \boldsymbol{E}_{i:}\boldsymbol{V}^{(t)}, \left([\tilde{\boldsymbol{S}}_{z(i):}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s\right) \quad \left([\tilde{\boldsymbol{S}}_{b:}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s\right)\right\rangle\right)^2$$

$$+ 8\left(\left\langle \boldsymbol{E}_{i:}(\boldsymbol{V} \quad \boldsymbol{V}^{(t)}), [\tilde{\boldsymbol{S}}_{z(i):}]^s \quad [\tilde{\boldsymbol{S}}_{b:}]^s\right\rangle\right)^2$$

$$8\left(k\boldsymbol{E}_{i:}\boldsymbol{V}k^2 + k\boldsymbol{E}_{i:}(\boldsymbol{V} \quad \boldsymbol{V}^{(t)})k^2\right) \max_{a\,2[r]} k[\tilde{\boldsymbol{S}}_{a:}]^s \quad [\boldsymbol{S}_{a:}^{(t)}]^s k$$

$$+ k\boldsymbol{E}_{i:}(\boldsymbol{V} \quad \boldsymbol{V}^{(t)})k^2 k[\tilde{\boldsymbol{S}}_{z(i):}]^s \quad [\tilde{\boldsymbol{S}}_{b:}]^s k. \tag{65}$$

Note that for all $a\,2\,[r]$,

$$k[\tilde{\boldsymbol{S}}_{a:}]^s \quad [\boldsymbol{S}_{a:}^{(t)}]^s k^2 = k[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2$$

$$2k[\boldsymbol{W}_{:a}^T\boldsymbol{Y}\boldsymbol{V}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s k^2 + 2k[\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2$$

$$\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} + \frac{rr^{2K} + pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}$$

$$rL^{(t)} + \frac{rr^{2K} + pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}$$

$$rL^{(t)}, \tag{66}$$

where the second inequality follows from the inequalities (80) and (81) in Lemma 11, the third inequality follows from the condition (44) in Condition 1, and the last inequality follows from the assumption that $\Delta_{\min}^2 \quad \tilde{C}p^{-K/2}\log p$.

Note that

$$k[\tilde{\boldsymbol{S}}_{z(i):}]^s \quad [\tilde{\boldsymbol{S}}_{b:}]^s k^2 = k[\tilde{\boldsymbol{S}}_{z(i):}]^s \quad [\boldsymbol{S}_{z(i):}]^s + [\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s + [\boldsymbol{S}_{b:}]^s \quad [\tilde{\boldsymbol{S}}_{b:}]^s k^2$$

$$k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 + \max_{a\,2[r]} k[\boldsymbol{S}_{a:}]^s \quad [\tilde{\boldsymbol{S}}_{a:}]^s k^2$$

$$k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2 + \max_{a\,2[r]}\frac{1}{k\boldsymbol{S}_{a:}k^2} k\boldsymbol{W}_{:a}^T\boldsymbol{E}\boldsymbol{V}k^2$$

$$k[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s k^2, \tag{67}$$

where the second inequality follows from Lemma 2, and the last inequality follows from the assumptions on $k\boldsymbol{S}_{a:}k$ in the parameter space (2), the inequality (40) in Condition 1 and the assumption $\Delta_{\min}^2 \& p^{-K/2}\log p$.

Therefore, we finish the proof of inequality (62) by plugging the inequalities (66) and (67) into the upper bound (65).

2. Upper bound for $G_{ib}^{(t)}$, i.e., inequality (63). By definition of $G_{ib}^{(t)}$, we rearrange terms and obtain

$$G_{ib}^{(t)} = \left(k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s k^2 \quad k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2\right)$$

$$\left(k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s k^2 \quad k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2\right)$$

$$= 2\left\langle [\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s, \left([\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s\right) \quad \left([\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s\right)\right\rangle$$

$$= G_1 + G_2 \quad G_3, \tag{68}$$

where

$$G_1 = k[\boldsymbol{W}_{:z(i)}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s k^2 \quad k[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s k^2,$$

33

$$G_2 = 2\left\langle [\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s, [\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s\right\rangle,$$

$$G_3 = 2\left\langle [\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s, [\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s\right\rangle.$$

For G_1, we have

$$jG_1 j^2 \quad \left| k[\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s k^2 \quad k[\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{b:}^{(t)}]^s k^2 \right|^2$$

$$\max_{a2[r]} k[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^4$$

$$C^4 \frac{r^4}{\Delta_{\min}^4}(L^{(t)})^4 + \frac{r^2 r^{4K} + p^2 r^{2K+4}}{p^{2K}}\frac{(L^{(t)})^2}{\Delta_{\min}^4}$$

$$C^4 \frac{\bar{C}}{\tilde{C}^3}\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right), \tag{69}$$

where the third inequality follows from the inequality (82) in Lemma 9 and the last inequality follows from the assumption that $\Delta_{\min}^2 \quad \tilde{C}p^{-K/2}\log p$ and inequality (44) in Condition 1.

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

$$jG_2 j^2 \quad 2k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2 k[\boldsymbol{W}_{:z(i)}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{S}_{z(i):}^{(t)}]^s k^2$$

$$\frac{2}{k\boldsymbol{W}_{z(i):}^T \boldsymbol{X}\boldsymbol{V}^{(t)} k^2}\max_{a2[r]} k\boldsymbol{W}_{:a}^T \boldsymbol{E}\boldsymbol{V}^{(t)} k^2 \max_{a2[r]} k[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2$$

$$C^0 \frac{r^{2K}\ {}^1 + Kpr^{K+1}}{p^K}\left(\frac{r^2}{\Delta_{\min}^2}(L^{(t)})^2 + \frac{rr^{2K} + pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\right)$$

$$\frac{C^0}{\tilde{C}^2}\Delta_{\min}^2 L^{(t)}, \tag{70}$$

where C^0 is a positive universal constant, the second inequality follows from Lemma 2, the third inequality follows from the inequality (41) in Condition 1, the inequalities (82) and (101) in the proof of Lemma 11, and the last inequality follows from the assumption $\Delta_{\min}^2 \quad \tilde{C}p^{-K/2}\log p$ and inequality (44) in Condition 1.

For G_3, note that by triangle inequality

$$k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s k^2 \quad k\boldsymbol{S}_{z(i):}^s \quad \boldsymbol{S}_{b:}^s k^2 + 2\max_{a2[r]} k[\boldsymbol{W}_{:a}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^T \boldsymbol{X}\boldsymbol{V}]^s k^2$$

$$k\boldsymbol{S}_{z(i):}^s \quad \boldsymbol{S}_{b:}^s k^2 + C\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2}, \tag{71}$$

where the last inequality follows from the inequality (100) in the proof of Lemma 11 and C is a positive constant. Then we have

$$jG_3 j^2 \quad 2k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2 \max_{a2[r]} k[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s k^2$$

$$2\left(k[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s k^2 + k[\boldsymbol{W}_{:b}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T \boldsymbol{X}\boldsymbol{V}^{(t)}]^s k^2\right)$$

$$\max_{a2[r]}\left\|[\boldsymbol{W}_{:a}^T \boldsymbol{Y}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\right\|^2$$

$$C^2\left(k\boldsymbol{S}_{z(i):}^s \quad \boldsymbol{S}_{b:}^s k^2 + C\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2}\right)\left(\frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} + \frac{rr^{2K} + pr^{K+2}}{p^K}\frac{L^{(t)}}{\Delta_{\min}^2}\right) + \frac{C^0}{\tilde{C}^2}\Delta_{\min}^2 L^{(t)}$$

$$\frac{C^2\bar{C}^2}{\tilde{C}}k\boldsymbol{S}_{z(i):}^s \quad \boldsymbol{S}_{b:}^s k^2(\Delta_{\min}^2 + L^{(t)}) + \frac{C^3 C^0 \bar{C}^2}{\tilde{C}^2}\left(\Delta_{\min}^4 + \Delta_{\min}^2 L^{(t)}\right), \tag{72}$$

where the third inequality follows from the same procedure to derive (69) and (70), and the last inequality follows from the assumption $\Delta_{\min}^2 \quad \tilde{C}p^{-K/2}\log p$ and inequality (44) in Condition 1.

34

Choose the $\tilde{C}$ such that

$$3\left(C^4\frac{\bar{C}}{\tilde{C}^3} + \frac{C^\theta}{\tilde{C}^2} + \frac{C^2\bar{C}^2}{\tilde{C}} + \frac{C^3C^\theta\bar{C}^2}{\tilde{C}^2}\right) \leq \frac{1}{512}. \tag{73}$$

Then, we finish the proof of inequality (63) by plugging the inequalities (69), (70), and (72) into the upper bound (68).

3. Upper bound for $H_{ib}^{(t)}$, i.e., the inequality (64). By definition of H_{ib}, we rearrange terms and obtain

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

where

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

For H_1, we have

$$|H_1| \leq \frac{4\max_{a\in[r]}\|\boldsymbol{W}_{:a}^T\boldsymbol{E}\boldsymbol{V}^{(t)}\|^2}{\|\boldsymbol{W}_{z(i):}^T\boldsymbol{X}\boldsymbol{V}^{(t)}\|^2} \quad \frac{r^{2K\ 1} + Kpr^{K+1}}{p^K} \quad \tilde{C}^{\ 2}\|[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{b:}]^s\|^2, \tag{74}$$

following the derivation of G_2 in inequality (70) and the assumption that $\Delta_{\min}^2 \geq \tilde{C}p^{\ K/2}\log p$.

For H_2, by the inequality (71), we have

$$|H_2| \leq 2\max_{a\in[r]}\|[\boldsymbol{W}_{:a}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:a}^T\boldsymbol{X}\boldsymbol{V}]^s\|^2 \cdot \frac{r^2(L^{(t)})^2}{\Delta_{\min}^2} \quad C\frac{\bar{C}^2}{\tilde{C}^2}\|[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{a:}]^s\|^2, \tag{75}$$

where the last inequality follows from the condition (44) in Condition 1.

For H_3, by Cauchy-Schwartz inequality, we have

$$|H_3| \leq \|[\boldsymbol{X}_{i:}\boldsymbol{V}^{(t)}]^s \quad [\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}^{(t)}]^s\||H_1|^{1/2} \quad 2\tilde{C}^{\ 1}\|[\boldsymbol{S}_{z(i):}]^s \quad [\boldsymbol{S}_{a:}]^s\|^2, \tag{76}$$

following the inequalities (71) and (74).

Choose $\tilde{C}$ such that

$$\tilde{C}^{\ 2} + C\frac{\bar{C}^2}{\tilde{C}^2} + \tilde{C}^{\ 1} \leq \frac{1}{4}. \tag{77}$$

Therefore, we finish the proof of inequality (64) combining inequalities (74), (75), and (76).

$\square$

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

$$\|\boldsymbol{V} \quad \boldsymbol{V}^{(t)}\|_\sigma \cdot \sqrt{\frac{r^{K\ 1}}{p^{K\ 1}}}\frac{r}{\Delta_{\min}^2}L^{(t)}, \tag{78}$$

$$\|\boldsymbol{E}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\|_\sigma \lesssim \sqrt{\frac{r^{K-1}(pr^{K-1}+pr)}{p^{K-1}}}\frac{r}{\Delta_{\min}^2}L^{(t)}, \tag{79}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}]^s - [\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s\| \leq C\left(\frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K}+pr^{K+1}}{p^K}}\frac{\rho\overline{L^{(t)}}}{\Delta_{\min}}\right), \tag{80}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}]^s - [\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\| \leq C\left(\sqrt{\frac{rr^{2K}+pr^{K+2}}{p^K}}\frac{\rho\overline{L^{(t)}}}{\Delta_{\min}} + \frac{rL^{(t)}}{\Delta_{\min}}\right), \tag{81}$$

$$\max_{b\in[r]}\|[\boldsymbol{W}_{:b}^T\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s - [\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}]^s\| \leq C\left(\frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{rr^{2K}+pr^{K+2}}{p^K}}\frac{\rho\overline{L^{(t)}}}{\Delta_{\min}}\right), \tag{82}$$

for some positive universal constant C. In addition, the inequality (81) also holds by replacing $\boldsymbol{W}_{:b}^{(t)}$ to $\boldsymbol{W}_{:b}$.

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

1. Inequality (78). By Han et al. (2020, Proof of Lemma 5), we have

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

Then, we complete the proof of inequality (78) by applying Lemma 8 to the above inequality.

2. Inequality (79). By Han et al. (2020, Proof of Lemma 5), we have

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

Also, we complete the proof of inequality (78) by applying Lemma 8 to the above inequality.

3. Inequality (80). We upper bound the desired quantity by triangle inequality,

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

where

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

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

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

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

For I_1, we further bound I_1 by triangle inequality,

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

where

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

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

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

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

$$\sum_{a \neq b, a, b \in [r]} D_{ab} \lesssim \frac{r}{p} \sum_{i:\, z^{(t)}(i) \neq z(i)} \theta(i) \lesssim \frac{L^{(t)}}{\Delta_{\min}^2} \lesssim \frac{1}{\log p}, \tag{83}$$

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

$$D_{bb} = \frac{\sum_{i \in [p]} \theta(i) \mathbb{1}\left\{z(i) = z^{(t)}(i) = b\right\}}{\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \geq \frac{c(\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\} - p\ell^{(t)})}{\sum_{i \in [p]} \mathbb{1}\left\{z^{(t)}(i) = b\right\}} \gtrsim 1 - \frac{1}{\log p}, \tag{84}$$

under the inequality (44) in Condition 1. By the definition of $W, W^{(t)}, V$, we have

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

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

$$\left\langle S_{b:}, D_{bb} S_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} S_{a:} \right\rangle = D_{bb} \|S_{b:}\|^2 + \sum_{a \neq b} D_{ab} \langle S_{b:}, S_{a:} \rangle$$

$$\geq D_{bb} \|S_{b:}\|^2 - \sum_{a \neq b, a \in [r]} D_{ab} \|S_{b:}\| \max_{a \in [r]} \|S_{a:}\|$$

$$\geq C,$$

where C is a positive constant, and the last inequality holds when p is large enough following the constraint of $\|S_{b:}\|$ in parameter space (2) and the bounds of D in (83) and (84).

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

$$\|[D_{bb} S_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} S_{a:}] \sin \alpha\| = \|(I_d - \text{Proj}(S_{b:})) \sum_{a \neq b, a \in [r]} D_{ab} S_{a:}\|$$

$$\leq \sum_{a \neq b, a \in [r]} D_{ab} \|(I_d - \text{Proj}(S_{b:})) S_{a:}\|$$

$$= \sum_{a \neq b, a \in [r]} D_{ab} \|S_{a:}\| \sin(\angle(S_{b:}, S_{a:}))$$

$$\lesssim \sum_{a \neq b, a \in [r]} D_{ab} \|S_{a:}\| \|S_{b:}^s - S_{a:}^s\|, \tag{85}$$

where the first inequality follows from the triangle inequality, and the last inequality follows from Lemma 2. Note that with bounds (83) and (84), when p is large enough, we have

$$\|W_{:b}^{(t),T} X V\| = \|D_{bb} S_{b:} + \sum_{a \neq b, a \in [r]} D_{ab} S_{a:}\| \geq D_{bb} \|S_{b:}\| - \sum_{a \neq b, a \in [r]} D_{ab} \|S_{a:}\| \geq C_1, \tag{86}$$

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

$$I_{11} \leq \sqrt{2} \sin \alpha$$

$$= \frac{k[D_{bb}\boldsymbol{S}_{b:} + \sum_{a \notin b, a \in [r]} D_{ab}\boldsymbol{S}_{a:}]\sin\alpha k}{k D_{bb}\boldsymbol{S}_{b:} + \sum_{a \notin b, a \in [r]} D_{ab}\boldsymbol{S}_{a:}k}$$

$$\frac{1}{C_1}\sum_{a \notin b, a \in [r]} D_{ab} \, k\boldsymbol{S}_{a:}k \, k\boldsymbol{S}_{b:}^s \quad \boldsymbol{S}_{a:}^s k$$

$$\cdot \frac{r}{p}\sum_{i \in [p]}\theta(i)\sum_{b \in [r]}\mathbb{1}\left\{z^{(t)}(i) = b\right\} k\boldsymbol{S}_{b:}^s \quad \boldsymbol{S}_{a:}^s k$$

$$\frac{rL^{(t)}}{\Delta_{\min}}, \tag{87}$$

where the second inequality follows from (85) and (86), and the last two inequalities follow by the definition of D_a and $L^{(t)}$, and the constraint of $k\boldsymbol{S}_{b:}k$ in parameter space (2).

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

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

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

$$k(\boldsymbol{W}_{:b}^T \quad \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}k \cdot \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\rho_{\overline{L^{(t)}}}}{\Delta_{\min}}. \tag{88}$$

Notice that

$$k(\boldsymbol{W}_{:b}^T \quad \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}k \quad k\boldsymbol{W}_{:b}^T \quad \boldsymbol{W}_{:b}^{(t),T}k \, k\boldsymbol{X}\boldsymbol{V}k_F \cdot \frac{r^{3/2}L^{(t)}}{\rho_{\overline{p}}\Delta_{\min}^2}k\boldsymbol{S}k k\boldsymbol{\Theta}\boldsymbol{M}k_\sigma \cdot \frac{\rho_{\overline{rL^{(t)}}}}{\Delta_{\min}}, \tag{89}$$

where the second inequality follows from Han et al. (2020, Inequality (121), Proof of Lemma 5) and the last inequality follows from Lemma 5 and (44) in Condition 1. Note that $\|\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}\| = k\boldsymbol{S}_{b:}k \quad c_3$ and $k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}k \quad C_1$ by inequality (86). Therefore, we have

$$I_{12} \cdot k(\boldsymbol{W}_{:b}^T \quad \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{E}\boldsymbol{V}k + k(\boldsymbol{W}_{:b}^T \quad \boldsymbol{W}_{:b}^{(t),T})\boldsymbol{X}\boldsymbol{V}k \, kk\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}k$$

$$\cdot \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\rho_{\overline{L^{(t)}}}}{\Delta_{\min}} + \frac{\rho_{\overline{rL^{(t)}}}}{\Delta_{\min}}\sqrt{\frac{r^{2K}}{p^K}}$$

$$\cdot \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\rho_{\overline{L^{(t)}}}}{\Delta_{\min}}, \tag{90}$$

where second inequality follows from the inequalities (88), (89), and (40) in Condition 1.

Hence, combining inequalities (87) and (90) yields

$$I_1 \cdot \frac{rL^{(t)}}{\Delta_{\min}} + \sqrt{\frac{r^{2K} + pr^{K+1}}{p^K}}\frac{\rho_{\overline{L^{(t)}}}}{\Delta_{\min}}. \tag{91}$$

For I_2 and I_3, recall that $\|\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}\| = k\boldsymbol{S}_{b:}k \quad c_3$ and $k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}k \quad C_1$ by inequality (86). By triangle inequality and (40) in Condition 1, we have

$$I_2 \quad \frac{k\boldsymbol{W}_{:b}^T\boldsymbol{E}\boldsymbol{V}k}{k\boldsymbol{W}_{:b}^T\boldsymbol{X}\boldsymbol{V}k} \cdot k\boldsymbol{W}_{:b}^T\boldsymbol{E}\boldsymbol{V}k \cdot \frac{r^K}{p^{K/2}}, \tag{92}$$

and

$$I_3 \quad \frac{k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}k}{k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}k} \cdot k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}k \cdot \frac{r^K}{p^{K/2}}. \tag{93}$$

Therefore, combining the inequalities (91), (92), and (93), we finish the proof of inequality (80).

38

4. Inequality (81). Here we only show the proof of inequality (81) with $\boldsymbol{W}_{:b}^{(t)}$. The proof also holds by replacing $\boldsymbol{W}_{:b}^{(t)}$ to $\boldsymbol{W}_{:b}$, and we omit the repeated procedures.

We upper bound the desired quantity by triangle inequality

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

where

$$J_1 = \left\| \frac{\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}}{k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}k} \quad \frac{\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{Y}\boldsymbol{V}^{(t)}}{k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}k} \right\|,$$

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

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

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

For J_1, by triangle inequality, we have

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

where

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

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

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

for $k = 1, \ldots, K\ 1$. Let β_k denote the angle between $\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^k$ and $\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{k+1}$. With the same idea to prove I_{11} in inequality (87), we bound J_{11}^k by the trigonometric function of β_k.

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

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

where

$$\mathcal{T}_k = \boldsymbol{X} \quad_1 \boldsymbol{W}_{:b}^{(t),T} \quad_2 \boldsymbol{W}^T \quad_3 \quad_k \boldsymbol{W}^T \quad_{k+1} \boldsymbol{W}^{(t),T} \quad_{k+1} \quad_K \boldsymbol{W}^{(t),T}$$

$$\mathcal{T}_{k+1} = \boldsymbol{X} \quad_1 \boldsymbol{W}_{:b}^{(t),T} \quad_2 \boldsymbol{W}^T \quad_3 \quad_k \boldsymbol{W}^T \quad_{k+1} \boldsymbol{W}^T \quad_{k+1} \quad_K \boldsymbol{W}^{(t),T}.$$

Recall the definition of confusion matrix $\boldsymbol{D} = \boldsymbol{M}^T\boldsymbol{\Theta}^T\boldsymbol{W}^{(t)} = \mathrm{J}D_{ab}\mathrm{K} \, 2 \, \mathbb{R}^{r \ r}$. We have

$$\left\langle \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^k, \boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{k+1} \right\rangle = \hbar\mathrm{Mat}_{k+1}(\mathcal{T}_k), \mathrm{Mat}_{k+1}(\mathcal{T}_{k+1})i$$

$$= \left\langle \boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^k, \boldsymbol{S}\boldsymbol{Z}^k \right\rangle$$

$$= \sum_{b2[r]} \left(D_{bb}k\boldsymbol{S}_{b:}\boldsymbol{Z}^k k^2 + \sum_{a\not=b, a2[r]} D_{ab}\left\langle \boldsymbol{S}_{a:}\boldsymbol{Z}^k, \boldsymbol{S}_{b:}\boldsymbol{Z}^k \right\rangle \right)$$

$$\&\ (1\quad \log p^{\ 1})\min_{a2[r]}k\boldsymbol{S}_{a:}\boldsymbol{Z}^k k^2 \quad \log p^{\ 1}\max_{a2[r]}k\boldsymbol{S}_{a:}\boldsymbol{Z}^k k^2, \tag{94}$$

39

where $\boldsymbol{Z}^k = \boldsymbol{D}_{:b} \otimes \boldsymbol{I}_r^{\otimes(k-1)} \otimes \boldsymbol{D}^{\otimes(K-k-1)}$, the equations follow by the tensor algebra and definitions, and the last inequality follows from the bounds of $\boldsymbol{D}$ in (83) and (84).

Note that

$$\|\boldsymbol{D}\|_\sigma \leq \|\boldsymbol{D}\|_F \leq \sqrt{\sum_{b \in [r]} D_{bb}^2 + (\sum_{a \neq b, a, b \in [r]} D_{ab})^2} \cdot \sqrt{r + \log^2 p} \lesssim 1, \tag{95}$$

where the second inequality follows from inequality (83), and the fact that for all $b \in [r]$,

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

Also, we have

$$\lambda_r(\boldsymbol{D}) \geq \lambda_r(\boldsymbol{W}^{(t)})\lambda_r(\boldsymbol{\Theta M}) \gtrsim 1, \tag{96}$$

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

$$1 \lesssim \|\boldsymbol{D}_{:b}\| \lambda_r(\boldsymbol{D})^{K-k-1} \leq \lambda_{r^{K-2}}(\boldsymbol{Z}^k) \leq \|\boldsymbol{Z}^k\|_\sigma \leq \|\boldsymbol{D}_{:b}\| \|\boldsymbol{D}\|_\sigma^{K-k-1} \lesssim 1. \tag{97}$$

Thus, we have bounds

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

Hence, when p is large enough, the inner product (94) is positive, which implies $\beta_k \in [0, \pi/2)$ and thus $2 \sin \frac{\beta_k}{2} \leq \sqrt{2} \sin \beta_k$.

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

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

where

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

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

To bound $\sin \beta_{k1}$, recall Definition 1 that for any cluster assignment $\bar{z}$ in the ε-neighborhood of true z,

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

Note that we have $\ell^{(t)} \leq \frac{L^{(t)}}{\Delta_{\min}^2} \leq \frac{C}{\mathcal{C}} r \log^{-1}(p)$ by Condition 1 and Lemma 8. Then, with the locally linear stability assumption, the is $\ell^{(t)}$-locally linearly stable; i.e.,

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

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

$$\min_{c \in \mathbb{R}} \|\boldsymbol{D} - c\tilde{\boldsymbol{D}}\|_F \leq \|\boldsymbol{I}_r - c_0 \mathrm{diag}(\boldsymbol{p}(z^{(t)}))\mathrm{diag}^{-1}(\boldsymbol{p}_\theta(z^{(t)}))\|_F \|\boldsymbol{D}\|_F$$

$$\leq \frac{\|\boldsymbol{p}(z^{(t)}) - c_0 \boldsymbol{p}_\theta(z^{(t)})\|}{\min_{a \in [r]} \|\boldsymbol{\theta}_{z^{(t),-1}(a)}\|_1}$$

40

$$= \frac{k\boldsymbol{p}(z^{(t)})k}{\min_{a\mathcal{2}[r]} k_{z^{(t)},\ ^{-1}(a)}k_1} \sin(\boldsymbol{p}(z^{(t)}),\boldsymbol{p}\ (z^{(t)}))$$

$$\cdot\ \frac{L^{(t)}}{\Delta_{\min}},$$

where the last inequality follows from Lemma 9, the constraint $\min_{i\mathcal{2}[p]}\theta(i)\quad c > 0$, $k\boldsymbol{p}(z^{(t)})k \quad p$ and $\min_{a\mathcal{2}[r]} k_{z^{(t)},\ ^{-1}(a)}k_1$ & p. By the geometry property of trigonometric function, we have

$$\sin\beta_{k1} = \min_{c\mathcal{2}\mathbb{R}} \frac{k\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\ k\ 1}\ (\boldsymbol{D}\quad c\tilde{\boldsymbol{D}})\ \boldsymbol{D}^{\ K\ k\ 1}k}{k\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\ k\ 1}\ \boldsymbol{D}^{\ K\ k}k} \tag{98}$$

$$\frac{k\boldsymbol{D}_{:b}^T\boldsymbol{S}kk\boldsymbol{D}\quad c_0\tilde{\boldsymbol{D}}k_\sigma k\boldsymbol{D}k_\sigma^{K\ k\ 1}}{k\boldsymbol{D}_{:b}^T\boldsymbol{S}k\lambda_r^{K\ k}(\boldsymbol{D})}$$

$$\cdot\quad k\boldsymbol{D}\quad c_0\tilde{\boldsymbol{D}}k_F$$

$$\cdot\ \frac{L^{(t)}}{\Delta_{\min}},$$

where the second inequality follows from the singular property of $\boldsymbol{D}$ in (95), (96) and the constraint of $\boldsymbol{S}$ in (2).

To bound $\sin\beta_{k2}$, let $\boldsymbol{C} = \mathrm{diag}(fk\boldsymbol{S}_{a:}kg_{a\mathcal{2}[r]})$. We have

$$\sin\beta_{k2}\ \cdot\ \frac{\left\|\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\ k\ 1}\ (\boldsymbol{I}_r\quad\tilde{\boldsymbol{D}})\ \boldsymbol{D}^{\ K\ k\ 1}\right\|}{k\boldsymbol{D}_{:b}^T\boldsymbol{S}\boldsymbol{I}_r^{\ k}\ \boldsymbol{D}^{\ K\ k\ 1}k} \tag{99}$$

$$\cdot\ \frac{k(\boldsymbol{I}_r\quad\tilde{\boldsymbol{D}}^T)\boldsymbol{S}\boldsymbol{Z}^k k_F}{k\boldsymbol{D}_{:b}^T\boldsymbol{S}k\lambda_r^{K\ k\ 1}(\boldsymbol{D})}$$

$$\cdot\ k(\boldsymbol{I}_r\quad\tilde{\boldsymbol{D}}^T)\boldsymbol{S}\boldsymbol{C}^{\ 1}k_F k\boldsymbol{C}\boldsymbol{Z}^k k_\sigma$$

$$\cdot\ \frac{r}{p}\sum_{i\mathcal{2}[p]}\theta(i)\sum_{b\mathcal{2}[r]}k\boldsymbol{S}_{b:}^s\quad\boldsymbol{S}_{z(i):}^s k$$

$$\cdot\ \frac{L^{(t)}}{\Delta_{\min}},$$

where the third inequality follows from the singular property of $\boldsymbol{D}$ and the boundedness of $\boldsymbol{S}$, and the fourth inequality follows from the definition of $\tilde{\boldsymbol{D}}$, boundedness of $\boldsymbol{S}$, the lower bound of , and the singular property of $\boldsymbol{Z}^k$ in inequality (97).

Combining (98) and (99) yields

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

Finally, by triangle inequality, we obtain

$$J_{11}\quad \sum_{k=1}^{K\ 1} J_{11}^k\ \cdot\ \sum_{k=1}^{K\ 1}\sin\beta_k\ \cdot\ (K\quad 1)\frac{rL^{(t)}}{\Delta_{\min}}. \tag{100}$$

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

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

Note that

$$k\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}k = k\boldsymbol{D}^T\boldsymbol{S}\boldsymbol{Z}^1 k\quad \lambda_r(\boldsymbol{D})\ k\boldsymbol{S}k\lambda_{r^{K}\ 2}(\boldsymbol{Z}^1)\ \&\ 1, \tag{101}$$

41

where the inequality follows from the bounds (96) and (97).

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

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\| \lesssim \sqrt{\frac{r^{2K+1} + pr^{2+K}}{p^K}}\frac{(K-1)\rho_{\overline{L^{(t)}}}}{\Delta_{\min}}. \tag{102}$$

Notice that

$$\begin{aligned}
\|\boldsymbol{X}(\boldsymbol{V}^k - \boldsymbol{V}^{k+1})\|_F &\le \|(\boldsymbol{I} - \boldsymbol{D}^T)\boldsymbol{S}(\boldsymbol{I}_r^{(k-1)} - \boldsymbol{D}^{(K-k-1)})\|_F \\
&\le \|(\boldsymbol{W}^T - \boldsymbol{W}^{(t),T})\boldsymbol{\Theta}\boldsymbol{M}\|_F \|\boldsymbol{S}\|_F \|\boldsymbol{D}\|_\sigma^{K-k-1} \\
&\lesssim \|\boldsymbol{W}^T - \boldsymbol{W}^{(t),T}\| \|\boldsymbol{\Theta}\boldsymbol{M}\|_\sigma \\
&\le \frac{\rho_{\overline{rL^{(t)}}}}{\Delta_{\min}},
\end{aligned} \tag{103}$$

where the first inequality follows from the tensor algebra in inequality (94), the second inequality follows from the fact that $\boldsymbol{I} = \boldsymbol{W}^T\boldsymbol{\Theta}\boldsymbol{M}$, and the last inequality follows from Han et al. (2020, Proof of Lemma 5). It follows from (103) and Lemma 9 that

$$\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}(\boldsymbol{V} - \boldsymbol{V}^{(t)})\| \le \|\boldsymbol{W}_{:b}^{(t),T}\|\sum_{k=1}^{K-1}\|\boldsymbol{X}(\boldsymbol{V}^k - \boldsymbol{V}^{k+1})\|_F \lesssim \frac{\rho_{\overline{rL^{(t)}}}}{\rho_{\overline{p}\Delta_{\min}}}. \tag{104}$$

Note that $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|$ are lower bounded by inequalities (86) and (101), respectively. We have

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

where the second inequality follows from inequalities (102), (104), and the inequality (40) in Condition 1.

For J_2 and J_3, recall that $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|$ and $\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|$ are lower bounded by inequalities (86) and (101), respectively. By triangle inequality and inequality (40) in Condition 1, we have

$$J_2 \le \frac{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\|}{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}\|}\lesssim \|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}, \tag{105}$$

and

$$J_3 \le \frac{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}^{(t)}\|}{\|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{X}\boldsymbol{V}^{(t)}\|}\lesssim \|\boldsymbol{W}_{:b}^{(t),T}\boldsymbol{E}\boldsymbol{V}\| \lesssim \frac{r^K}{p^{K/2}}. \tag{106}$$

Therefore, combining the inequalities (100), (105), and (106), we finish the proof of inequality (81).

5. Inequality (82). By triangle inequality, we upper bound the desired quantity

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

following the inequalities (80) and (81). Therefore, we finish the proof of inequality (82).

□