Generalized Low-rank plus Sparse Tensor Estimation by Fast Riemannian Optimization

Jian-Feng Cai*, Jingyang Li and Dong Xia†
Hong Kong University of Science and Technology

Abstract

We investigate a generalized framework to estimate a latent low-rank plus sparse tensor, where the low-rank tensor often captures the multi-way principal components and the sparse tensor accounts for potential model mis-specifications or heterogeneous signals that are unexplainable by the low-rank part. The framework is flexible covering both linear and non-linear models, and can easily handle continuous or categorical variables. We propose a fast algorithm by integrating the Riemannian gradient descent and a novel gradient pruning procedure. Under suitable conditions, the algorithm converges linearly and can simultaneously estimate both the low-rank and sparse tensors. The statistical error bounds of final estimates are established in terms of the gradient of loss function. The error bounds are generally sharp under specific statistical models, e.g., the robust tensor PCA and the community detection in hypergraph networks with outlier vertices. Moreover, our method achieves non-trivial error bounds for heavy-tailed tensor PCA whenever the noise has a finite $2+\varepsilon$ moment. We apply our method to analyze the international trade flow dataset and the statistician hypergraph co-authorship network, both yielding new and interesting findings.

1 Introduction

In recent years, massive *multi-way* datasets, often called *tensor* data, have routinely arisen in diverse fields. An mth-order tensor is a multilinear array with m ways, e.g., matrices are second order tensors. These multi-way structures often emerge when, to name a few, information features are collected from distinct domains (Chen et al., 2020a; Liu et al., 2017; Han et al., 2020; Bi et al., 2018; Zhang et al., 2020b; Wang and Zeng, 2019), the multi-relational interactions or higher-order interactions of entities are present (Ke et al., 2019; Jing et al., 2020; Kim et al., 2017; Paul and

*Jian-Feng Cai's research was partially supported by Hong Kong RGC Grant GRF 16310620 and GRF 16309219.
†Dong Xia's research was partially supported by Hong Kong RGC Grant ECS 26302019 and GRF 16303320.

Chen, 2020; Luo and Zhang, 2020; Wang and Li, 2020; Ghoshdastidar and Dukkipati, 2017; Pensky and Zhang, 2019), or the higher-order moments of data are explored (Anandkumar et al., 2014; Sun et al., 2017; Hao et al., 2020). There is an increasing demand for effective methods to analyze large and complex tensorial datasets. However, tensor data are often ultra high-dimensional because tensor size grows exponentially fast with respect to its order. Analyzing large tensor datasets is thus challenging.

Low-rank tensor models are a class of statistical models for describing and analyzing tensor datasets. At its core is the assumption that the observed data obeys a distribution that is characterized by a *latent* low-rank tensor $\boldsymbol{\mathcal{T}}^*$. Generalized from matrices, a tensor is low-rank if its fiber spaces (akin to the row and column spaces of a matrix) have low ranks. These are called the *Tucker* ranks (the formal definition is deferred to Section 1.2). Low-rank structures provide the benefit of dimension reduction. Indeed, the ambient dimension of $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ is $d_1 \cdots d_m$, but its actual degree of freedom is only $O(d_1 + \cdots + d_m)$ if the ranks of $\boldsymbol{\mathcal{T}}^*$ are $O(1)$. Under low-rank tensor models, the latent tensor $\boldsymbol{\mathcal{T}}^*$ usually preserves the principal components of each dimension as well as their multi-way interactions. Oftentimes, analyzing tensor datasets boils down to estimating the low-rank $\boldsymbol{\mathcal{T}}^*$. This procedure is usually referred to as the *low-rank tensor estimation*. Together with specifically designed algorithms, low-rank tensor methods have demonstrated encouraging performances on many real-world applications and datasets such as the spatial and temporal pattern analysis of human brain developments (Liu et al., 2017), community detection on multi-layer networks and hypergraph networks (Jing et al., 2020; Ke et al., 2019; Wang and Li, 2020), multi-dimensional recommender system (Bi et al., 2018), comprehensive climate data analysis (Chen et al., 2020a), learning the hidden components of mixture models (Anandkumar et al., 2014), analysis of brain dynamic functional connectivity(Sun and Li, 2019), image denoising and recovery (Xia et al., 2021; Liu et al., 2012) and etc.

However, the exact low-rank assumption is stringent and sometimes untrue, making low-rank tensor methods vulnerable under model misspecification or in the existence of outliers or heterogeneous signals. While low-rank structure underscores the multi-way principal components, it fails to capture the dimension-specific outliers or heterogeneous signals that often carry distinctive and useful information. Consider the international trade flow dataset (see Section 8) that forms a third-order tensor by the dimensions countries $\times$ countries $\times$ commodities. On the one hand we observe that the low-rank structure is capable to reflect the shared similarities among countries such as their geographical locations and economic structures, but on the other hand the low-rank structure tends to disregard the heterogeneity in the trading flows of different countries. This vital heterogeneity often reveals distinctive trading patterns of certain commodities for some countries. Moreover, the heterogeneous signals are usually full-rank and strong that can deteriorate the es-

timates of the multi-way low-rank principal components. We indeed observe that by filtering out these outliers or heterogeneous signals, the resultant low-rank estimates become more insightful. It is therefore advantageous to decouple the low-rank signal and the heterogeneous one in the procedure of low-rank tensor estimation. Fortunately, these outliers or heterogeneous signals are usually representable by a *sparse* tensor, which, is identifiable in generalized low-rank tensor models under suitable conditions.

In this paper, we propose a generalized low-rank plus sparse tensor model to analyze tensorial datasets. Our fundamental assumption is that the observed data is sampled from a statistical model characterized by the latent tensor $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$. We assume $\boldsymbol{\mathcal{T}}^*$ to be low-rank capturing the multi-way principal components, and $\boldsymbol{\mathcal{S}}^*$ to be sparse (the precise definition of being "sparse" can be found in Section 2) addressing potential model mis-specifications, outliers or heterogeneous signals that are unexplainable by the low-rank part. The goal is to estimate both $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ when provided with the observed dataset. Our framework is very flexible which covers both linear and non-linear models, and can easily handle both quantitative and categorical data. We note that our framework reduces to the typical (exact) low-rank tensor estimation if the sparse component $\boldsymbol{\mathcal{S}}^*$ is absent. Compared with existing literature on low-rank tensor methods (Han et al., 2020; Xia et al., 2021; Zhang and Xia, 2018; Yuan and Zhang, 2017; Sun et al., 2017; Wang and Li, 2020; Hao et al., 2020), our framework and method are more robust, particularly when the latent tensor is only *approximately* low-rank or when the noise have heavy tails.

With a properly chosen loss function $\mathfrak{L}(\cdot)$, we formulate the generalized low-rank plus sparse tensor estimation into an optimization framework, which aims at minimizing $\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}})$ subject to the low-rank and sparse constraints on $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{S}}$, respectively. This optimization program is highly non-convex and can be solved only locally. We propose a new and fast algorithm to solve for the underlying tensors of interest. The algorithm is iterative and consists of two main ingredients: the *Riemannian gradient descent* and the *gradient pruning*. By viewing the low-rank solution as a point on the Riemannian manifold, we adopt Riemannian gradient descent to update the low-rank estimate. Basically, the Riemannian gradient is the projection of the *vanilla gradient* $\nabla\mathfrak{L}$ onto the tangent space of a Riemannian manifold. Unlike the vanilla gradient that is usually full-rank, the Riemannian gradient is often low-rank which can significantly boost up the speed of updating the low-rank estimate. Provided with a reliable estimate of the low-rank tensor $\boldsymbol{\mathcal{T}}^*$, the gradient pruning is a fast procedure to update our estimate of the sparse tensor $\boldsymbol{\mathcal{S}}^*$. It is based on the belief that, under suitable conditions, if the current estimate $\widehat{\boldsymbol{\mathcal{T}}}$ is close to $\boldsymbol{\mathcal{T}}^*$ entry-wisely, the entries of the gradient $\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}})$ should have small magnitudes on the complement of the support of $\boldsymbol{\mathcal{S}}^*$. Then it suffices to run a screening of the entries of $\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}})$, locate its entries with large magnitudes and choose $\widehat{\boldsymbol{\mathcal{S}}}$ to minimize the magnitudes of those entries of $\nabla\mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}} + \widehat{\boldsymbol{\mathcal{S}}})$.

The procedure looks like pruning the gradient $\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}})$ – thus the name gradient pruning. The algorithm alternates between Riemannian gradient descent and gradient pruning until reaching a locally optimal solution. Provided with a warm initialization of $\boldsymbol{\mathcal{T}}^*$ and assuming certain regularity conditions of the loss function, we prove that our algorithm converges with a constant contraction rate that is strictly smaller than 1. We also establish the error bounds of the final estimates in terms of the gradient of loss at $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$, yielding many interesting implications in specific statistical models.

Due to the benefit of fast computations, Riemannian optimization and its applications in matrix and tensor related problems (Wei et al., 2016; Cai et al., 2019b, 2020c; Vandereycken, 2013; Kressner et al., 2014; Absil et al., 2009; Vandereycken, 2013; Edelman et al., 1998) have been intensively investigated in recent years. However, the extensions of Riemannian gradient descent to a non-linear framework, in the existence of a sparse component or stochastic noise and their statistical performances, are relatively unknown. To compare with the existing literature on statistical tensor analysis and Riemannian optimization, we highlight our main contributions as follows.

1.1 Our Contributions

We propose a novel and generalized framework to analyze tensor datasets. Compared with the existing low-rank tensor literature, our framework allows the latent tensor to be high-rank, as long as it is within the sparse perturbations of a low-rank tensor. The sparse tensor can account for potential mis-specifications of the exact low-rank tensor models. This makes our framework robust to outliers, heavy-tailed distributions, heterogeneous signals and etc. Meanwhile, our framework is flexible which covers both linear and non-linear models, and is applicable to both continuous and categorical variables.

We develop a new and fast algorithm which can simultaneously estimate both the low-rank and the sparse tensors. The algorithm is based on the integration of Riemannian gradient descent and a novel gradient pruning procedure. The existing literature on the Riemannian optimization for low-rank tensor estimation primarily focus on linear models of exact low-rank tensors. Our proposed algorithm works for both linear and non-linear models, adapts to additional sparse perturbations, and is reliable in the existence of stochastic noise. We prove, in a general framework, that our algorithm converges fast even with fixed step-sizes, and establish the statistical error bounds of final estimates. The error bounds are sharp and proportional to the intrinsic degrees of freedom under many specific statistical models.

To showcase the superiority of our methods, we consider applying our framework to four interesting examples. The first application is on the robust tensor principal component analysis (PCA) where the observation is simply $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$ with additive sub-Gaussian noise. We show that

our algorithm can recover both $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ with sharp error bounds, and recover the support of $\boldsymbol{\mathcal{S}}^*$ under fairly weak conditions. The second example is on the tensor PCA when the noise have heavy tails. We show that our framework is naturally immune to the potential outliers caused by the heavy-tailed noise, and demonstrate that our method achieves non-trivial error bounds as long as the noise have a finite $2+\varepsilon$ moment. This bridges a fundamental gap in the understanding of tensor PCA since the existing methods are usually effective only under sub-Gaussian or sub-Exponential noise. We then apply our framework to learn the latent low-rank structure $\boldsymbol{\mathcal{T}}^*$ from a binary tensorial observation, assuming the Bernoulli tensor model with a general link function, e.g., the logistic and probit link. Compared with the existing literature on binary tensor analysis, our method is robust and allows an arbitrary but sparse corruption $\boldsymbol{\mathcal{S}}^*$. Finally, based on the generalized framework, we derive a robust method to detect communities in hypergraph networks in the existence of outlier vertices. We propose a hypergraph generalized stochastic block model that incorporates a group of outlier vertices. These outlier vertices can connect others (be them normal or outlier) by hyperedges in an arbitrary way. Under fairly weak conditions, our method can consistently estimate the latent low-rank tensor and recover the underlying communities. More details on these applications and related literature reviews can be found in Section 5.

Lastly, we employ our method to analyze two real-world datasets: the international commodity trade flow network (continuous variables) and the statistician hypergraph co-authorship network (binary variables). We observe that the low-rank plus sparse tensor framework yields intriguing and new findings that are unseen by the exact low-rank tensor methods. Interestingly, the sparse tensor can nicely capture informative patterns which are overlooked by the multi-way principal components.

1.2 Notations and Preliminaries of Tensor

Throughout the paper, we use calligraphic-font bold-face letters (e.g. $\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{T}}_1$) to denote tensors, bold-face capital letters (e.g. $\mathbf{T}, \mathbf{X}, \mathbf{T}_1$) for matrices, bold-face lower-case letters (e.g. $\mathbf{t}, \mathbf{x}, \mathbf{t}_1$) for vectors and blackboard bold-faced letters (e.g. $\mathbb{R}, \mathbb{M}, \mathbb{U}, \mathbb{T}$) for sets. We use square brackets with subscripts (e,g. $[\boldsymbol{\mathcal{T}}]_{i_1,i_2,i_3}, [\mathbf{T}]_{i_1,i_2}, [\mathbf{t}]_{i_1}$) to represent corresponding entries of tensors, matrices and vectors, respectively. Denote $[\boldsymbol{\mathcal{T}}]_{i_1,:,:}$ and $[\mathbf{T}]_{i_1,:}$ the i_1-th frontal-face and i_1-th row-vector of $\boldsymbol{\mathcal{T}}$ and $\mathbf{T}$, respectively. We use $\|\cdot\|_{\mathrm{F}}$ to denote the Frobenius norm of matrices and tensors, and use $\|\cdot\|_{\ell_p}$ to denote the ℓ_p-norm of vectors or vectorized tensors for $0 \le p \le \infty$. Thus, $\|\mathbf{v}\|_{\ell_0}$ represents the number of non-zero entries of $\mathbf{v}$, and $\|\mathbf{v}\|_{\ell_\infty}$ denotes the largest magnitude of the entries of $\mathbf{v}$. The j-th canonical basis vector is written as $\mathbf{e}_j$ whose actual dimension might vary at different appearances. We also use $C, C_1, C_2, c, c_1, c_2 \cdots$ to denote some absolute constants whose actual values can change at different lines.

An m-th order tensor is an m-way array, e.g., $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ means that its j-th dimension has size d_j. Thus, $\boldsymbol{\mathcal{T}}$ has in total $d_1 \cdots d_m$ entries. The j-th matricization (also called unfolding) $\mathcal{M}_j(\cdot) : \mathbb{R}^{d_1 \times \cdots \times d_m} \mapsto \mathbb{R}^{d_j \times d_j^-}$ with $d_j^- = (d_1 \cdots d_m)/d_j$ is a linear mapping so that, for example if $m = 3$, $[\mathcal{M}_1(\boldsymbol{\mathcal{T}})]_{i_1,(i_2-1)d_3+i_3} = [\boldsymbol{\mathcal{T}}]_{i_1,i_2,i_3}$ for $\forall i_j \in [d_j]$. Then, the collection $\text{rank}(\boldsymbol{\mathcal{T}}) := \big(\text{rank}(\mathcal{M}_1(\boldsymbol{\mathcal{T}})), \cdots, \text{rank}(\mathcal{M}_m(\boldsymbol{\mathcal{T}}))\big)^\top$ is called the multi-linear ranks or *Tucker ranks* of $\boldsymbol{\mathcal{T}}$. Given a matrix $\mathbf{W}_j \in \mathbb{R}^{p_j \times d_j}$ for any $j \in [m]$, the multi-linear product, denoted by $\times_j$, between $\boldsymbol{\mathcal{T}}$ and $\mathbf{W}_j$ is defined by $[\boldsymbol{\mathcal{T}} \times_j \mathbf{W}_j]_{i_1,\cdots,i_m} := \sum_{k=1}^{d_j} [\boldsymbol{\mathcal{T}}]_{i_1,\cdots,i_{j-1},k,i_{j+1},\cdots,i_m} \cdot [\mathbf{W}_j]_{i_j,k}$, $\forall i_{j'} \in [d_{j'}]$ for $j' \neq j; \forall i_j \in [p_j]$. If $\boldsymbol{\mathcal{T}}$ has Tucker ranks $\mathbf{r} = (r_1, \cdots, r_m)^\top$, there exist $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ and $\mathbf{U}_j \in \mathbb{R}^{d_j \times r_j}$ satisfying $\mathbf{U}_j^\top \mathbf{U}_j = \mathbf{I}_{r_j}$ for all $j \in [m]$ such that

$$\boldsymbol{\mathcal{T}} = \mathbf{C} \cdot [\![\mathbf{U}_1, \cdots, \mathbf{U}_m]\!] := \boldsymbol{\mathcal{C}} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m. \tag{1.1}$$

The representation (1.1) is referred to as the Tucker decomposition of a low-rank tensor. Tucker ranks and decomposition are well-defined. Readers are suggested to refer (Kolda and Bader, 2009) for more details and examples on tensor decomposition and tensor algebra.

2 General Low-rank plus Sparse Tensor Model

Suppose that we observe data $\mathfrak{D}$, which can be, for instance, simply a tensorial observation such as the binary adjacency tensor of a hypergraph network or multi-layer network (Ke et al., 2019; Jing et al., 2020; Kim et al., 2017; Paul and Chen, 2020; Luo and Zhang, 2020; Wang and Li, 2020); a real-valued tensor describing multi-dimensional observations (Han et al., 2020; Chen et al., 2020a; Sun et al., 2017; Sun and Li, 2019; Liu et al., 2017); or a collection of pairs of tensor covariate and real-valued response (Hao et al., 2020; Zhang et al., 2020a; Xia et al., 2020; Raskutti et al., 2019; Chen et al., 2019). At the core of our model is the assumption that the observed $\mathfrak{D}$ is sampled from a distribution characterized by a latent large tensor. In high-dimensional settings, the latent tensor is often assumed structural such as being low-rank. This gives rise to a large literature on low-rank tensor estimation from diverse fields. See, e.g., (Liu et al., 2012; Bi et al., 2018, 2020; Xia et al., 2021; Hao et al., 2020; Sun et al., 2017; Cai et al., 2019a; Yuan and Zhang, 2017; Wang and Li, 2020) and references therein. While various exact low-rank tensor models have been proposed and investigated, the exact low-rank assumption is sometimes stringent and may be untrue in reality, making them vulnerable and sensitive under model mis-specifications and outliers.

To develop more robust tensor methods, the more reasonable assumption is that the data $\mathfrak{D}$ is driven by a latent tensor that is only approximately low-rank. More exactly, we assume that the latent tensor is decomposable as the sum of a low-rank tensor and a sparse tensor, denoted by $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$, where $\boldsymbol{\mathcal{T}}^*$ has small multi-linear ranks and $\boldsymbol{\mathcal{S}}^*$ is sparse. Unlike the exact low-rank

tensor models, the additional sparse tensor $\boldsymbol{\mathcal{S}}^*$ can account for potential model mis-specifications and outliers. Consider that $\boldsymbol{\mathcal{T}}^*$ has multi-linear ranks $\mathbf{r} = (r_1, \cdots, r_m)^\top$ with $r_j \ll d_j$ so that $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ where

$$\mathbb{M}_{\mathbf{r}} := \big\{\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_1 \times \cdots \times d_m} : \mathrm{rank}\big(\mathcal{M}_j(\boldsymbol{\mathcal{W}})\big) \le r_j,\ \forall j \in [m]\big\}. \tag{2.1}$$

As for the sparse tensor, we assume that each slice of $\boldsymbol{\mathcal{S}}^*$ has at most α-portion of entries being non-zero for some $\alpha \in (0, 1)$. We write $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$ where the latter is defined by

$$\mathbb{S}_\alpha := \big\{\boldsymbol{\mathcal{S}} \in \mathbb{R}^{d_1 \times \cdots \times d_m} : \|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{S}})\|_{\ell_0} \le \alpha d_j^-,\ \forall j \in [m], i \in [d_j]\big\} \tag{2.2}$$

where $\mathbf{e}_i$ denotes the i-th canonical basis vector whose dimension varies at different appearances. We note that another popular approach for modeling sparse tensors is to assume a sampling distribution on the support of $\boldsymbol{\mathcal{S}}^*$ (e.g., (Candès et al., 2011; Chen et al., 2020b; Cherapanamjeri et al., 2017)). The deterministic model $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$ is more general and has been widely explored in the literature (Hsu et al., 2011; Netrapalli et al., 2014; Yi et al., 2016; Klopp et al., 2017).

When the low-rank tensor $\boldsymbol{\mathcal{T}}^*$ is also sparse, it is generally impossible to distinguish between $\boldsymbol{\mathcal{T}}^*$ and its sparse counterpart $\boldsymbol{\mathcal{S}}^*$. To make $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ identifiable, we assume that $\boldsymbol{\mathcal{T}}^*$ satisfies the *spikiness condition* meaning that the information it carries spreads fairly across nearly all its entries. Put differently, the spikiness condition enforces $\boldsymbol{\mathcal{T}}^*$ to be dense – thus distinguishable from the sparse $\boldsymbol{\mathcal{S}}^*$. This is a typical condition in robust matrix estimation (Yi et al., 2016; Xu et al., 2012; Candès et al., 2011) and tensor completion (Xia and Yuan, 2019; Xia et al., 2021; Cai et al., 2020a, 2019a; Montanari and Sun, 2018; Potechin and Steurer, 2017). We emphasize that Assumption 1 is necessary only for guaranteeing the identifiability of $\boldsymbol{\mathcal{S}}^*$. For exact low-rank tensor models where $\boldsymbol{\mathcal{S}}^*$ is absent, this assumption is generally not required. See Section 6 for more details.

Assumption 1. *Let $\boldsymbol{\mathcal{T}}^\star \in \mathbb{M}_{\mathbf{r}}$, and suppose there exists $\mu_1 > 0$ such that the following holds:*

$$\mathit{Spiki}(\boldsymbol{\mathcal{T}}^*) := (d^*)^{1/2}\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \mu_1,$$

where $d^* = d_1 \cdots d_m$.

We denote $\mathbb{U}_{\mathbf{r},\mu_1} := \big\{\boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}} : \mathsf{Spiki}(\boldsymbol{\mathcal{T}}) \le \mu_1\big\}$ the set of low-rank tensors with spikiness bounded by μ_1.

Relation between spikiness condition and incoherence condition Let $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ admits a Tucker decomposition $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot [\![\mathbf{U}_1^*, \cdots, \mathbf{U}_m^*]\!]$ with $\boldsymbol{\mathcal{C}}^* \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ and $\mathbf{U}_j^* \in \mathbb{R}^{d_j \times r_j}$ satisfying $\mathbf{U}_j^{*\top}\mathbf{U}_j^* = \mathbf{I}_{r_j}$ for all $j \in [m]$. Suppose that there exists $\mu_0 > 0$ so that

$$\mu(\boldsymbol{\mathcal{T}}^*) := \max_{j\in[m]} \max_{i\in[d_j]} \|\mathbf{e}_i^\top \mathbf{U}_j^*\|_{\ell_2} \cdot (d_j/r_j)^{1/2} \le \sqrt{\mu_0},$$

where $\mathbf{e}_i$ denotes the i-th canonical basis vector. Then, $\boldsymbol{\mathcal{T}}^*$ is said to satisfy the incoherence condition with constant μ_0. Notice the spikiness condition forces the "energy" of the tensor to spread fairly across all its entries. It also implies the incoherence condition *and vice versa*. The details about the relation between spikiness and incoherence are summarized in Lemma B.5 or the Proposition 2 in (Xia et al., 2021).

After observing data $\mathfrak{D}$, our goal is to estimate the underlying $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_\alpha)$. Oftentimes, the problem is formulated as an optimization program equipped with a properly chosen loss function. More specifically, let $\mathfrak{L}(\cdot) := \mathfrak{L}_{\mathfrak{D}}(\cdot) : \mathbb{R}^{d_1 \times \cdots \times d_m} \mapsto \mathbb{R}$ be a smooth (see Assumption 2) loss function whose actual form depends on the particular applications. The estimators of $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ are then defined by

$$(\widehat{\boldsymbol{\mathcal{T}}}_\gamma,\ \widehat{\boldsymbol{\mathcal{S}}}_\gamma) := \underset{\boldsymbol{\mathcal{T}} \in \mathbb{U}_{\mathbf{r},\mu_1}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}}{\arg\min} \mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}), \tag{2.3}$$

where $\gamma > 1$ is a tuning parameter determining the desired sparsity level of $\widehat{\boldsymbol{\mathcal{S}}}_\gamma$. For ease of exposition, we tentatively assume that the true ranks are known. In real-world applications, $\mathbf{r}, \alpha$ and γ are all tuning parameters. We discuss in Section 8 general approaches for choosing these parameters in practice.

Remark 2.1. *(The role of $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$ in the loss function) In view of the definition of $(\widehat{\boldsymbol{\mathcal{T}}}_\gamma, \widehat{\boldsymbol{\mathcal{S}}}_\gamma)$ as (2.3), one might regard $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ as the true minimizer of the objective function in (2.3). This is, however, generally untrue. Indeed, for some applications where no stochastic noise exists, $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ is often the unique global minimizer of (2.3). See, for instance, the noiseless robust tensor PCA in Example 2.1. However, for many applications where stochastic noise or random sampling exists, $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ is not necessarily the minimizer of (2.3). As shown in Theorem 4.1, our algorithm simply requires that the function $\mathfrak{L}(\cdot)$ is strongly convex and smooth in a small neighbourhood of $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$. Nevertheless, it worth to mention that, in many applications, $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ is the minimizer of $\mathbb{E}\mathfrak{L}(\cdot)$ where the expectation is w.r.t. the stochastic noise or random sampling. See, for instance, Example 2.1 and 2.2.*

The above generalized framework covers many interesting and important examples as special cases. These examples are investigated more closely in Section 5.

Example 2.1. (Robust tensor principal component analysis) For (robust) tensor PCA, the data observed is simply a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$. The basic assumption of tensor PCA is the existence of a low-rank tensor $\boldsymbol{\mathcal{T}}^*$, called the "signal", planted inside of $\boldsymbol{\mathcal{A}}$. See, e.g. (Zhang and Xia, 2018; Richard and Montanari, 2014) and references therein. The exact low-rank condition on the "signal" is sometimes stringent. Robust tensor PCA (Candès et al., 2011; Lu et al., 2016; Robin et al., 2020; Tao and Yuan, 2011) relaxes this condition by assuming that the "signal" is the sum of a low-rank

tensor $\boldsymbol{\mathcal{T}}^*$ and a sparse tensor $\boldsymbol{\mathcal{S}}^*$. With additional additive stochastic noise, the robust tensor PCA model assumes $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^* + \boldsymbol{\mathcal{Z}}$ with $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_\alpha)$ and $\boldsymbol{\mathcal{Z}}$ being a noise tensor having i.i.d. random centered sub-Gaussian entries. Given $\boldsymbol{\mathcal{A}}$, the goal is to estimate $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$. Without knowing the distribution of noise, a suitable loss function is $\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) := \frac{1}{2}\|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2$, which measures the goodness-of-fit by $\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}$ to data. In this case, the estimator $(\widehat{\boldsymbol{\mathcal{T}}}_\gamma, \widehat{\boldsymbol{\mathcal{S}}}_\gamma)$ is defined by

$$(\widehat{\boldsymbol{\mathcal{T}}}_\gamma, \widehat{\boldsymbol{\mathcal{S}}}_\gamma) := \underset{\boldsymbol{\mathcal{T}} \in \mathbb{U}_{\mathbf{r},\mu_1}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}}{\arg\min} \frac{1}{2}\|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2. \tag{2.4}$$

Clearly, as long as $\boldsymbol{\mathcal{Z}}$ is non-zero, the underlying truth $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ is generally not the minimizer of (2.4), but is the minimizer of the expected loss $\mathbb{E}\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}})$ where the expectation is w.r.t. the randomness of $\boldsymbol{\mathcal{Z}}$. Unlike the existing literature in tensor PCA (Richard and Montanari, 2014; Zhang and Xia, 2018; Liu et al., 2017), the solution to (2.4) is more robust to model mis-specifications and outliers. We remark that the least squares estimator as (2.4) is the maximum likelihood estimator (MLE) if $\boldsymbol{\mathcal{Z}}$ has i.i.d. standard normal entries. If the distribution of $\boldsymbol{\mathcal{Z}}$ is known, we can replace the objective function in (2.4) by its respective negative log-likelihood function, and the aforementioned properties of $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ continues to hold.

Example 2.2. (Learning low-rank structure from binary tensor) In many applications, the observed data $\boldsymbol{\mathcal{A}}$ is merely a binary tensor. Examples include the adjacency tensor in multi-layer networks (Jing et al., 2020; Paul and Chen, 2020), temporal networks (Nickel et al., 2011; Matias and Miele, 2016)), brain structural connectivity networks (Wang et al., 2019; Wang and Li, 2020) and etc. Following the Bernoulli tensor model proposed in (Wang and Li, 2020) or generalizing the 1-bit matrix completion model (Davenport et al., 2014), we assume that there exist $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_\alpha)$ satisfying

$$[\boldsymbol{\mathcal{A}}]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(p([\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*]_\omega)\big), \quad \forall \omega \in [d_1] \times \cdots \times [d_m], \tag{2.5}$$

where $p(\cdot) : \mathbb{R} \mapsto [0,1]$ is a suitable inverse link function. Popular choices of $p(\cdot)$ include the logistic link $p(x) = (1 + e^{-x/\sigma})^{-1}$ and probit link $p(x) = 1 - \Phi(-x/\sigma)$ where $\sigma > 0$ is a scaling parameter. We note that, due to potential symmetry in networks, the independence statement in (2.5) might only hold for a subset of its entries (e.g., off-diagonal entries in a single-layer undirected network). Compared with the exact low-rank Bernoulli tensor model (Wang and Li, 2020), our model (2.5) is more robust to model mis-specifications and outliers. For any pair $(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}}) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_{\gamma\alpha})$, a suitable loss function is the negative log-likelihood $\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) := -\sum_\omega \big([\boldsymbol{\mathcal{A}}]_\omega \log p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega) + \big(1 - [\boldsymbol{\mathcal{A}}]_\omega\big) \log(1 - p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega))\big)$. By maximizing the log-likelihood, we define

$$(\widehat{\boldsymbol{\mathcal{T}}}_\gamma, \widehat{\boldsymbol{\mathcal{S}}}_\gamma) := \underset{\boldsymbol{\mathcal{T}} \in \mathbb{U}_{\mathbf{r},\mu_1}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}}{\arg\min} -\sum_\omega \Big([\boldsymbol{\mathcal{A}}]_\omega \log p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega) + \big(1 - [\boldsymbol{\mathcal{A}}]_\omega\big) \log\big(1 - p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega)\big)\Big). \tag{2.6}$$

The underlying truth $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*)$ is generally not the minimizer to (2.6), but is the minimizer of the expected loss $\mathbb{E}\mathfrak{L}(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}})$ where the expectation is w.r.t. the randomness of $\boldsymbol{\mathcal{A}}$.

Example 2.3. (Community detection in hypergraph networks with outlier vertices) A special case of learning binary tensor is to uncover community structures in hypergraph networks. A hypergraph network $\mathcal{G}$ consists of n vertices $\mathcal{V} = [n]$ and a subset of hyperedges $\mathcal{H}$ where each hyperedge is a subset of $\mathcal{V}$. Hypergraph studies higher-order interactions among vertices, which often brings in new insights to scientific researches. See, e.g., (Ke et al., 2019; Benson et al., 2016; Ghoshdastidar and Dukkipati, 2017; Kim et al., 2018; Yuan et al., 2018) and references therein. Without loss of generality, we assume $\mathcal{G}$ is m-uniform meaning that each hyperedge $\mathbf{h} \in \mathcal{H}$ links exactly m vertices so that $\mathcal{G}$ can be encoded into an m-th order adjacency tensor $\boldsymbol{\mathcal{A}}$. Usually, the observed hypergraph is *undirected* implying that $\boldsymbol{\mathcal{A}}$ is a symmetric tensor. Suppose that these n vertices belong to K disjoint normal communities denoted by $\mathcal{V}_1, \cdots, \mathcal{V}_K$, and *one* outlier community denoted by $\mathcal{O}$ so that $\mathcal{V} = \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_K \cup \mathcal{O}$. Denote $n_o = |\mathcal{O}|$ the number of outlier vertices. For any $\omega = (i_1, \cdots, i_m)$ with $1 \le i_1 \le \cdots \le i_m \le n$, we assume

$$[\boldsymbol{\mathcal{A}}]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(\nu_n \cdot [\boldsymbol{\mathcal{C}}]_{k_1,\cdots,k_m}\big), \quad \text{if } i_j \in \mathcal{V}_{k_j} \text{ for } \forall j \in [m] \text{ and } k_j \in [K], \tag{2.7}$$

where $\nu_n \in (0, 1]$ describes the global network sparsity and $\boldsymbol{\mathcal{C}}$ is an m-th order $K \times \cdots \times K$ non-negative tensor. Tensor $\boldsymbol{\mathcal{C}}$ characterizes the connection intensity among vertices of normal communities, where we assume $\|\boldsymbol{\mathcal{C}}\|_{\ell_\infty} = 1$ for identifiability. Condition (2.7) means that the probability of generating the hyperedge $(i_1, \cdots, i_m)$ only depends on the community memberships of the normal vertices $\{i_j\}_{j=1}^m$. This is often referred to as the hypergraph stochastic block model (Ke et al., 2019; Ghoshdastidar and Dukkipati, 2017). On the other hand, if some vertex is an outlier, say $i_1 \in \mathcal{O}$, we assume that there exists a deterministic subset $\mathbb{N}_{i_1} \subset [n]^{m-1}$ such that $|\mathbb{N}_{i_1}| \le \alpha n^{m-1}$ and

$$\mathbb{P}([\boldsymbol{\mathcal{A}}]_\omega = 1) \text{ is arbitrary but } \text{supp}\big([\boldsymbol{\mathcal{A}}]_{i_1,:,\cdots,:}\big) \subset \mathbb{N}_{i_1} \textit{ almost surely}, \quad \text{if } i_1 \in \mathcal{O}. \tag{2.8}$$

Condition (2.8) dictates that an outlier vertex can connect other vertices (be them normal or outlier) in an arbitrary way, but it participates in at most $\lfloor \alpha n^{m-1} \rfloor$ hyperedges. Define a binary matrix (usually called *membership matrix*) $\mathbf{Z} \in \{0, 1\}^{n \times K}$ so that, for any $i \in \mathcal{V} \setminus \mathcal{O}$, $\mathbf{e}_i^\top \mathbf{Z} = \mathbf{e}_{k_i}^\top$ if[1] $i \in \mathcal{V}_{k_i}$; for any $i \in \mathcal{O}$, $\mathbf{e}_i^\top \mathbf{Z} = \mathbf{0}^\top$. Let $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}} \cdot [\![\mathbf{Z}, \cdots, \mathbf{Z}]\!]$. By condition (2.8), there exists an $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$ such that

$$[\boldsymbol{\mathcal{A}}]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}\big([\nu_n \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*]_\omega\big), \quad \forall \omega = (i_1, \cdots, i_m) \text{ with } 1 \le i_1 \le \cdots \le i_m \le n. \tag{2.9}$$

[1] We slightly abuse the notations so that $\mathbf{e}_i$ denotes the i-th canonical basis in $\mathbb{R}^n$, while $\mathbf{e}_{k_i}$ is the k_i-th canonical basis in $\mathbb{R}^K$.

We refer to (2.9) as the hypergraph generalized stochastic block model, or hGSBM in short. Similar models have been proposed for graph network analysis (see e.g., (Cai and Li, 2015; Hajek et al., 2016)). By observing the hypergraph $\mathcal{G}$ generated from (2.9), the goal is to recover the normal community memberships $\{\mathcal{V}_k\}_{k=1}^K$ and detect the set of outlier vertices $\mathcal{O}$. Since the singular vectors of $\nu_n \boldsymbol{\mathcal{T}}^\star$ directly reflect the community assignments, it suffices to estimate the latent low-rank tensor $\nu_n \boldsymbol{\mathcal{T}}^*$ from $\boldsymbol{\mathcal{A}}$. Unlike the existing literature in hypergraph network analysis, it is of great importance to decouple the sparse component $\boldsymbol{\mathcal{S}}^*$ for analyzing hGSBM.

3 Estimating by Fast Riemannian Optimization

Unfortunately, the optimization program (2.3) is highly non-convex. It can usually be optimized only locally, meaning that an iterative algorithm is designed to find a local minimum of (2.3) within a neighbourhood of good initializations.

Suppose that a pair[2] of initializations near the ground truth is provided. Our estimating procedure adopts a gradient-based iterative algorithm to search for a local minimum of the loss. Since the problem (2.3) is a constrained optimization, the major difficulty is on the enforcement of constraints during gradient descent updates. To ensure low-rankness, we apply the Riemannian gradient descent algorithm that is fast and simple to implement. Meanwhile, we enforce the sparsity constraint via a gradient-based pruning algorithm. The details of these algorithms are provided in Section 3.1 and 3.2, respectively.

3.1 Riemannian Gradient Descent Algorithm

Provided with $(\widehat{\boldsymbol{\mathcal{T}}}_l, \widehat{\boldsymbol{\mathcal{S}}}_l)$ at the l-th iteration, the *vanilla* gradient of the loss function is $\boldsymbol{\mathcal{G}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l)$. The naive gradient descent updates the low-rank part to $\widehat{\boldsymbol{\mathcal{T}}}_l - \beta \boldsymbol{\mathcal{G}}_l$ with a carefully chosen stepsize $\beta > 0$, and then projects it back into the set $\mathbb{M}_{\mathbf{r}}$. This procedure is sometimes referred to as the projected gradient descent (Chen et al., 2019). Oftentimes, the gradient $\boldsymbol{\mathcal{G}}_l$ has full ranks and thus the subsequent low-rank projection is computationally expensive. Observe that $\widehat{\boldsymbol{\mathcal{T}}}_l$ is an element in the smooth manifold $\mathbb{M}_{\mathbf{r}}$. Meanwhile, due to the smoothness of loss function, it is well recognized that the optimization problem can be solved by Riemannian optimization (Absil et al., 2009; Vandereycken, 2013; Edelman et al., 1998) on the respective smooth manifold, e.g. $\mathbb{M}_{\mathbf{r}}$. Therefore, instead of using the vanilla gradient $\boldsymbol{\mathcal{G}}_l$, it suffices to take the *Riemannian gradient*, which corresponds to the steepest descent of the loss but is restricted to the tangent space of $\mathbb{M}_{\mathbf{r}}$ at the point $\widehat{\boldsymbol{\mathcal{T}}}_l$. Fortunately, the Riemannian gradient is low-rank rendering amazing speed-up of

[2]We will show, in Section 4, that obtaining a good initialization for $\boldsymbol{\mathcal{S}}$ is, under suitable conditions, easy once a good initialization for $\boldsymbol{\mathcal{T}}$ is available.

the subsequent computations. See, e.g., (Wei et al., 2016; Cai et al., 2019b; Vandereycken, 2013; Kressner et al., 2014) for its applications in low-rank matrix and tensor problems.

An essential ingredient of Riemannian gradient descent is to project the vanilla gradient onto the tangent space of $\mathbb{M}_{\mathbf{r}}$. Let $\mathbb{T}_l$ denote the tangent space of $\mathbb{M}_{\mathbf{r}}$ at $\widehat{\boldsymbol{\mathcal{T}}}_l$. Suppose that $\widehat{\boldsymbol{\mathcal{T}}}_l$ admits a Tucker decomposition $\widehat{\boldsymbol{\mathcal{T}}}_l = \widehat{\boldsymbol{\mathcal{C}}}_l \cdot [\![\widehat{\mathbf{U}}_{l,1}, \cdots, \widehat{\mathbf{U}}_{l,m}]\!]$. The tangent space $\mathbb{T}_l$ (Cai et al., 2020c; Kressner et al., 2014) has an explicit form written as

$$\mathbb{T}_l = \left\{ \boldsymbol{\mathcal{D}}_l \times_{i\in[m]} \widehat{\mathbf{U}}_{l,i} + \sum_{i=1}^{m} \widehat{\boldsymbol{\mathcal{C}}}_l \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i : \boldsymbol{\mathcal{D}}_l \in \mathbb{R}^{\mathbf{r}}, \mathbf{W}_i \in \mathbb{R}^{d_i \times r_i}, \mathbf{W}_i^\top \widehat{\mathbf{U}}_{l,i} = \mathbf{0} \right\} \quad (3.1)$$

Clearly, all elements in $\mathbb{T}_l$ has their multi-linear ranks upper bounded by $2\mathbf{r}$. Given the vanilla gradient $\boldsymbol{\mathcal{G}}_l$, its projection onto $\mathbb{T}_l$ is defined by $\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l) := \arg\min_{\boldsymbol{\mathcal{X}}\in\mathbb{T}_l} \|\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^2$. Note that the summands in eq. (3.1) are all orthogonal to each other. This benign property allows fast computation for $\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l)$. More details on the computation of this projection can be found in Remark 3.1.

By choosing a suitable stepsize $\beta > 0$, the update by Riemannian gradient descent yields $\boldsymbol{\mathcal{W}}_l := \widehat{\boldsymbol{\mathcal{T}}}_l - \beta \mathcal{P}_{\mathbb{T}_l} \boldsymbol{\mathcal{G}}_l$. But $\boldsymbol{\mathcal{W}}_l$ may fail to be an element in $\mathbb{M}_{\mathbf{r}}$. To enforce the low-rank constraint, another key step in Riemannian optimization is the so-called *retraction* (Vandereycken, 2013), which projects a general tensor $\boldsymbol{\mathcal{W}}$ back to the smooth manifold $\mathbb{M}_{\mathbf{r}}$. This procedure amounts to a low-rank approximation of the tensor $\boldsymbol{\mathcal{W}}_l$. However, in addition, recall that we also need to enforce the spikiness (or incoherent) condition on the low-rank estimate. While in some particular applications (Cai et al., 2020b) or through sophisticated analysis (Cai et al., 2019a), it is possible to directly prove the incoherence property of $\widehat{\boldsymbol{\mathcal{T}}}_{l+1}$, this approach is, if not impossible, unavailable in our general framework. Instead, we first truncate $\boldsymbol{\mathcal{W}}_l$ entry-wisely by $\zeta_{l+1}/2$ for some carefully chosen threshold ζ_{l+1} and obtain $\widetilde{\boldsymbol{\mathcal{W}}}_l$. We then retract the truncated tensor $\widetilde{\boldsymbol{\mathcal{W}}}_l$ back to the manifold $\mathbb{M}_{\mathbf{r}}$, although the best low-rank approximation of a given tensor is generally NP-hard (Hillar and Lim, 2013). Fortunately, we show that a low-rank approximation of $\widetilde{\boldsymbol{\mathcal{W}}}_l$ by a simple higher order singular value decomposition (HOSVD) guarantees the convergence of Riemannian gradient descent algorithm. More specifically, for all $j \in [m]$, compute $\mathbf{V}_{l,j}$ which is the top-r_j left singular vectors of $\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}}_l)$. The HOSVD approximation of $\widetilde{\boldsymbol{\mathcal{W}}}_l$ with multi-linear ranks $\mathbf{r}$ is obtained by

$$\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\widetilde{\boldsymbol{\mathcal{W}}}_l) := (\widetilde{\boldsymbol{\mathcal{W}}}_l \times_{j=1}^{m} \mathbf{V}_{l,j}^\top) \cdot [\![\mathbf{V}_{l,1}, \cdots, \mathbf{V}_{l,m}]\!]. \quad (3.2)$$

Basically, retraction by HOSVD is the generalization of low-rank matrix approximation by singular value thresholding, although (3.2) is generally not the optimal low-rank approximation of $\widetilde{\boldsymbol{\mathcal{W}}}_l$. See, e.g. (Zhang and Xia, 2018; Xia and Zhou, 2019; Liu et al., 2017; Richard and Montanari, 2014) for

more explanations. Now put these two steps together and we define a trimming operator $\mathsf{Trim}_{\zeta,\mathbf{r}}$.

$$\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) := \mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\widetilde{\boldsymbol{\mathcal{W}}}), \quad \text{where } [\widetilde{\boldsymbol{\mathcal{W}}}]_\omega = \begin{cases} (\zeta/2)\cdot \mathrm{Sign}([\boldsymbol{\mathcal{W}}]_\omega), & \text{if } |[\boldsymbol{\mathcal{W}}]_\omega| > \zeta/2 \\ [\boldsymbol{\mathcal{W}}]_\omega, & \text{otherwise} \end{cases} \tag{3.3}$$

Equipped by the retraction (3.2) and the entry-wise truncation, the Riemannian gradient descent algorithm updates the low-rank estimate by

$$\widehat{\boldsymbol{\mathcal{T}}}_{l+1} = \mathsf{Trim}_{\zeta_{l+1},\mathbf{r}}(\boldsymbol{\mathcal{W}}_l), \tag{3.4}$$

with a properly chosen ζ_{l+1}. We show, in Lemma B.5, that the output tensor $\widehat{\boldsymbol{\mathcal{T}}}_{l+1}$ satisfies the spikiness condition (and incoherent condition), and moves closer to $\boldsymbol{\mathcal{T}}^*$ than $\widehat{\boldsymbol{\mathcal{T}}}_l$.

In the absence of sparse components, the updating rule (3.4) is the building block of Riemannian optimization for exact low-rank tensor estimation. See Algorithm 3 in Section 6.

Necessity of trimming We note that the trimming step is necessary because the spikiness condition is required for identifiability issue. For an exact low-rank tensor model (i.e., $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$), the spikiness condition is often unnecessary and thus the trimming treatment is not required.

Remark 3.1. *(Computation details of Riemannian gradient) By the definition of $\mathbb{T}_l$ as (3.1), there exist $\boldsymbol{\mathcal{D}}_l \in \mathbb{R}^{r_1\times\cdots\times r_m}$ and $\boldsymbol{\mathcal{W}}_i \in \mathbb{R}^{d_i\times r_i}$ satisfying*

$$\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l) = \boldsymbol{\mathcal{D}}_l \cdot [\![\widehat{\mathbf{U}}_{l,1},\cdots,\widehat{\mathbf{U}}_{l,m}]\!] + \sum\nolimits_{i=1}^m \widehat{\boldsymbol{\mathcal{C}}}_l \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i.$$

Observe that the summands are orthogonal to each other. Together with the definition of $\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l)$, it is easy to verify $\boldsymbol{\mathcal{D}}_l = \arg\min_{\boldsymbol{\mathcal{C}}} \|\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{C}}\cdot[\![\widehat{\mathbf{U}}_{l,1},\cdots,\widehat{\mathbf{U}}_{l,m}]\!]\|_{\mathrm{F}}$ which admits a closed-form solution $\boldsymbol{\mathcal{D}}_l = \boldsymbol{\mathcal{G}}_l\cdot[\![\widehat{\mathbf{U}}_{l,1}^\top,\cdots,\widehat{\mathbf{U}}_{l,m}^\top]\!]$. Similarly, the matrix $\mathbf{W}_i$ is the solution to

$$\mathbf{W}_i = \mathop{\arg\min}_{\mathbf{W}\in\mathbb{R}^{d_i\times r_i}, \mathbf{W}^\top\widehat{\mathbf{U}}_{l,i}=\mathbf{0}} \big\|\boldsymbol{\mathcal{G}}_l - \widehat{\boldsymbol{\mathcal{C}}}_l \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i\big\|_{\mathrm{F}}^2,$$

which also admits an explicit formula $\mathbf{W}_i = (\mathbf{I} - \widehat{\mathbf{U}}_{l,i}\widehat{\mathbf{U}}_{l,i}^\top)\mathcal{M}_i(\boldsymbol{\mathcal{G}}_l)\big(\otimes_{j\neq i}\widehat{\mathbf{U}}_{l,j}\big)^\top\mathcal{M}_i^\top(\widehat{\boldsymbol{\mathcal{C}}}_l)$.

3.2 Gradient Pruning

After updating the low-rank estimate, the next step is to update the estimate of sparse tensor $\boldsymbol{\mathcal{S}}^*$. Provided with the updated $\widehat{\boldsymbol{\mathcal{T}}}_l$ at the l-th iteration, an ideal estimator of the sparse tensor $\boldsymbol{\mathcal{S}}^*$ is to find $\arg\min_{\boldsymbol{\mathcal{S}}\in\mathbb{S}_{\gamma\alpha}} \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \boldsymbol{\mathcal{S}})$. Unfortunately, solving this problem is NP-hard for a general loss function.

Interestingly, if the loss function is entry-wise meaning that $\mathfrak{L}(\boldsymbol{\mathcal{T}}) = \sum_\omega \mathfrak{l}_\omega([\boldsymbol{\mathcal{T}}]_\omega)$ where $\mathfrak{l}_\omega(\cdot): \mathbb{R}\mapsto\mathbb{R}$ for each $\omega\in[d_1]\times\cdots\times[d_m]$, the computation of sparse estimate becomes tractable. Indeed,

we observe that a fast pruning algorithm on the gradient, under suitable conditions, suffices to guarantee appealing convergence performances. Basically, this algorithm finds the large entries of $|\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)|$ and chooses a suitable $\widehat{\boldsymbol{\mathcal{S}}}_l$ to prune the gradient on these entries. More exactly, given a tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$, we denote $|\boldsymbol{\mathcal{G}}|^{(n)}$ the value of its n-th largest entry in absolute value for $\forall n \in [d_1 \cdots d_m]$. Thus, $|\boldsymbol{\mathcal{G}}|^{(1)}$ denotes its largest entry in absolute value. The *level-α active indices* of $\boldsymbol{\mathcal{G}}$ is defined by

$$\mathsf{Level}\text{-}\alpha\ \mathsf{AInd}(\boldsymbol{\mathcal{G}}) := \Big\{\omega = (i_1, \cdots, i_m) : \big|[\boldsymbol{\mathcal{G}}]_\omega\big| \geq \max\nolimits_{j\in[m]} \big|\mathbf{e}_{i_j}^\top \mathcal{M}_j(\boldsymbol{\mathcal{G}})\big|^{(\lfloor \alpha d_j^- \rfloor)}\Big\}.$$

By definition, the *level-α active indices* of $\boldsymbol{\mathcal{G}}$ are those entries whose absolute value is no smaller than the $(1-\alpha)$-th percentile in absolute value on each of its corresponding slices. Clearly, for any $\boldsymbol{\mathcal{S}} \in \mathbb{S}_\alpha$, the support of $\boldsymbol{\mathcal{S}}$ belongs to the $\mathsf{Level}\text{-}\alpha\ \mathsf{AInd}(\boldsymbol{\mathcal{S}})$.

Provided with the trimmed $\widehat{\boldsymbol{\mathcal{T}}}_l$, we compute the vanilla gradient $\widehat{\boldsymbol{\mathcal{G}}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)$ so that $[\widehat{\boldsymbol{\mathcal{G}}}_l]_\omega = \mathfrak{l}'_\omega([\widehat{\boldsymbol{\mathcal{T}}}_l]_\omega)$ and find $\mathbb{J} = \mathsf{Level}\text{-}\alpha\ \mathsf{AInd}(\widehat{\boldsymbol{\mathcal{G}}}_l)$. Intuitively, the indices in $\mathbb{J}$ have the greatest potential in decreasing the value of loss function. The gradient pruning algorithm sets $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega = 0$ if $\omega \notin \mathbb{J}$. On the other hand, for $\omega \in \mathbb{J}$, ideally, the entry $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega$ is chosen to vanish the gradient in that $\mathfrak{l}'_\omega\big([\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l]_\omega\big) = 0$ (assuming that such an $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega$ is easily accessible). However, for functions with always-positive gradient (e.g. e^x), it is impossible to vanish the gradient. Generally, we choose a pruning parameter $\mathsf{k_{pr}} > 0$ and set

$$[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega := \arg\min\nolimits_{s:|s+[\widehat{\boldsymbol{\mathcal{T}}}_l]_\omega| \leq \mathsf{k_{pr}}} \big|\mathfrak{l}'_\omega([\widehat{\boldsymbol{\mathcal{T}}}_l]_\omega + s)\big|, \quad \forall \omega \in \mathbb{J}. \tag{3.5}$$

Basically, eq. (3.5) chooses $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega$ from the closed interval $\big[-\mathsf{k_{pr}} - [\widehat{\boldsymbol{\mathcal{T}}}_l]_\omega, \mathsf{k_{pr}} - [\widehat{\boldsymbol{\mathcal{T}}}_l]_\omega\big]$ to minimize the gradient. For a properly selected loss function $\mathfrak{l}_\omega(\cdot)$, searching for the solution $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega$ is usually fast. Moreover, for entry-wise square loss, the pruning parameter $\mathsf{k_{pr}}$ can be ∞ and $[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega$ has a closed-form solution. See Section 5.1 for more details.

The procedure of gradient pruning is summarized in Algorithm 1.

We note that Algorithm 1 is partially inspired by (Yi et al., 2016) for the fast estimation of matrix robust PCA. But our proposal is novel in three crucial aspects. First, our method can handle higher order tensors while (Yi et al., 2016) only treats matrices. The analysis for tensor-related algorithms is generally more challenging. Second, our framework allows stochastic noise rendering more realistic models and methods for statisticians while (Yi et al., 2016) only studies the noiseless case. Lastly, our method is able to deal with general non-linear relation between the observation and the underlying tensor parameters. By contrast, (Yi et al., 2016) merely focuses on a linear function.

Final algorithm Putting together the Riemannian gradient descent algorithm in Section 3.1 and the gradient pruning algorithm, we propose to solve problem (2.3) by Algorithm 2. The

Algorithm 1 Gradient Pruning for Sparse Estimate

Input: $\widehat{\boldsymbol{\mathcal{T}}}_l$ and parameters $\gamma > 1, \alpha, \mathsf{k_{pr}} > 0$

Calculate the gradient $\widehat{\boldsymbol{\mathcal{G}}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)$

Find $\mathbb{J} = \mathsf{Level}\text{-}\gamma\alpha\ \mathsf{AInd}(\widehat{\boldsymbol{\mathcal{G}}}_l)$

for $\omega \in [d_1] \times \cdots \times [d_m]$ **do**

$$[\widehat{\boldsymbol{\mathcal{S}}}_l]_\omega = \begin{cases} \text{by (3.5)}, & \text{if } \omega \in \mathbb{J} \\ 0, & \text{if } \omega \notin \mathbb{J} \end{cases}$$

end for

Output: $\widehat{\boldsymbol{\mathcal{S}}}_l$

algorithm alternatingly updates the low-rank estimate and the sparse estimate. We emphasize that the notations α and μ_1 in Algorithm 2 do not have to be exactly the model parameters α and μ_1 in eq. (2.2) and the Assumption 1, respectively. These notations are used in Algorithm 2 for the ease of exposition. In theory and applications, we only require them to be larger than the true model parameters α and μ_1, respectively.

Algorithm 2 Riemannian Gradient Descent and Gradient Pruning

Initialization: $\widehat{\boldsymbol{\mathcal{T}}}_0 \in \mathbb{U}_{\mathbf{r},\mu_1}$, stepsize β and parameters $\alpha, \gamma, \mu_1, \mathsf{k_{pr}} > 0$

Apply Algorithm 1 with input $\widehat{\boldsymbol{\mathcal{T}}}_0$ and parameters $\alpha, \gamma, \mathsf{k_{pr}}$ to obtain $\widehat{\boldsymbol{\mathcal{S}}}_0$

for $l = 0, 1, \cdots, l_{\max} - 1$ **do**

$\boldsymbol{\mathcal{G}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l + \widehat{\boldsymbol{\mathcal{S}}}_l)$

$\boldsymbol{\mathcal{W}}_l = \widehat{\boldsymbol{\mathcal{T}}}_l - \beta \mathcal{P}_{\mathbb{T}_l} \boldsymbol{\mathcal{G}}_l$

$\zeta_{l+1} = \frac{16}{7}\mu_1 \frac{\|\boldsymbol{\mathcal{W}}_l\|_{\mathrm{F}}}{\sqrt{d^*}}$

$\widehat{\boldsymbol{\mathcal{T}}}_{l+1} = \mathsf{Trim}_{\zeta_{l+1},\mathbf{r}}(\boldsymbol{\mathcal{W}}_l)$

Apply Algorithm 1 with input $\widehat{\boldsymbol{\mathcal{T}}}_{l+1}$ and parameters $\alpha, \gamma, \mathsf{k_{pr}}$ to obtain $\widehat{\boldsymbol{\mathcal{S}}}_{l+1}$

end for

Output: $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ and $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}$

4 General Convergence and Statistical Guarantees

In this section, we investigate the local convergence of Algorithm 2 in a general framework, and characterize the error of final estimates in terms of the gradient of loss function. Their applications on more specific examples are collected in Section 5. Our theory relies crucially on the regularity

of loss function. Recall that Algorithm 2 involves: routine 1. Riemannian gradient descent for the low-rank estimate; and routine 2. gradient pruning for the sparse estimate. It turns out that these two routines generally require different regularity conditions on the loss function, although these conditions can be equivalent in special cases (e.g. see Section 5.1). Recall that $\gamma > 1$ is the tuning parameter in Algorithm 2.

Assumption 2. *(Needed for Low-rank Estimate) There exist $b_l, b_u > 0$ such that the loss function $\mathfrak{L}(\cdot)$ is b_l-strongly convex and b_u-smooth in a subset $\mathbb{B}_2^* \subset \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$ meaning that*

$$\langle \boldsymbol{\mathcal{X}} - (\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*), \nabla\mathfrak{L}(\boldsymbol{\mathcal{X}}) - \nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\rangle \geq b_l \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \tag{4.1}$$

$$\|\nabla\mathfrak{L}(\boldsymbol{\mathcal{X}}) - \nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}} \leq b_u \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \tag{4.2}$$

for all $\boldsymbol{\mathcal{X}} \in \mathbb{B}_2^*$. Note that b_l and b_u may depend on $\mathbb{B}_2^*$.

Note that the explicit form of subset $\mathbb{B}_2^*$ in Assumption 2 is usually determined by the actual problems (see examples in Section 5). For the main theorem in this section (Theorem 4.1), we consider $\mathbb{B}_2^*$ to be a small neighbour around the truth $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$. In this case, Assumption 2 requires the loss function to be *locally* strongly convex and smooth, which is a typical assumption for generalized low-rank matrix and tensor estimation. See, e.g. (Wang et al., 2017; Zhao et al., 2015; Bhojanapalli et al., 2016; Li and Tang, 2017; Han et al., 2020) and references therein.

Assumption 3. *(Needed for Sparse Estimate) Suppose that $\mathfrak{L}$ is an entry-wise loss meaning $\mathfrak{L}(\boldsymbol{\mathcal{T}}) = \sum_\omega \mathfrak{l}_\omega([\boldsymbol{\mathcal{T}}]_\omega)$ where $\mathfrak{l}_\omega(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ for any $\omega \in [d_1] \times \cdots \times [d_m]$. There exist a subset $\mathbb{B}_\infty^* \subset \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$ and $b_l, b_u > 0$ such that*

$$\langle [\boldsymbol{\mathcal{X}}]_\omega - [\boldsymbol{\mathcal{Z}}]_\omega, \nabla\mathfrak{l}_\omega([\boldsymbol{\mathcal{X}}]_\omega) - \nabla\mathfrak{l}_\omega([\boldsymbol{\mathcal{Z}}]_\omega)\rangle \geq b_l |[\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}]_\omega|^2 \tag{4.3}$$

$$\big|\nabla\mathfrak{l}_\omega([\boldsymbol{\mathcal{X}}]_\omega) - \nabla\mathfrak{l}_\omega([\boldsymbol{\mathcal{Z}}]_\omega)\big| \leq b_u \big|[\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}]_\omega\big| \tag{4.4}$$

for $\forall \omega \in [d_1] \times \cdots \times [d_m]$ and any $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Z}} \in \mathbb{B}_\infty^*$. Similarly, b_l and b_u may depend on $\mathbb{B}_\infty^*$.

Assumption 3 dictates that the loss function is *entry-wise*, and on each entry the loss is strongly-convex and smooth within a subset $\mathbb{B}_\infty^*$. These conditions in Assumption 3 are unique in our framework, due to the existence of the sparse additive tensor $\boldsymbol{\mathcal{S}}^*$. The gradient pruning Algorithm 1 operates on entries of the gradients. Intuitively, entry-wise loss not only simplifies the computation but also helps characterize the performance of gradient pruning algorithm. If the sparse component is absent in our model, i.e. the underlying tensor is exactly low-rank, Assumption 3 will be unnecessary. See Section 6 for more details. We remark that the actual form of $\mathbb{B}_\infty^*$ can vary for solving different problems (see examples in Section 5). For simplicity, our Theorem 4.1 considers $\mathbb{B}_\infty^*$ to be a small neighbour around $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$.

Notice that the same parameters b_l, b_u are both used in Assumption 2 and Assumption 3. This slightly abuse of notations is for the ease of exposition. These parameters are not necessarily equal. For an entry-wise loss, condition (4.3) and (4.4) imply the condition (4.1) and (4.2), respectively. Therefore, Assumption 2 can be a by-product of Assumption 3, if we ignore the possible differences between the two neighbours $\mathbb{B}_2^*$ and $\mathbb{B}_\infty^*$. In this way, these two assumptions can be merged into one single assumption. However, we state them separately for several purposes. First, they highlight the differences of theoretical requirements between Riemannian gradient descent and gradient pruning algorithms. Second, the neighbours in these assumptions ($\mathbb{B}_2^*$ and $\mathbb{B}_\infty^*$) can be drastically different. Third, keeping them separate eases subsequent applications for special cases (e.g., for exact low-rank estimate in Section 6).

The signal strength $\underline{\lambda}$ is defined by $\underline{\lambda} := \lambda_{\min}(\boldsymbol{\mathcal{T}}^*) := \min_{j\in[m]} \lambda_{r_j}\big(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\big)$. Here $\lambda_r(\cdot)$ denotes the r-th largest singular value of a matrix. Thus, $\underline{\lambda}$ represents the smallest non-zero singular value among all the matricizations of $\boldsymbol{\mathcal{T}}^*$. Similarly, denote $\overline{\lambda} := \lambda_{\max}(\boldsymbol{\mathcal{T}}^*) := \max_{j\in[m]} \lambda_1\big(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\big)$ and define $\kappa_0 := \overline{\lambda}\underline{\lambda}^{-1}$ to be the condition number of $\boldsymbol{\mathcal{T}}^*$.

The error of final estimates produced by Algorithm 2 is characterized by the following quantities. Define

$$\mathsf{Err}_{2\mathbf{r}} := \sup_{\boldsymbol{\mathcal{M}}\in\mathbb{M}_{2\mathbf{r}}, \|\boldsymbol{\mathcal{M}}\|_{\mathrm{F}}\le 1} \big\langle \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*), \boldsymbol{\mathcal{M}}\big\rangle \tag{4.5}$$

and

$$\mathsf{Err}_\infty := \max\Big\{ \|\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\ell_\infty}, \min_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty}\le \mathsf{k}_{\mathsf{pr}}} \|\nabla \mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty} \Big\},$$

where k_{pr} is the tuning parameter in gradient pruning Algorithm 1. The quantity $\mathsf{Err}_{2\mathbf{r}}$ is typical in the aforementioned literature in exact low-rank matrix and tensor estimation. But the special quantity Err_∞ appears in our paper for investigating the performance of gradient pruning algorithm. The first term in Err_∞ comes from the gradient of loss function at the ground truth characterizing the stochastic error in many statistical models, while the second term is due to the setting of tuning parameter k_{pr}.

Theorem 4.1 displays the general performance bounds of Algorithm 2. For simplicity, we denote Ω^* the support of $\boldsymbol{\mathcal{S}}^*$, $\bar{r} = \max_j r_j, \bar{d} = \max_j d_j, \underline{d} = \min_j d_j, r^* = r_1\cdots r_m$ and $d^* := d_j d_j^{-1} = d_1\cdots d_m$. Let $\|\cdot\|_{\ell_\infty}$ denote the vectorized ℓ_∞-norm of tensors.

Theorem 4.1. *Let $\gamma > 1, k_{pr} > 0$ be the parameters used in Algorithm 2. Suppose that Assumption 1, 2 and 3 hold with $\mathbb{B}_2^* = \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \le C_{0,m}\underline{\lambda}, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$, $\mathbb{B}_\infty^* = \big\{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le k_\infty, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\big\}$ where $k_\infty = C_{1,m}\mu_1^{2m}(\bar{r}^{m-1}/\underline{d}^{m-1})^{1/2}\underline{\lambda} + k_{pr} + \|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty}$, $0.36 b_l (b_u^2)^{-1} \le 1$ and $b_u b_l^{-1} \le 0.4(\sqrt{\delta})^{-1}$ for some $\delta \in (0,1]$ and large absolute constants $C_{0,m}, C_{1,m} > 0$ depending only on m. Assume that*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda}\cdot\min\big\{\delta^2\bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\big\}$

(b) Signal-to-noise ratio:

$$\frac{\mathsf{Err}_{2\mathbf{r}}}{\underline{\lambda}} + \frac{\mathsf{Err}_{\infty}}{\underline{\lambda}} b_l^{-1}(b_u+1)(|\Omega^*| + \gamma\alpha d^*)^{1/2} \le c_{2,m} \cdot \min\left\{\delta^2 \bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\right\}$$

(c) Sparsity condition: $\alpha \le c_{3,m}(\kappa_0^{4m}\mu_1^{4m}\bar{r}^m b_u^4 b_l^{-4})^{-1}$ *and* $\gamma \ge 1 + c_{4,m}^{-1} b_u^4 b_l^{-4}$

where $c_{1,m}, c_{2,m}, c_{3,m}, c_{4,m} > 0$ *are small constants depending only on* m. *If the stepsize* $\beta \in [0.005 b_l/(b_u^2), 0.36 b_l/(b_u^2)]$, *then we have*

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &\le (1-\delta^2)\|\widehat{\boldsymbol{\mathcal{T}}}_{l} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{1,\delta}\mathsf{Err}_{2\mathbf{r}}^2 + C_{1,b_u,b_l}\left(|\Omega^*| + \gamma\alpha d^*\right)\mathsf{Err}_{\infty}^2 \\
\|\widehat{\boldsymbol{\mathcal{S}}}_{l+1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 &\le \frac{b_u^2}{b_l^2}\left(C_{2,m}\frac{1}{\gamma-1} + C_{3,m}(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\right)\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2}(|\Omega^*| + \gamma\alpha d)\mathsf{Err}_{\infty}^2
\end{aligned} \tag{4.6}$$

where $C_{1,\delta} = 6\delta^{-1}$ *and* $C_{1,b_u,b_l} = (C_2 + C_3 b_u + C_4 b_u^2) b_l^{-2}$ *for absolute constants* $C_1, \cdots, C_4 > 0$ *and* $C_{2,m}, C_{3,m} > 0$ *depending only on* m. *Therefore,*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le (1-\delta^2)^l\|\widehat{\boldsymbol{\mathcal{T}}}_{0} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_{1,\delta}\mathsf{Err}_{2\mathbf{r}}^2 + C_{1,b_u,b_l}\left(|\Omega^*| + \gamma\alpha d^*\right)\mathsf{Err}_{\infty}^2}{\delta^2} \tag{4.7}$$

for all $l \in [l_{\max}]$.

By eq. (4.6) of Theorem 4.1, as long as α is small (i.e., $\boldsymbol{\mathcal{S}}^*$ is sufficiently sparse) and γ is large, the error of $\widehat{\boldsymbol{\mathcal{S}}}_{l+1}$ is dominated by the error of $\widehat{\boldsymbol{\mathcal{T}}}_{l+1}$. This is interesting since it implies that the sparse estimate $\widehat{\boldsymbol{\mathcal{S}}}_{l+1}$ can indeed absorb outliers when $\boldsymbol{\mathcal{S}}^*$ has extremely large-valued entries. See also Theorem 4.3 for the ℓ_∞-norm upper bound of $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*$.

By eq. (4.7), after suitably chosen $l_{\max}$ iterations and treating b_l, b_u, δ as constants, we conclude with the following error bounds:

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le C_1\mathsf{Err}_{2\mathbf{r}}^2 + C_2(|\Omega^*| + \gamma\alpha d^*)\mathsf{Err}_{\infty}^2 \tag{4.8}$$

and

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \le \frac{\alpha(\mu_1\kappa_0)^{4m}\bar{r}^m(\gamma-1)+1}{\gamma-1} \cdot \left(C_5\mathsf{Err}_{2\mathbf{r}}^2 + C_6(|\Omega^*| + \gamma\alpha d^*)\mathsf{Err}_{\infty}^2\right) + C_7(|\Omega^*| + \gamma\alpha d^*)\mathsf{Err}_{\infty}^2.$$

There exist two types of error as illustrated on the RHS of (4.8). The first term $\mathsf{Err}_{2\mathbf{r}}^2$ comes from the model complexity of low-rank $\boldsymbol{\mathcal{T}}^*$, and the term $|\Omega^*|\mathsf{Err}_{\infty}^2$ is related to the model complexity of sparse $\boldsymbol{\mathcal{S}}^*$. These two terms both reflect the intrinsic complexity of our model. On the other hand, the last term $\gamma\alpha d^*\mathsf{Err}_{\infty}^2$ is a human-intervened complexity which originates from the tuning parameter γ in the algorithm design. If the cardinality of Ω^* happens to be of the same order as αd^* (it is the worse-case cardinality of Ω^* for $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$), the error bound is simplified into the following corollary. It is an immediate result from Theorem 4.1 and we hence omit the proof.

Corollary 4.2. *Suppose that the conditions of Theorem 4.1 hold and assume that* $|\Omega^*| \asymp \alpha d^*$. *Then for all* $l = 1, \cdots, l_{\max}$,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le (1-\delta^2)^l \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_{1,\delta}\mathsf{Err}_{2\mathbf{r}}^2 + C_{1,b_u,b_l}|\Omega^*|\mathsf{Err}_\infty^2}{\delta^2}.$$

If $(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)$ is the unconstrained global minimizer of the loss function, we have $\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*) = \mathbf{0}$ by the first order optimality condition. In these case, the error term $\mathsf{Err}_{2\mathbf{r}}$ is simply zero, and $\mathsf{Err}_\infty = \inf_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty} \le \mathsf{k_{pr}}} \|\nabla\mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty}$. If $\mathfrak{L}$ is entry-wisely convex, the term Err_∞ can be negligibly small if the tuning parameter $\mathsf{k_{pr}}$ is chosen large enough. Therefore, Corollary 4.2 suggests that Algorithm 2 can exactly recover (with negligible error) the ground truth after a finite number of iterations.

Remarks on the conditions of Theorem 4.1 Theorem 4.1 requires the initialization to be as close to $\boldsymbol{\mathcal{T}}^*$ as $o(\underline{\lambda})$, if b_l, b_u, κ_0 and $\bar{r}$ are all $O(1)$ constants. It is a common condition for non-convex methods for low-rank matrix and tensor related problems. Interestingly, we observe that, under the conditions of Theorem 4.1, it is unnecessary to specify an initialization for the sparse component $\boldsymbol{\mathcal{S}}^*$. Actually, according to (4.6), a good initialization $\widehat{\boldsymbol{\mathcal{S}}}_0$ is attainable by Algorithm 1 once $\widehat{\boldsymbol{\mathcal{T}}}_0$ is sufficiently good. Concerning the signal-to-noise ratio condition, Theorem 4.1 requires $\underline{\lambda}$ to dominate $\mathsf{Err}_{2\mathbf{r}}$ and $(|\Omega^*| + \gamma\alpha d^*)^{1/2}\mathsf{Err}_\infty$ if $b_l, b_u, \kappa_0, \bar{r} = O(1)$. This condition is mild and perhaps minimal. Otherwise, in view of eq. (4.8), simply estimating $\boldsymbol{\mathcal{T}}^*$ by zeros yields a sharper bound than (4.8). Oftentimes, the signal-to-noise ratio condition required by warm initialization is the primary bottleneck in tensor-related problems. See, e.g. (Zhang and Xia, 2018; Xia et al., 2021). The sparsity requirement on $\boldsymbol{\mathcal{S}}^*$ is also mild. Assuming $b_l, b_u, \kappa_0, \bar{r}, \mu_1 = O(1)$, Theorem 4.1 merely requires $\alpha \le c$ for a sufficiently small $c > 0$ which depends only on m. It suggests that Algorithm 2 allows a wide range of sparsity on $\boldsymbol{\mathcal{S}}^*$. Similarly, Theorem 4.1 only requires $\gamma \ge C$ for a sufficiently large $C > 0$ which depends on m only.

We now investigate the recovery of the support of $\boldsymbol{\mathcal{S}}^*$. As mentioned after Corollary 4.2, in the case $\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*) = \mathbf{0}$ such that the sparse component $\boldsymbol{\mathcal{S}}^*$ is exactly recoverable, the support of $\boldsymbol{\mathcal{S}}^*$ is also recoverable. However, if there exists stochastic noise or random sampling such that $\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*) \ne \mathbf{0}$, Algorithm 2 usually over-estimates the size of the support of $\boldsymbol{\mathcal{S}}^*$ since the Level-$\gamma\alpha$ active indices are used for a γ strictly greater than 1. Fortunately, we have the following results controlling the sup-norm error of $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}$.

Theorem 4.3. *Suppose conditions of Theorem 4.1 hold,* $b_l, b_u = O(1)$, $|\Omega^*| \asymp \alpha d^*$ *and* $l_{\max}$ *is chosen such that (4.8) holds. Then,*

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le C_{1,m}\kappa_0^{2m}\mu_1^{2m}\Big(\frac{\bar{r}^m}{\underline{d}^{m-1}}\Big)^{1/2} \cdot \big(\mathsf{Err}_{2\mathbf{r}} + (\gamma|\Omega^*|)^{1/2}\mathsf{Err}_\infty\big) + C_{2,m}\mathsf{Err}_\infty,$$

where $C_{1,m}, C_{2,m} > 0$ only depend on m.

If the non-zero entries of $\boldsymbol{\mathcal{S}}^*$ satisfy $|[\boldsymbol{\mathcal{S}}^*]_\omega| > 2\delta^*$ for all $\omega \in \Omega^*$ where $\delta^* := C_{1,m}\mu_1^{2m}\kappa_0^{2m}(\bar{r}^m/\underline{d}^{m-1})^{1/2}$. $(\mathsf{Err}_{2\mathbf{r}} + (\gamma|\Omega^*|)^{1/2}\mathsf{Err}_\infty) + C_{2,m}\mathsf{Err}_\infty$, we obtain $\widehat{\boldsymbol{\mathcal{S}}}$ by a final-stage hard thresholding on $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}$ so that

$$[\widehat{\boldsymbol{\mathcal{S}}}]_\omega := [\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}]_\omega \cdot \mathbb{1}\big(|[\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}]_\omega| > \delta^*\big), \quad \omega \in [d_1] \times \cdots \times [d_m].$$

By Theorem 4.3, we get $\mathrm{supp}(\widehat{\boldsymbol{\mathcal{S}}}) = \Omega^*$ and thus recovering the support of $\boldsymbol{\mathcal{S}}^*$.

More specific examples and applications are studied in Section 5. Numerical results in Section 7 also confirm our findings.

Convergence behavior of the sparse estimate Although bound (4.6) seems to imply that the error $\|\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ decreases as l increases, the error of sparse estimates roughly stays the same as the initial estimate $\widehat{\boldsymbol{\mathcal{S}}}_0$. Due to the gradient pruning algorithm, for any $\omega \in \mathrm{supp}(\widehat{\boldsymbol{\mathcal{S}}}_0)$, the gradient of loss function $[\nabla(\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0)]_\omega$ is negligible. Consequently, on the support of $\widehat{\boldsymbol{\mathcal{S}}}_0$, the low-rank estimate $\widehat{\boldsymbol{\mathcal{T}}}_l$ is roughly unchanged. This implies that the corresponding entries of $\widehat{\boldsymbol{\mathcal{S}}}_l$ are also roughly unchanged.

5 Applications

We now apply the established results in Section 4 to more specific examples and elaborate the respective statistical performances. Some examples have been briefly introduced in Section 2. Our framework certainly covers many other interesting examples but we do not intend to exhaust them.

5.1 Robust sub-Gaussian Tensor PCA

As introduced in Example 2.1, the goal of tensor PCA is to extract low-rank *signal* from a noisy tensor observation $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$. Tensor PCA has been proven effective in learning hidden components in Gaussian mixture models (Anandkumar et al., 2014), denoising electron microscopy data (Zhang et al., 2020b) and in the inference of spatial and temporal patterns of gene regulation during brain development (Liu et al., 2017). Tensor PCA has been intensively studied from various aspects by both statistics and machine learning community resulting into a rich literature (e.g. (Richard and Montanari, 2014; Zhang and Xia, 2018; Xia et al., 2020; Arous et al., 2019; De Lathauwer et al., 2000; Dudeja and Hsu, 2020; Hopkins et al., 2015; Huang et al., 2020; Vannieuwenhoven et al., 2012)).

However, the exact low-rank assumption is often too stringent and the standard tensor PCA is sensitive to model mis-specifications and outliers. To develop statistically more robust methods

for tensor PCA, we assume that the underlying *signal* is an approximately low-rank tensor. More exactly, assume

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^* + \boldsymbol{\mathcal{Z}} \tag{5.1}$$

where $\boldsymbol{\mathcal{T}}^* \in \mathbb{U}_{\mathbf{r},\mu_1}$ is low-rank, $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$ is heterogeneous signal or a sparse corrupt, $\boldsymbol{\mathcal{Z}}$ is a noise tensor having i.i.d. centered sub-Gaussian entries. We refer to the model (5.1) as *robust tensor PCA* where the underlying *signal* is $\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*$. The sparse component $\boldsymbol{\mathcal{S}}^*$ accounts for potential model mis-specifications and outliers.

Due to the linearity, we use the loss function $\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) := \frac{1}{2}\|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2$ for estimating the underlying signal in model (5.1). Clearly, this loss is an entry-wise loss function, and satisfies the strongly-convex and smoothness conditions of Assumption 2 and 3 with constants $b_l = b_u = 1$ within any subsets $\mathbb{B}_2^*$ and $\mathbb{B}_\infty^*$, or simply $\mathbb{B}_2^* = \mathbb{B}_\infty^* = \mathbb{R}^{d_1 \times \cdots \times d_m}$. As a result, Theorem 4.1 and Theorem 4.3 are readily applicable by choosing $\delta = 0.15$, and setting the tuning parameter $\mathsf{k_{pr}} = \infty$ in the gradient pruning Algorithm 1.

Theorem 5.1. *Suppose Assumption 1 holds and there exists $\sigma_z > 0$ such that $\mathbb{E}\exp\{t[\boldsymbol{\mathcal{Z}}]_\omega\} \le \exp\{t^2\sigma_z^2/2\}$ for $\forall t \in \mathbb{R}$ and $\forall \omega \in [d_1] \times \cdots \times [d_m]$. Let $r^* = r_1 \cdots r_m$ and $\gamma > 1$ be the tuning parameter in Algorithm 2. Assume $|\Omega^*| \asymp \alpha d^*$ and*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda} \cdot (\kappa_0^{2m}\bar{r}^{1/2})^{-1}$

(b) Signal-to-noise ratio: $\underline{\lambda}/\sigma_z \ge C_{1,m}\kappa_0^{2m}\bar{r}^{1/2} \cdot (\bar{d}\bar{r} + r^* + \gamma|\Omega^*|\log\bar{d})^{1/2}$

(c) Sparsity condition: $\alpha \le c_{2,m}(\mu_1^{4m}\kappa_0^{4m}\bar{r}^m)^{-1}$ and $\gamma \ge 1 + c_{3,m}^{-1}$

where $c_{1,m}, c_{2,m}, c_{3,m}, C_{1,m} > 0$ are constants depending only on m. If the step size $\beta \in [0.005, 0.36]$, then after $l_{\max} > 1$ iterations, with probability at least $1 - \bar{d}^{-2}$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le& 0.98^{l_{\max}}\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{2,m}(\bar{d}\bar{r} + r^* + \gamma|\Omega^*|\log\bar{d})\sigma_z^2 \\
\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \le& \big(C_{3,m}\alpha\bar{r}^m\mu_1^{4m}\kappa_0^{4m} + C_{4,m}(\gamma-1)^{-1}\big) \cdot \|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{5,m}\sigma_z^2 \cdot \gamma|\Omega^*|\log\bar{d}
\end{aligned} \tag{5.2}$$

where $C_{2,m}, C_{3,m}, C_{4,m}, C_{5,m} > 0$ are constants depending only on m. Moreover, If $l_{\max}$ is chosen large enough such that the second term on RHS of (5.2) dominates and assume $\mu_1^{4m}\kappa_0^{4m}\bar{r}^m(\bar{r}\bar{d} + r^*) \le C_{9,m}\underline{d}^{m-1}$, we get with probability at least $1 - \bar{d}^{-2}$ that

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le \Big(C_{6,m}\kappa_0^{2m}\mu_1^{2m}\bar{r}^{m/2}(\gamma|\Omega^*|)^{1/2}/\underline{d}^{(m-1)/2} + C_{7,m}\Big) \cdot \sigma_z \log^{1/2}\bar{d}$$

where $C_{6,m}, C_{7,m} > 0$ are constants depending only on m.

Theorem 5.1 has several interesting implications. If the noise is absent meaning $\sigma_z = 0$, eq. (5.2) implies that, for an arbitrary $\varepsilon > 0$, after $l_{\max} \asymp \log(\varepsilon^{-1})$ iterations, Algorithm 2 outputs a

$\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ satisfying $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} = O(\varepsilon)$. Therefore, Algorithm 2 can exactly recover the low-rank and sparse component, separately.

On the other hand, if $\sigma_z > 0$ and $l_{\max} \asymp \log(\underline{\lambda}\sigma_z^{-1})$, eq. (5.2) implies that Algorithm 2 produces, with probability at least $1 - \bar{d}^{-2}$,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le C_{2,m}\sigma_z^2(\bar{d}\bar{r} + r^* + \gamma|\Omega^*|\log\bar{d}). \tag{5.3}$$

Since the intrinsic model complexity is of order $\bar{d}\bar{r} + r^* + |\Omega^*|$, the bound (5.3) is sharp up to logarithmic factors. Similar bounds also hold for $\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2$. In addition, if $\mu_1^{4m}\kappa_0^{4m}\bar{r}^m(\bar{d}\bar{r} + r^* + \gamma|\Omega^*|) \le C_{9,m}\underline{d}^{m-1}$, we get with probability at least $1 - \bar{d}^{-2}$,

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le C_{7,m}\sigma_z\log^{1/2}\bar{d}. \tag{5.4}$$

Bound (5.4) is nearly optimal. To see it, consider the simpler model that $\boldsymbol{\mathcal{T}}^* = \mathbf{0}$. Then, it is equivalent to estimate a sparse tensor from the data $\boldsymbol{\mathcal{S}}^* + \boldsymbol{\mathcal{Z}}$. Without further information, $O(\sigma_z\log^{1/2}\bar{d})$ is the best sup-norm performance one can expect in general.

Comparison with existing literature In (Lu et al., 2016, 2019), the authors studied noiseless robust tensor PCA assuming low tubal rank and proved that a convex program can exactly recover the underlying parameters. They used the averaged nuclear norms of the matrix slices of $\boldsymbol{\mathcal{T}}$ as a convex relaxation for the tubal rank. Their method is not applicable to low Tucker-rank tensors, and there exists no statistical guarantee for the noisy setting. Interestingly, by setting $|\Omega^*| = 0$, our Theorem 5.1 degrades to well-established results for tensor PCA in the literature. In this case, eq. (5.2) coincides with the convergences of higher order orthogonal iteration in (Zhang and Xia, 2018) and regularized jointly gradient descent in (Han et al., 2020).

Comparison with prior works on *matrix* robust PCA If $m = 2$, problem (5.1) reduces to the matrix robust PCA. In the noiseless case $\sigma_z = 0$, the sparsity requirement in Theorem 5.1 almost matches the best known bound in (Chandrasekaran et al., 2011; Candès et al., 2011; Chen et al., 2013; Hsu et al., 2011; Ganesh et al., 2010; Yi et al., 2016; Chen et al., 2020b; Netrapalli et al., 2014; Cherapanamjeri et al., 2017) amongst others. Unlike many prior works, we impose no sampling distribution on the support of $\boldsymbol{\mathcal{S}}^*$. In the noisy case, the error bound (5.3) improves those in (Zhou et al., 2010; Wong and Lee, 2017; Koltchinskii et al., 2011), assuming that $\bar{d}\bar{r}$ dominates the others in RHS of (5.3). More recently, a sharper bound for the low-rank estimate is derived in (Chen et al., 2020b) by a more sophisticated analysis and more stringent model on $\boldsymbol{\mathcal{S}}^*$. Nonetheless, we note that our framework does not consider missing values, which we leave as a future work.

5.2 Tensor PCA under Heavy-tailed Noise

Most aforementioned literature in Section 5.1 on tensor PCA focus on sub-Gaussian or sub-exponential noise (Chen et al., 2020a). Nowadays, heavy-tailed noise routinely arise in diverse fields. However, the performances of most existing approaches for tensor PCA significantly deteriorate when noise have heavy tails. On the technical front, concentration inequalities of random tensors fail to hold under heavy-tailed distributions. Consequently, tensor PCA under heavy-tailed noise is deemed difficult.

Interestingly, tensor PCA under heavy-tailed noise can be regarded as a special case of robust tensor PCA. Suppose that the observed tensorial data $\boldsymbol{\mathcal{A}}$ satisfies $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{Z}}$ with $\boldsymbol{\mathcal{T}}^* \in \mathbb{U}_{\mathbf{r},\mu_1}$. The noise tensor $\boldsymbol{\mathcal{Z}}$ satisfies the following tail assumption.

Assumption 4. *(θ-tailed noise) The entries of $\boldsymbol{\mathcal{Z}}$ are i.i.d. with $\mathbb{E}[\boldsymbol{\mathcal{Z}}]_\omega = 0$ and $\mathrm{Var}([\boldsymbol{\mathcal{Z}}]_\omega) \le \sigma_z^2$. There exist $\theta > 2$ such that $\mathbb{P}(|[\boldsymbol{\mathcal{Z}}]_\omega| \ge \sigma_z \cdot t) \le t^{-\theta}$ for all $t > 1$.*

Assumption 4 implies that $\mathbb{E}|[\boldsymbol{\mathcal{Z}}]_\omega|^{\theta'} < \infty$ for any $0 < \theta' < \theta$. If θ is only moderately large (e.g., $\theta = 3$), many entries of $\boldsymbol{\mathcal{Z}}$ can have large magnitudes such that the typical concentration properties of $\boldsymbol{\mathcal{Z}}$ (e.g., the bounds of $\mathsf{Err}_{2\mathbf{r}}$ and Err_∞ in Section 5.1) disappear.

Fix any $\alpha > 1$, we decompose $[\boldsymbol{\mathcal{Z}}]_\omega = [\boldsymbol{\mathcal{S}}_\alpha]_\omega + [\widetilde{\boldsymbol{\mathcal{Z}}}]_\omega$ such that

$$[\widetilde{\boldsymbol{\mathcal{Z}}}]_\omega = \mathbb{1}(|[\boldsymbol{\mathcal{Z}}]_\omega| \ge \alpha\sigma_z) \cdot (\alpha\sigma_z)\mathrm{sign}([\boldsymbol{\mathcal{Z}}]_\omega) + \mathbb{1}(|[\boldsymbol{\mathcal{Z}}]_\omega| < \alpha\sigma_z) \cdot [\boldsymbol{\mathcal{Z}}]_\omega, \quad \forall \omega \in [d_1] \times \cdots \times [d_m].$$

By definition, the entry $[\boldsymbol{\mathcal{S}}_\alpha]_\omega \neq 0$ if and only if $|[\boldsymbol{\mathcal{Z}}]_\omega| > \alpha\sigma_z$. Now, we write

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}_\alpha + \widetilde{\boldsymbol{\mathcal{Z}}} \tag{5.5}$$

The model (5.5) satisfies the robust PCA model (5.1) under mild conditions.

Lemma 5.2. *Suppose Assumption 4 holds and the distribution of $[\boldsymbol{\mathcal{Z}}]_\omega$ is symmetric. For any $\alpha > 1$, we have $\mathbb{E}\widetilde{\boldsymbol{\mathcal{Z}}} = \mathbf{0}$. There exists an event $\mathfrak{E}_1$ with $\mathbb{P}(\mathfrak{E}_1) \ge 1 - \bar{d}^{-2}$ such that $\boldsymbol{\mathcal{S}}_\alpha \in \mathbb{S}_{\alpha'}$ in the event $\mathfrak{E}_1$ where $\alpha' = \max\{2\alpha^{-\theta}, 10(\bar{d}/d^*)\log(m\bar{d}^3)\}$.*

By Lemma 5.2, in the event $\mathfrak{E}_1$, model (5.5) satisfies the robust sub-Gaussian PCA model (5.1) such that $\boldsymbol{\mathcal{T}}^* \in \mathbb{U}_{\mathbf{r},\mu_1}, \boldsymbol{\mathcal{S}}_\alpha \in \mathbb{S}_{\alpha'}$. Meanwhile, the entries of $\widetilde{\boldsymbol{\mathcal{Z}}}$ are sub-Gaussian for being uniformly bounded by $\alpha\sigma_z$. Therefore, conditioned on $\mathfrak{E}_1$, Theorem 5.1 is readily applicable and we end up with the following results.

Theorem 5.3. *Suppose Assumption 1 and the conditions of Lemma 5.2 hold. Choose $\alpha \asymp (d^*/\bar{d})^{1/\theta}, \gamma > 1$ as the tuning parameters in Algorithm 2. Assume $(\bar{d}/d^*)\log(m\bar{d}^3) \le c_{2,m}(\mu_1^{4m}\kappa_0^{4m}\bar{r}^m)^{-1}$, $\gamma \ge 1 + c_{3,m}^{-1}$ and*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda} \cdot (\kappa_0^{2m}\bar{r}^{1/2})^{-1}$

(b) Signal-to-noise ratio: $\underline{\lambda}/(\alpha\sigma_z) \ge C_{1,m}\gamma^{1/2}\kappa_0^{2m}\bar{r}^{1/2} \cdot \big(\bar{d}\bar{r} + r^* + \bar{d}\log(m\bar{d})\big)^{1/2}$

where $c_{1,m}, c_{2,m}, c_{3,m}, C_{1,m} > 0$ *are constants depending only on* m*. If the step size* $\beta \in [0.005, 0.36]$*, then after* $l_{\max} > 1$ *iterations, with probability at least* $1 - 2\bar{d}^{-2}$*, we have*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le 0.98^{l_{\max}}\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{2,m}\gamma\big(\bar{d}\bar{r} + r^* + \bar{d}\log(m\bar{d})\big)\alpha^2\sigma_z^2, \tag{5.6}$$

where $C_{2,m} > 0$ *is a constant depending only on* m*.*

By Theorem 5.3 and (5.6), if $\kappa_0, \gamma = O(1)$, $d_j \asymp d$ with $\alpha \asymp d^{(m-1)/\theta}$ and $l_{\max}$ is properly chosen, we get with probability at least $1 - 2d^{-2}$ that

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le C_{2,m}\big(d\bar{r} + r^* + d\log(md)\big)d^{2(m-1)/\theta} \cdot \sigma_z^2. \tag{5.7}$$

The bound (5.7) decreases as θ increases implying that the final estimate $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ becomes more accurate as the noise tail gets lighter. Moreover, if Assumption 4 holds with $\theta = 2(m-1)\log d$ so that $d^{2(m-1)/\theta} = O(1)$, bound (5.7) implies $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2/\sigma_z^2 = O\big(d\bar{r} + r^* + d\log(md)\big)$ which is sharp up to logarithmic factors. Similarly as Theorem 5.1, under Assumption 1, it is possible to derive an ℓ_∞-norm bound for $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*$.

Tensor PCA under heavy-tailed noise is largely unknown in the literature. While the bound (5.7) is novel and interesting, we suspect that this rate is generally sub-optimal. Indeed, matrix PCA ($m = 2$) under heavy-tailed noise has been intensively studied. See, e.g., (Fan et al., 2016; Minsker, 2018; Alquier et al., 2019; Elsener and van de Geer, 2018). For instance, it is proved in (Minsker, 2018) that, under suitable conditions, an $O(d\bar{r})$ rate is attainable as long as Assumption 4 holds with $\theta = 2$.

5.3 Learning from Binary Tensor

We now apply Algorithm 2 to learning the low-rank plus sparse tensor from a binary tensorial observation $\boldsymbol{\mathcal{A}} \in \{0,1\}^{d_1 \times \cdots \times d_m}$. Following the Bernoulli tensor model in Example 2.2, we assume

$$[\boldsymbol{\mathcal{A}}]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(p([\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*]_\omega)\big), \quad \forall \omega \in [d_1] \times \cdots \times [d_m],$$

where $p(x)$ is the inverse link function and $(\boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{S}}^*) \in (\mathbb{U}_{\mathbf{r},\mu_1}, \mathbb{S}_\alpha)$. The loss function is the *negative* log-likelihood (without loss of generality, we set the scale parameter $\sigma = 1$ for ease of exposition)

$$\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) = -\sum\nolimits_\omega \Big([\boldsymbol{\mathcal{A}}]_\omega \log p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega) + (1 - [\boldsymbol{\mathcal{A}}]_\omega)\log\big(1 - p([\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}]_\omega)\big)\Big). \tag{5.8}$$

The RHS of (5.8) is an entry-wise loss, and Assumption 2 and 3 are determined by the entry-wise second order derivatives. For $\forall \zeta > 0$, define

$$b_{u,\zeta} := \max\left\{\sup_{|x|\le\zeta} \frac{(p'(x))^2}{p^2(x)} - \frac{p''(x)}{p(x)}, \sup_{|x|\le\zeta} \frac{(p'(x))^2}{(1-p(x))^2} + \frac{p''(x)}{1-p(x)}\right\}$$

$$b_{l,\zeta} := \min\left\{ \inf_{|x|\le\zeta} \frac{(p'(x))^2}{p^2(x)} - \frac{p''(x)}{p(x)},\ \inf_{|x|\le\zeta} \frac{(p'(x))^2}{(1-p(x))^2} + \frac{p''(x)}{1-p(x)} \right\}$$

Assuming $b_{l,\zeta}, b_{u,\zeta} > 0$, then the loss function (5.8) satisfies Assumption 2 and 3 with constants $b_{l,\zeta}$ and $b_{u,\zeta}$ for $\mathbb{B}_2^* = \mathbb{B}_\infty^* = \{\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}} : \|\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}\|_{\ell_\infty} \le \zeta, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}, \boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}\}$ (more precisely, the low-rank and sparse conditions are unnecessary).

Notice that $b_{l,\zeta}$ and $b_{u,\zeta}$ can be extremely sensitive to large ζ. For instance (Wang and Li, 2020), we have

$$b_{l,\zeta} = \begin{cases} \frac{e^\zeta}{(1+e^\zeta)^2}, & \text{if } p(x) = (1+e^{-x})^{-1} \\ \gtrsim \frac{\zeta+1/6}{\sqrt{2\pi}} e^{-\zeta^2}, & \text{if } p(x) = \Phi(x) \end{cases} \quad \text{and} \quad b_{u,\zeta} = \begin{cases} \frac{1}{4}, & \text{if } p(x) = (1+e^{-x})^{-1} \\ \ge 0.6, & \text{if } p(x) = \Phi(x) \end{cases}$$

implying that $b_{u,\zeta} b_{l,\zeta}^{-1}$ increases very fast as ζ becomes larger. Therefore, the underlying ground truth should satisfy $\|\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le \zeta$ for a small ζ. Toward that end, we impose the following Assumption.

Assumption 5. *There exists a small $\zeta > 0$ such that $\|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} \le \zeta/2$, $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1 and its largest singular value $\overline{\lambda} \le \mu_1^{-1}\zeta/2 \cdot (\sqrt{d^*/\bar{r}})$ where $r^* = r_1 \cdots r_m$ and $d^* = d_1 \cdots d_m$.*

Assumption 5 implies $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \zeta/2$ so that $\|\boldsymbol{\mathcal{S}}^* + \boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \zeta$. Meanwhile, we shall guarantee that the iterates $(\widehat{\boldsymbol{\mathcal{T}}}_l, \widehat{\boldsymbol{\mathcal{S}}}_l)$ produced by our algorithm satisfy $\|\widehat{\boldsymbol{\mathcal{T}}}_l\|_{\ell_\infty}, \|\widehat{\boldsymbol{\mathcal{S}}}_l\|_{\ell_\infty} = O(\zeta)$. This can be guaranteed by the following lemma.

Lemma 5.4. *Suppose that Assumption 1 and 5 hold. Given any $\boldsymbol{\mathcal{W}}$ such that $\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, if we choose $\eta = 16\mu_1\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}/(7\sqrt{d^*})$, then $\|\mathsf{Trim}_{\eta,\mathbf{r}}(\boldsymbol{\mathcal{W}})\|_{\ell_\infty} \le (9\zeta/16) \cdot (\kappa_0\mu_1)^m$.*

By Lemma 5.4, if $\kappa_0\mu_1, m = O(1)$ and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, the trimming operator guarantees $\|\mathsf{Trim}_{\zeta,\mathbf{r}}(\widehat{\boldsymbol{\mathcal{T}}}_l)\|_{\ell_\infty} = O(\zeta)$. Equipped with (3.3) and by setting $\mathsf{k_{pr}} = C_1\zeta$ for some absolute $C_1 > 1$ depending only on $\kappa_0\mu_1$ and m, we apply Algorithm 2 to minimize the RHS of (5.8).

Similarly, the error of final estimate relies on $\mathsf{Err}_{2\mathbf{r}}$ and Err_∞, both of which are related to the gradient of loss (5.8). Denote

$$L_\zeta := \sup_{|x|\le\zeta} \left| \frac{p'(x)}{p(x)\big(1-p(x)\big)} \right|.$$

Suppose $\mathsf{k_{pr}} := \zeta'/2 > \zeta$, by definition of Err_∞, we have

$$\mathsf{Err}_\infty \le \max\left\{ L_\zeta, \min_{|x|\le\mathsf{k_{pr}}} \left| \frac{p'(x)}{p(x)\big(1-p(x)\big)} \right| \right\} \le L_\zeta. \tag{5.9}$$

We note that the error of final estimate depends on L_ζ rather than $L_{\zeta'}$, namely the true ζ rather than the algorithm-relevant ζ'.

Theorem 5.5. *Let $\gamma > 1, k_{pr} = \zeta'/4 := C_1\zeta$ be the parameters used in Algorithm 2 for a constant $C_1 > 1$ depending only on $\kappa_0\mu_1$ and m via Lemma 5.4. Suppose Assumption 1 and 5 hold. Assume $|\Omega^*| \asymp \alpha d^*$, $0.36 b_{l,\zeta'} b_{u,\zeta'}^{-2} \le 1$, $b_{u,\zeta'} b_{l,\zeta'}^{-1} \le 0.4(\sqrt{\delta})^{-1}$ for some $\delta \in (0,1]$, and*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda} \cdot \min\{\delta^2 \bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\}$ and $\|\widehat{\boldsymbol{\mathcal{T}}}_0\|_{\ell_\infty} \le c_{2,m}\zeta$

(b) Signal-to-noise ratio:

$$\underline{\lambda} \cdot \min\{\delta^2 \bar{r}^{-1/2}, (\kappa_0^{2m}\bar{r}^{1/2})^{-1}\} \ge C_{2,m}\Big(d_1 \bar{r} + r^* + \gamma|\Omega^*|\frac{1 + b_{u,\zeta'}}{b_{l,\zeta'}}\Big) \cdot L_\zeta$$

(c) Sparsity condition: $\alpha \le c_{3,m} b_{l,\zeta'}^4 (b_{u,\zeta'}^4 \kappa_0^{4m} \mu_1^{4m} \bar{r}^m)^{-1}$ and $\gamma \ge 1 + c_{4,m}^{-1} b_{u,\zeta'}^4 b_{l,\zeta'}^{-4}$

where $c_{1,m}, c_{2,m}, c_{3,m}, c_{4,m}, C_{2,m} > 0$ are some constants depending on m only. If the stepsize $\beta \in [0.005 b_{l,\zeta'}/(b_{u,\zeta'})^2, 0.36 b_{l,\zeta'}/(b_{u,\zeta'})^2]$, after $l_{\max}$ iterations, with probability at least $1 - \bar{d}^{-2}$,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le (1-\delta^2)^{l_{\max}} \cdot \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_3 L_\zeta^2 \cdot (\bar{d}\bar{r} + r^* + \gamma|\Omega^*|) \tag{5.10}$$

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \le \frac{b_{u,\zeta'}^2}{b_{l,\zeta'}^2}\big(C_{4,m}\alpha\bar{r}^m(\mu_1\kappa_0)^{4m} + C_{5,m}(\gamma-1)^{-1}\big)\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_{6,m}}{b_{l,\zeta'}^2} L_\zeta^2 \cdot \gamma|\Omega^*|$$

where $C_3 > 0$ depends only on $\delta, b_{l,\zeta'}, b_{u,\zeta'}, m$, and $C_{4,m}, C_{5,m}, C_{6,m} > 0$ are constants depending only on m. Moreover, if $l_{\max}$ is chosen large enough such that the second term on RHS of (5.10) dominates and assume $\kappa_0^{4m}\mu_1^{4m}\bar{r}^m(\bar{r}\bar{d} + r^*) \lesssim_m O(\underline{d}^{m-1})$, we get with probability at least $1 - \bar{d}^{-2}$ that

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} &\le C_6\kappa_0^{2m}\mu_1^{2m}(\bar{r}^m/\underline{d}^{m-1})^{1/2}(\bar{d}\bar{r} + r^* + \gamma|\Omega^*|)^{1/2} \cdot L_\zeta \\
\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}} - \boldsymbol{\mathcal{S}}^*\|_{\ell_\infty} &\le \Big(C_7\kappa_0^{2m}\mu_1^{2m}\bar{r}^{m/2}|\Omega^*|^{1/2}/\underline{d}^{(m-1)/2} + C_8\Big) \cdot L_\zeta
\end{aligned}$$

where $C_6, C_7, C_8 > 0$ depend only on $\delta, b_{l,\zeta'}, b_{u,\zeta'}, m$.

By Theorem 5.5, after a properly chosen $l_{\max}$ iterations and treating γ as a bounded constant, bound (5.10) implies $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 = O\big(L_\zeta^2 \cdot (\bar{d}\bar{r} + r^* + |\Omega^*|)\big)$. Note that the term $\bar{d}\bar{r} + r^* + |\Omega^*|$ is the model complexity and thus this rate is sharp in general. If $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$ so that $|\Omega^*| = 0$, this rate is comparable to the existing ones in generalized low-rank tensor estimation (Wang and Li, 2020; Han et al., 2020). Additionally, for the matrix case ($m = 2$), this rate matches the well-known results in (Davenport et al., 2014; Robin et al., 2020; Wang et al., 2017).

5.4 Community Detection in Hypergraph Networks with Outlier Vertices

Suppose an undirected hypergraph network $\mathcal{G}$ following the hGSBM model (2.9) is observed such that its adjacency tensor $\boldsymbol{\mathcal{A}}$ is symmetric and $[\boldsymbol{\mathcal{A}}]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}([\nu_n\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*]_\omega), \forall \omega = (i_1, \cdots, i_m)$ for $1 \le i_1 \le \cdots \le i_m \le n$, where $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}} \cdot (\mathbf{Z}, \cdots, \mathbf{Z})$ and $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$. The m-th order connection intensity tensor $\boldsymbol{\mathcal{C}}$ has size $K \times \cdots \times K$, whose entries are non-negative and upper bounded by 1.

The parameter $\nu_n \in (0,1]$ characterizes the overall network sparsity, and the membership matrix $\mathbf{Z} \in \{0,1\}^{n\times K}$. As explained in Example 2.3, we assume the vertex set $\mathcal{V} = \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_K \cup \mathcal{O}$ where $\mathcal{O}$ is the subset of outlier vertices. The sizes of communities are denoted by $n_j = |\mathcal{V}_j|$ for $j \in [K]$ and $n_o = |\mathcal{O}|$ such that $n = n_o + \sum_{j=1}^k n_j$. Assume $n_o \le n/2$ such that $\operatorname{rank}(\mathbf{Z}) = K$. Then, we write

$$\boldsymbol{\mathcal{A}} = \nu_n \boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^* + \big(\boldsymbol{\mathcal{A}} - \mathbb{E}\boldsymbol{\mathcal{A}}\big), \tag{5.11}$$

where $\nu_n \boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{K}}$ and the stochastic noise $\boldsymbol{\mathcal{A}} - \mathbb{E}\boldsymbol{\mathcal{A}}$ derives from Bernoulli distribution. Our goal is to estimate the underlying parameters $\nu_n \boldsymbol{\mathcal{T}}^*$ and thus recover the underlying community memberships. Toward that end, we use the loss function

$$\mathfrak{L}(\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{S}}) = \frac{1}{2}\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{S}}\|_{\mathrm{F}}^2 \tag{5.12}$$

for $\boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{K}}$ and $\boldsymbol{\mathcal{S}} \in \mathbb{S}_{\gamma\alpha}$ with $\mathbf{K} = (K, \ldots, K)^T$.

We apply the Riemannian gradient descent Algorithm 2. Since the loss function is quadratic, we set the pruning parameter $\mathsf{k_{pr}}$ to be ∞. To summarize, we apply Algorithm 2 with loss function (5.12), $\mathsf{k_{pr}} = \infty$, $\zeta = O(\nu_n)$ and $\gamma > 1$.

For simplicity, we only focus on the cases of balanced communities: $n_j \asymp n/K, \quad \forall j \in [K]$. Denote $\underline{\lambda}_c := \lambda_{\min}(\boldsymbol{\mathcal{C}})$ and $\overline{\lambda}_c := \lambda_{\max}(\boldsymbol{\mathcal{C}})$. We assume $c_0 \le \underline{\lambda}_c \le \overline{\lambda}_c \le C_0$ for some absolute constants $c_0, C_0 > 0$ implying that $c_0'\nu_n\big((n-n_o)/K\big)^{m/2} \le \underline{\lambda} \le \bar{\lambda} \le C_0'\nu_n\big((n-n_o)/K\big)^{m/2}$ for constants $c_0', C_0' > 0$. Under these conditions, we conclude $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{K}}$, $\|\nu_n\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \nu_n$, $\|\nu_n\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \ge c_0\nu_n(n-n_o)^{m/2}K^{(1-m)/2}$. Therefore, $\nu_n\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1 with $\mu_1 = O_m(K^{(m-1)/2})$ and its condition number is bounded by $O(1)$. By Lemma 5.4, all the iteration $\widehat{\boldsymbol{\mathcal{T}}}_l$ satisfies the spikiness condition. This property enables us to derive sharper bounds for $\mathsf{Err}_{\mathbf{2r}}$ in the sparse regime.

Theorem 5.6. *Let $(\widehat{\boldsymbol{\mathcal{T}}}_l, \widehat{\boldsymbol{\mathcal{S}}}_l)$ be the l-th update by Algorithm 2 with loss function (5.12), $\mathsf{k_{pr}} = \infty, \zeta = C_0\nu_n$ and $\gamma > C_1$ where $C_0, C_1 > 1$ are some absolute constants. Assume $n_o \le n/2$ and*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \nu_n\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m}\underline{\lambda}/\sqrt{K}$

(b) Network sparsity condition:

$$\nu_n(n/K)^{m/2} \ge C_{2,m}\sqrt{\nu_n n}K^{3m/2}(\log n)^{(m+8)/2} + C_{3,m}(Km\alpha n_o n^{m-1} + K\gamma\alpha n^m)^{1/2}$$

(c) Edge intensity of outlier vertices: $\alpha \le c_{3,m}K^{-m}$

where $c_{1,m}, c_{3,m}, C_{2,m}, C_{3,m} > 0$ depend on m only. If the step size $\beta \in [0, 005, 0.36]$, then after $l_{\max}$ iterations, with probability at least $1 - n^{-2}$, we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \nu_n\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le 0.98^{l_{\max}}\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \nu_n\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_{4,m}\nu_n n K^{3m}(\log n)^{m+8} + C_{5,m}(Km\alpha n_o n^{m-1} + K\gamma\alpha n^m),$$

where $C_{4,m}, C_{5,m} > 0$ depend on m only.

By Theorem 5.6, after a suitable chosen $l_{\max}$ iterations, the relative error is, with probability at least $1 - n^{-2}$, bounded by

$$\frac{\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \nu_n \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\lambda_{\min}(\nu_n \boldsymbol{\mathcal{T}}^*)} \leq C_{4,m} \frac{K^{2m}(\log n)^{(m+8)/2}}{\sqrt{\nu_n n^{m-1}}} + C_{5,m} \frac{K^{(m+1)/2}(\alpha n_o n^{m-1} + \gamma \alpha n^m)^{1/2}}{\nu_n n^{m/2}} \tag{5.13}$$

The relative error (5.13) plays the essential role in the consistency of spectral clustering based on the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$. See the subsequent K-means spectral clustering for community detection in (Ke et al., 2019). Interested readers can follow the routine procedures (Jing et al., 2020; Rohe et al., 2011; Lei and Rinaldo, 2015) to explicitly derive the mis-clustering error rate of K-means algorithm based on (5.13). We now focus on the implications of bound (5.13).

The RHS of (5.13) converges to zero when (i) $K^{2m}(\log n)^{(m+8)/2}/\sqrt{\nu_n n^{m-1}} \to 0$ and (ii) $K^{(m+1)/2}(\gamma \alpha n^m)^{1/2}/(\nu_n n^{m/2}) \to 0$ as $n \to \infty$. The first condition requires that the average node degree $\nu_n n^{m-1} \gg K^{4m}(\log n)^{m+8}$, which matches (up to the powers of K and $\log n$) the best known results on the network sparsity for consistent community detection in hypergraph SBM. See, e.g., (Ghoshdastidar and Dukkipati, 2017; Ke et al., 2019; Kim et al., 2018). Interestingly, the second condition requires, by setting $\gamma = O(1)$, that $\nu_n \gg \alpha^{1/2} K^{(m+1)/2}$. It means that the edge intensity of outlier vertices should be small enough to ensure that the signal strength of normal communities dominates that of the outlier community. Recall $\lambda_{\min}(\nu_n \boldsymbol{\mathcal{T}}^*) \gtrsim \nu_n (n/K)^{m/2}$ and a simple upper bound for $\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ is $O\big((\alpha n^m)^{1/2}\big)$. Therefore, condition (ii) implies that $\lambda_{\min}(\nu_n \boldsymbol{\mathcal{T}}^*) \gg K^{1/2}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$. Ignoring $K^{1/2}$, this condition is perhaps minimal to guarantee the consistent recovery of $\nu_n \boldsymbol{\mathcal{T}}^*$, in the sense that the relative error goes to 0 as $n \to \infty$.

6 When Sparse Component is Absent

In this section, we consider the special case when the sparse component is absent, i.e., $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$. For the exact low-rank tensor model, we observe that many conditions in Section 4 can be relaxed. A major difference is that the spikiness condition is generally not required for exact low-rank model. Consequently, the trimming step in Algorithm 2 is unnecessary. Therefore, it suffices to simply apply the Riemannian gradient descent algorithm to solve for the underlying low-rank tensor $\boldsymbol{\mathcal{T}}^*$. For ease of exposition, the procedure is summarized in Algorithm 3 (largely the same as Algorithm 2).

Algorithm 3 runs fast and guarantees favourable convergence performances under weaker conditions than Theorem 4.1. Indeed, since there is no sparse component, only Assumption 2 is required to guarantee the convergence of Algorithm 3. Similarly as Section 4, the error of final estimate produced by Algorithm 3 is characterized by the gradient at $\boldsymbol{\mathcal{T}}^*$. With a slightly abuse of notation, denote

$$\mathsf{Err}_{\mathbf{2r}} = \sup_{\boldsymbol{\mathcal{X}} \in \mathbb{M}_{\mathbf{2r}}, \|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} \leq 1} \langle \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^*), \boldsymbol{\mathcal{X}} \rangle.$$

Algorithm 3 Riemannian Gradient Descent for Exact Low-rank Estimate

Initialization: $\widehat{\boldsymbol{\mathcal{T}}}_0 \in \mathbb{M}_{\mathbf{r}}$ and stepsize $\beta > 0$

for $l = 0, 1, \cdots, l_{\max}$ **do**

$\boldsymbol{\mathcal{G}}_l = \nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_l)$

$\boldsymbol{\mathcal{W}}_l = \widehat{\boldsymbol{\mathcal{T}}}_l - \beta \mathcal{P}_{\mathbb{T}_l} \boldsymbol{\mathcal{G}}_l$

$\widehat{\boldsymbol{\mathcal{T}}}_{l+1} = \mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\boldsymbol{\mathcal{W}}_l)$

end for

Output: $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$

Theorem 6.1. *Suppose Assumption 2 holds with* $\boldsymbol{\mathcal{S}}^* = \mathbf{0}$ *and* $\mathbb{B}_2^* = \{\boldsymbol{\mathcal{T}} : \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{0,m}\underline{\lambda}, \boldsymbol{\mathcal{T}} \in \mathbb{M}_{\mathbf{r}}\}$ *for a small constant* $c_{0,m} > 0$ *depending on* m *only, also suppose* $1.5 b_l b_u^{-2} \le 1$ *and* $0.75 b_l b_u^{-1} \ge \delta^{1/2}$ *for some* $\delta \in (0, 1]$ *and the stepsize* $\beta \in [0.4 b_l b_u^{-2}, 1.5 b_l b_u^{-2}]$ *in Algorithm 3. Assume*

(a) Initialization: $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda} \cdot c_{1,m} \delta \bar{r}^{-1/2}$

(b) Signal-to-noise ratio: $\mathsf{Err}_{2\mathbf{r}}/\underline{\lambda} \le c_{2,m} \delta^2 \bar{r}^{-1/2}$

where $c_{1,m}, c_{2,m} > 0$ *are small constants depending only on* m*. Then for all* $l = 1, \cdots, l_{\max}$*,*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le (1 - \delta^2)^l \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_\delta \mathsf{Err}_{2\mathbf{r}}^2$$

where $C_\delta > 0$ *is a constant depending only on* δ*. Then after at most* $l_{\max} \asymp \log\left(\underline{\lambda}/\mathsf{Err}_{2\mathbf{r}}\right)$ *iterations (also depends on* $b_l, b_u, m, \bar{r}$ *and* β*), we get*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le C \cdot \mathsf{Err}_{2\mathbf{r}},$$

where the constant $C > 0$ *depends on only* $b_l, b_u, m, \bar{r}$ *and* β*.*

Note that Theorem 6.1 holds without spikiness condition in contrast with Theorem 4.1. It makes sense for the model has no missing values or sparse corruptions. The assumptions on loss function are also weaker (e.g., no need to be an entry-wise loss or entry-wisely smooth) than those in Theorem 4.1. As a result, Theorem 6.1 is also applicable to the low-rank tensor regression model among others. See (Han et al., 2020; Chen et al., 2019; Xia et al., 2020) and references therein. The initialization and signal-to-noise conditions are similar to those in Theorem 4.1, e.g., by setting $|\Omega^*| = \alpha = 0$ there. In addition, the error of final estimate depends only on $\mathsf{Err}_{2\mathbf{r}}$. Interestingly, the contraction rate does not depend on the condition number κ_0.

Comparison with existing literature In (Han et al., 2020), the authors proposed a general framework for exact low-rank tensor estimation based on regularized jointly gradient descent on the core tensor and associated low-rank factors. Their method is fast and achieves statistical optimality

in various models. In contrast, our algorithm is based on Riemannian gradient descent, requires no regularization and also runs fast. An iterative tensor projection algorithm was studied in (Yu and Liu, 2016). But their method only applies to tensor regression. Other notable works focusing only on tensor regression include (Zhang et al., 2020a; Zhou et al., 2013; Hao et al., 2020; Sun et al., 2017; Li et al., 2018; Pan et al., 2018; Rauhut et al., 2017). A general projected gradient descent algorithm was proposed in (Chen et al., 2019) for generalized low-rank tensor estimation. For Tucker low-rank tensors, their algorithm is similar to our Algorithm 3 except that they use vanilla gradient $\boldsymbol{\mathcal{G}}_l$ while we use the Riemannian gradient $\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l$. As explained in Section 3, using the vanilla gradient can cause heavy computation burdens in the subsequent steps.

Riemannian gradient descent algorithm for tensor completion was initially proposed by (Kressner et al., 2014). They focused only on tensor completion model and did not investigate its theoretical guarantees. Recently in (Cai et al., 2020c), the Riemannian gradient descent algorithm is applied for noiseless tensor regression and its convergence analysis is proved. In (Mu et al., 2014), the authors proposed a matrix-based method for tensor recovery which is, generally, computationally slow and statistically sub-optimal.

7 Numerical Experiments

In this subsection, we test the performances of our algorithms on synthetic datasets, specifically for the four applications studied in Section 5.

7.1 Robust Tensor PCA

The low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$ is generated from the HOSVD of a trimmed standard normal tensor. It satisfies the spikiness condition, with high probability, and has singular values $\bar{\lambda} \approx 3$ and $\underline{\lambda} \approx 1$. Given a sparsity level $\alpha \in (0, 1)$, the entries of sparse tensor $\boldsymbol{\mathcal{S}}^*$ are i.i.d. sampled from $\mathrm{Be}(\alpha) \times \mathrm{N}(0, 1)$, which ensures $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_{O(\alpha)}$ with high probability. This ensures that the non-zero entries of $\boldsymbol{\mathcal{S}}^*$ have typically much larger magnitudes than the entries of $\boldsymbol{\mathcal{T}}^*$. The noise tensor $\boldsymbol{\mathcal{Z}}$ has i.i.d. entries sampled from $\mathrm{N}(0, \sigma_z^2)$. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the left panels of Figure 1.

The top-left plot in Figure 1 displays the effects of α on the convergence of Algorithm 2. It shows that the convergence speed of Algorithm 2 is insensitive to α, while the error of final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. This is consistent with the claims of Theorem 5.1. In the middle-left plot of Figure 1, we observe that, for a fixed sparsity level α, the error of final estimates grows as the tuning parameter γ becomes larger. The bottom-left plot of Figure 1 shows the convergence of Algorithm 2

for different noise levels. All these plots confirm the fast convergence of our Riemannian gradient descent algorithm. In particular, there are stages during which the log relative error decreases linearly w.r.t. the number of iterations, as proved in Theorem 4.1.

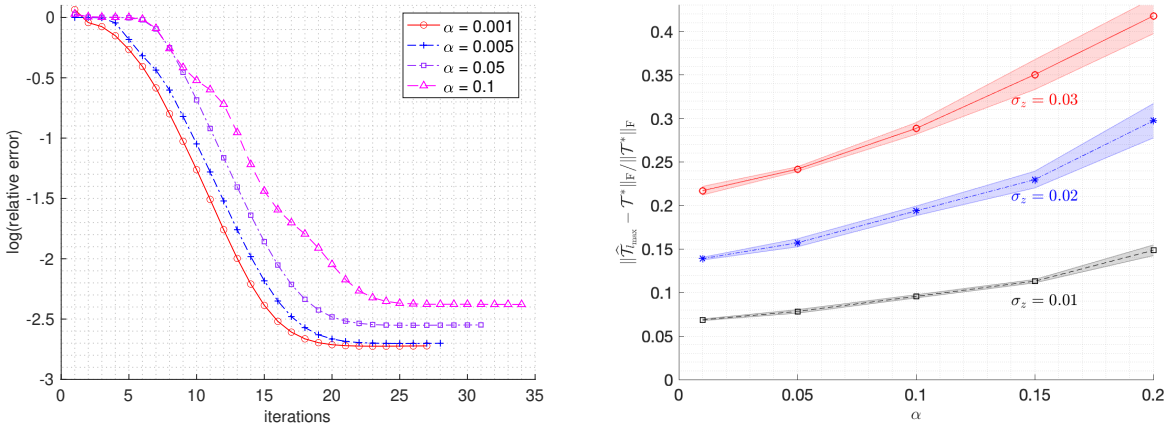
The statistical stability of the final estimates by Algorithm 2 is demonstrated in the right panels of Figure 1. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 10 simulations, and the error bar shows the confidence region by one empirical standard deviation. Based on these plots, we observe that the standard deviations of $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ grow as the noise level σ_z, the sparsity level α or the tuning parameter γ increases.

7.2 Tensor PCA with Heavy-tailed Noise

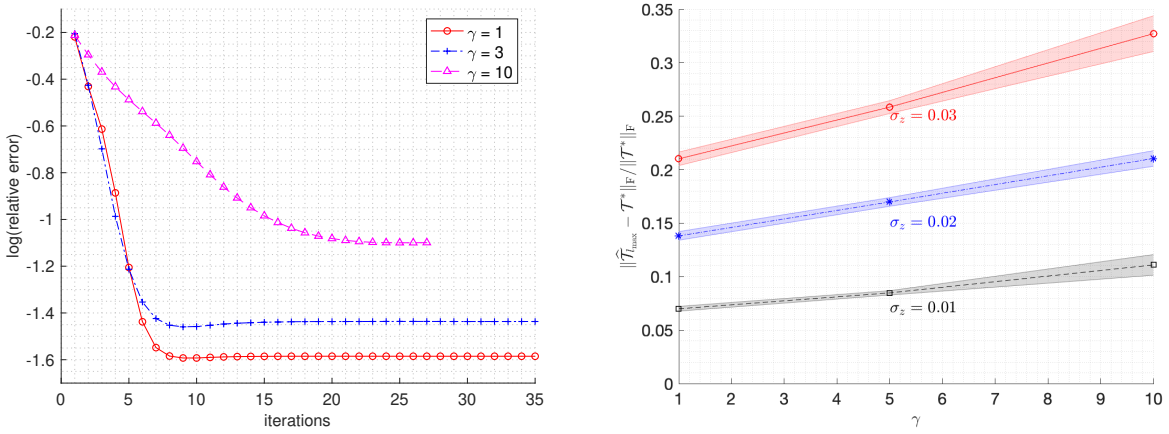
The low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$ is generated from the HOSVD of a trimmed standard normal tensor, as in Section 7.1. Given a parameter θ, we generate the noisy tensor whose entries are i.i.d. and satisfy the Student-t distribution with degree of freedom θ. But notice here we also apply a global scaling to better control the noise standard deviation. We denote the noisy tensor after scaling by $\boldsymbol{\mathcal{Z}}$. This generated tensor $\boldsymbol{\mathcal{Z}}$ satisfies Assumption 4 with the same parameter θ. Once the parameter θ and global scaling are given, we are able to calculate the variance σ_z^2. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the upper panels of Figure 2.

The top-left plot in Figure 2 displays the effects of α on the convergence of Algorithm 2. The case $\alpha = 0$ reduces to the normal Riemannian gradient descent, which cannot output a satisfiable result due to the heavy-tailed noise, even if a warm initialization is provided. This shows the importance of gradient pruning in Algorithm 2. When $\alpha > 0$, the convergence speed of the algorithm is insensitive to α, but the final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. In the top-right plot of Figure 2, we observe the error becomes larger as θ decreases (or equivalently, as σ_z^2 increases). All these results match the claim of Theorem 5.3 and confirm the fast convergence of Riemannian gradient descent. And there are indeed stages where the log relative error decreases linearly w.r.t. the number of iterations.

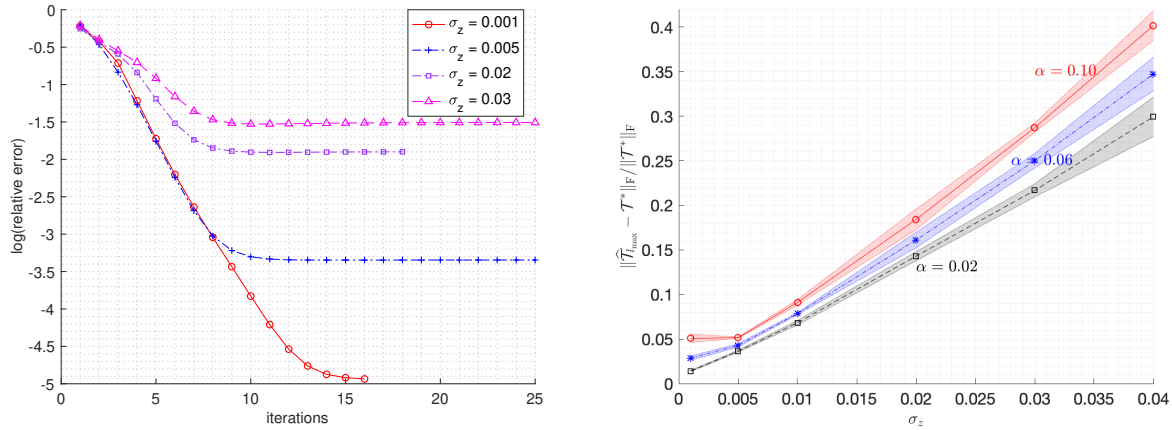
The statistical stability of the final estimates by Algorithm 2 applied to tensor PCA with heavy-tailed noise is demonstrated in the bottom panel of Figure 2. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 5 simulations, and the error bar shows the confidence region by one empirical standard deviation. Based on these plots, we observe that for each fixed θ (or σ_z^2, equivalently), we need to choose α carefully to achieve the best performance. This is reasonable since in the heavy-tail noise setting, we do not know the sparsity of outliers. Also, the figure shows that Algorithm 2 is stable for different α and θ.



(a) Change of sparsity α. Left: $\sigma_z = 0.01, \gamma = 2$; Right: $\gamma = 2$

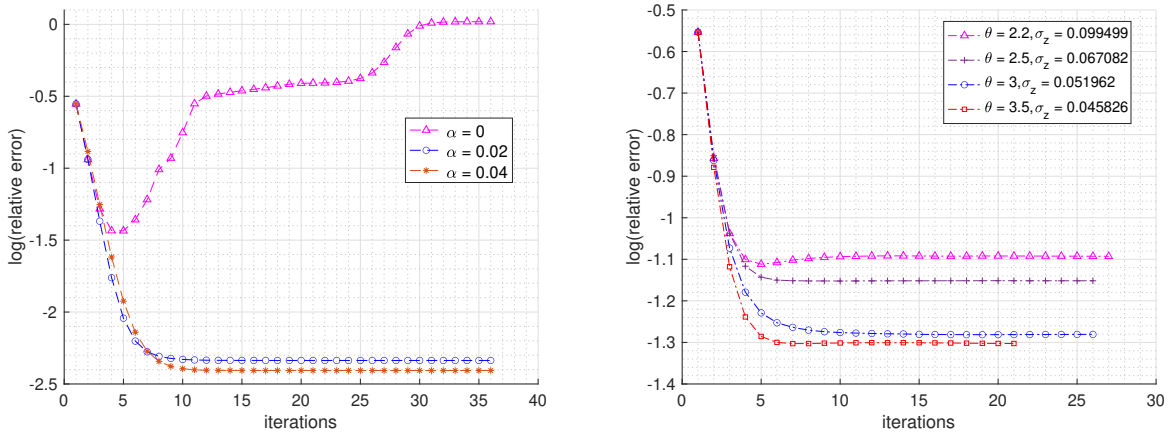


(b) Change of γ. Left: $\alpha = 0.02, \sigma_z = 0.01$; Right: $\alpha = 0.02$

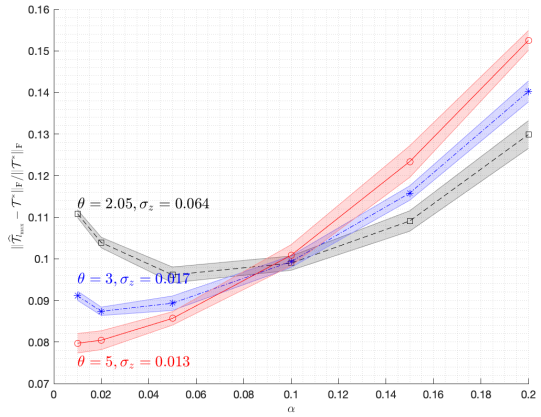


(c) Change of noise size σ_z. Left: $\alpha = 0.02, \gamma = 2$; Right: $\gamma = 2$

Figure 1: Performances of Algorithm 2 for robust tensor PCA. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on left panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} / \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the right panels are based on 1 standard deviation from 10 replications.



(a) Left: Change of α, $\theta = 2.2(\sigma_z = 0.332)$; Right: Change of θ, $\alpha = 0.01$



(b) Change of θ

Figure 2: Performances of Algorithm 2 for tensor PCA with heavy-tailed noise. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on upper panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the lower panels are based on 1 standard deviation from 5 replications.

7.3 Binary Tensor

In the binary tensor setting, we generate the low-rank tensor $\boldsymbol{\mathcal{T}}^* \in \mathbb{R}^{d\times d\times d}$ with $d = 100$ and Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$ from the HOSVD of a trimmed standard normal tensor. But here we did a scaling to $\boldsymbol{\mathcal{T}}^*$ so that the singular value $\bar{\lambda} \approx 300$ and $\underline{\lambda} \approx 100$. Given a sparsity level $\alpha \in (0, 1)$, the entries of sparse tensor $\boldsymbol{\mathcal{S}}^*$ are i.i.d. sampled from $\mathrm{Be}(\alpha) \times \mathrm{N}(0, 1)$, which ensures $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_{O(\alpha)}$ with high probability. We generate the tensor $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ in this way in order to meet the requirements of Assumption 5. In the following experiments, we are considering the logistic link function with the scaling parameter σ, i.e., $p(x) = (1 + e^{-x/\sigma})^{-1}$. The convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ by Algorithm 2 are examined and presented in the top two panels of Figure 3.

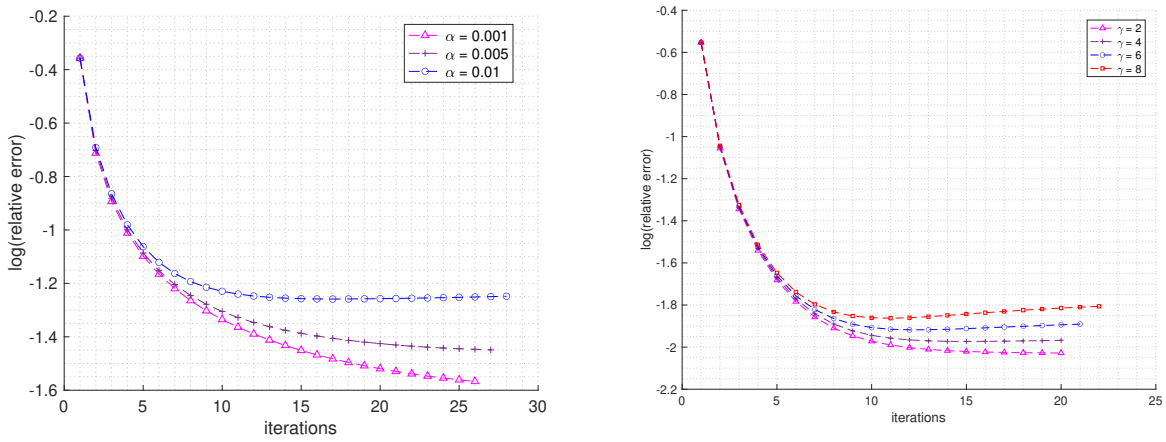
The top-left plot in Figure 3 shows the effect of α on the convergence of Algorithm 2. From the figure, it is clear that the error of final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ is related to α. This again verifies the results in Theorem 5.5. In the top-right plot in Figure 3, we can see the error of the final estimates increases as the parameter γ becomes larger. All these experiments show that Riemannian gradient descent converges fast and there are stages when the log relative error decreases linearly w.r.t. the number of iterations.

The statistical stability of the final estimates by Algorithm 2 is demonstrated in the bottom panel of Figure 3. Each curve represents the average relative error of $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ based on 5 simulations, and the error bar shows the confidence region by one empirical standard deviation. From these plots, we observe that the standard deviations of $\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ grow as the noise level, the sparsity level α or the tuning parameter γ increases.

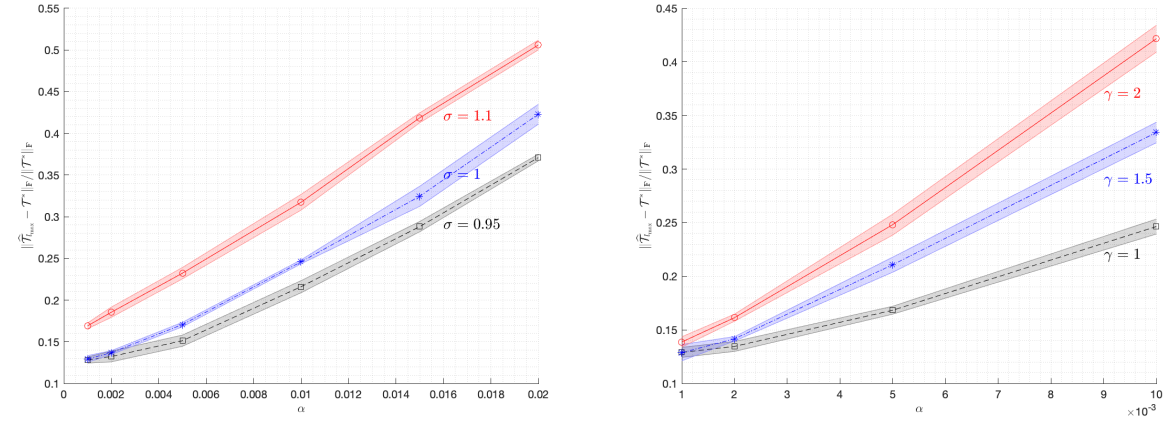
7.4 Community Detection

In the community detection experiment, we consider the 3-uniform graph $\mathcal{G}$ of $n = 100$ vertices. And we assume there are $K = 2$ normal communities. We manually generate the tensor $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{2\times 2\times 2}$ by setting $[\boldsymbol{\mathcal{C}}]_{1,1,1} = [\boldsymbol{\mathcal{C}}]_{2,2,2} = 1$ and setting the other entries to be 0.5. The outlier tensor $\boldsymbol{\mathcal{S}}^*$ is generated in the following way: whenever the subscript contains an outlier, we set the corresponding entry to be $\mathrm{Be}(\alpha)$. The tensor $\boldsymbol{\mathcal{S}}^*$ generated in this way ensures $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_{O(\alpha)}$ with high probability. We now perform experiments by modifying α, the network sparsity ν and the number of outliers. We apply the Algorithm 2 and the convergence performances of $\log(\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}})$ are examined and presented in Figure 4.

The top-left plot in Figure 4 displays the effect of α on the convergence of Algorithm 2. It shows that the convergence speed of Algorithm 2 is insensitive to α, while the error of final estimates $\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}$ increases as α increases. This is consistent with the claims of Theorem 5.6. In the top-right plot



(a) Left: Change of sparsity α, $\sigma = 1, \gamma = 1$; Right: Change of γ, $\alpha = 0.001, \sigma = 1$



(b) Left: Change of σ, $\gamma = 1$; Right: Change of γ, $\alpha = 0.001$

Figure 3: Performances of Algorithm 2 for binary tensor. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on left panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} / \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. The error bars on the bottom panels are based on 1 standard deviation from 5 replications.

of Figure 4, we examine the influence of network sparsity ν on the convergence of the Algorithm 2. It shows that when the parameter ν is too small, the algorithm performs poorly. As ν becomes larger, the error of the final estimates decreases. And in the bottom plot of Figure 4, we examine the impact of the number of outliers on the converge performance. When the number of outliers increases, the error of the final estimates also increases.

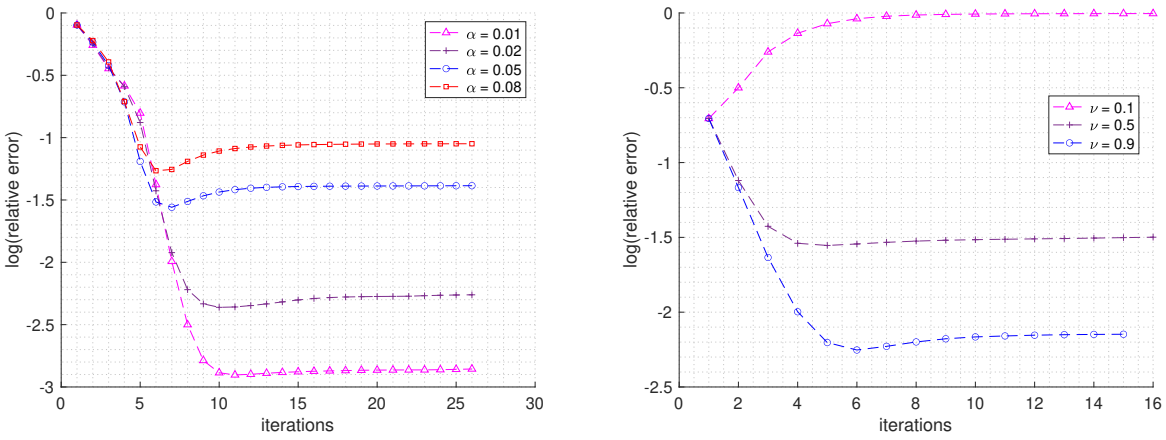
8 Real Data Analysis

8.1 International Commodity Trade Flows

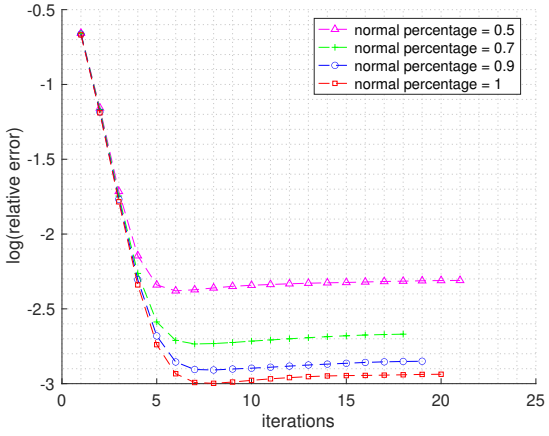
We collected the international commodity trade data from the API provided by UN website *https://comtrade.un.org*. The dataset contains the monthly information of imported commodities by countries from Jan. 2010 to Dec. 2016 (84 months in total). For simplicity, we focus on 50 countries among which 35 are from Europe, 9 from America, 5 from Asia[3] and 1 from Africa. All the commodities are classified into 100 categories based on the 2-digit HS code (*https://www.foreign-trade.com/reference/hscode.htm*). Thus, the raw data is a $50 \times 50 \times 100 \times 84$ tensor. At any month and for any category of commodity, there is a directed and weighted graph of size 50×50 depicting the trade flow between countries. The international trade has cyclic pattern annually. Since we are less interested in the time domain, we eliminate the fourth dimension by simply adding up the entries. Finally, we end up with a tensor $\boldsymbol{\mathcal{A}}$ of size $50 \times 50 \times 100$.

In Figure 5, circular plots are presented for illustrating the special trade patterns of some commodities. The countries are grouped and coloured by continent, i.e., Europe by red, Asia by green, America by blue and Africa by black. The links represent the directional trade flow between nations and are coloured based on the starting end of the link. The position of starting end of the link is shorter than the other end to give users the feeling that the link is moving out. The thickness of link indicates the volume of trade. From the top-left plot, we observe that Japan imports a large volume of tobacco related commodities; Germany is the largest exporter; Poland and Brazil are the second and third largest exporter; USA both import and export a large quantity of tobacco commodity. The top-right plot shows that USA and Canada import and export large volumes of mineral fuels; Malaysia exports lots of mineral fuels to Japan; Algeria exports a large quantity of miner fuels which plays the major role of international trade of this Africa country. The middle-left plot shows that Portugal is the largest exporter of Cork, and European countries are the major exporter and importer of this commodity. From the middle-right plot, we observe that Pakistan is the major exporter of Cotton in Asia; the European countries Turkey, Italy and Germany all

[3]Egypt is at the cross of Eastern Africa and Western Asia. For simplicity, we treat it as an Asian country. In addition, Turkey is treated as an Eastern European country rather than a Western Asian country.



(a) Left: Change of α, $\nu = 0.5$, 10% of outliers; Right: Change of ν, $\alpha = 0.05$, 10% of outliers



(b) Change of number of outliers, $\nu = 0.5, \alpha = 0.01$

Figure 4: Performances of Algorithm 2 for community detection. The low-rank $\boldsymbol{\mathcal{T}}^*$ has size $d \times d \times d$ with $d = 100$ and has Tucker ranks $\mathbf{r} = (2, 2, 2)^\top$. The relative error on upper panels is defined by $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \nu\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/\|\nu\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ where ν describes the overall network sparsity. The error bars on the lower panels are based on 1 standard deviation from 5 replications.

export and import large volumes of cotton; USA exports a great deal of cotton to Mexico,Turkey and Philippines. The bottom-left plot shows that Malaysia and Belgium are the largest exporter of Tin and USA is the major importer. Finally, the bottom-right plot shows that Switzerland is the single largest exporter of clocks and watches; USA is the major importer; France and Germany both export and import large quantities of clocks and watches.

We implement the robust PCA framework as Section 5.1 to analyze the tensor $\log(1+\boldsymbol{\mathcal{A}})$, where a logarithmic transformation helps shrink the extremely large entries. The Tucker ranks are set as $(3,3,3)$, although most of the results seem insensitive to the ranks so long as they are bounded by 5. Nevertheless, the sparsity ratio α of $\boldsymbol{\mathcal{S}}$ significantly impacts the outputs of Algorithm 2. The algorithm is initialized by a higher-order singular value decomposition, which finally produces a low-rank $\widehat{\boldsymbol{\mathcal{T}}}$ and a sparse $\widehat{\boldsymbol{\mathcal{S}}}$. We shall use $\widehat{\boldsymbol{\mathcal{T}}}$ to uncover underlying relations among countries, and $\widehat{\boldsymbol{\mathcal{S}}}$ to examine distinctive trading patterns of certain commodities.

In particular, the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$ are utilized to illustrate the community structure of nations. Note that the 1st-dim and 2nd-dim singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$ are distinct because the trading flows are directed. We observe that the 2nd-dim singular vectors often render better results. Then, a procedure of multi-dimensional scaling is adopted to visualize the rows of these singular vectors. The results are presented in Figure 6 for four choices of $\alpha \in \{0.03, 0.1, 0.2, 0.3\}$. All the plots in Figure 6 reveal certain degrees of the geographical relations among countries. It is reasonable since regional trade partnerships generally dominate the inter-continental trade relations. The European countries (coloured in blue) are mostly separated from the others. Overall, countries from America (coloured in red) and Asia (coloured in magenta) are less separable especially when α is large. For small α like 0.03, the 5 Asian countries are clustered together and the major 8 American nations lie on the top-right corner of the plot. The two geographically close African countries Algeria and Egypt are also placed together in the top-left plot of Figure 6, as is the case with the Western European nations such as United Kingdom, Spain, France, Germany and Italy.

The four plots in Figure 6 show that the low-rank estimate is sensitive to the choices of sparsity ratio α. Interesting shifts appear as α increases. Indeed, the geographical relations become a less important factor but the economic similarity plays the dominating role. For instance, some Asian and American nations split and merge into two clusters. The three large economies US, Canada and Japan are merged into one cluster, while the other small and less-developed Asian and American countries are merged into another cluster. It may be caused by that these three large economies are better at advanced technology and share similar structures in exporting high end commodities. Moreover, as α increases, the African country Algeria moves closer to the less-developed American and Asian nations. All these nations including Algeria rely heavily on exporting natural resources even if Algeria is geographically far from the others. Another significant shift is that the European

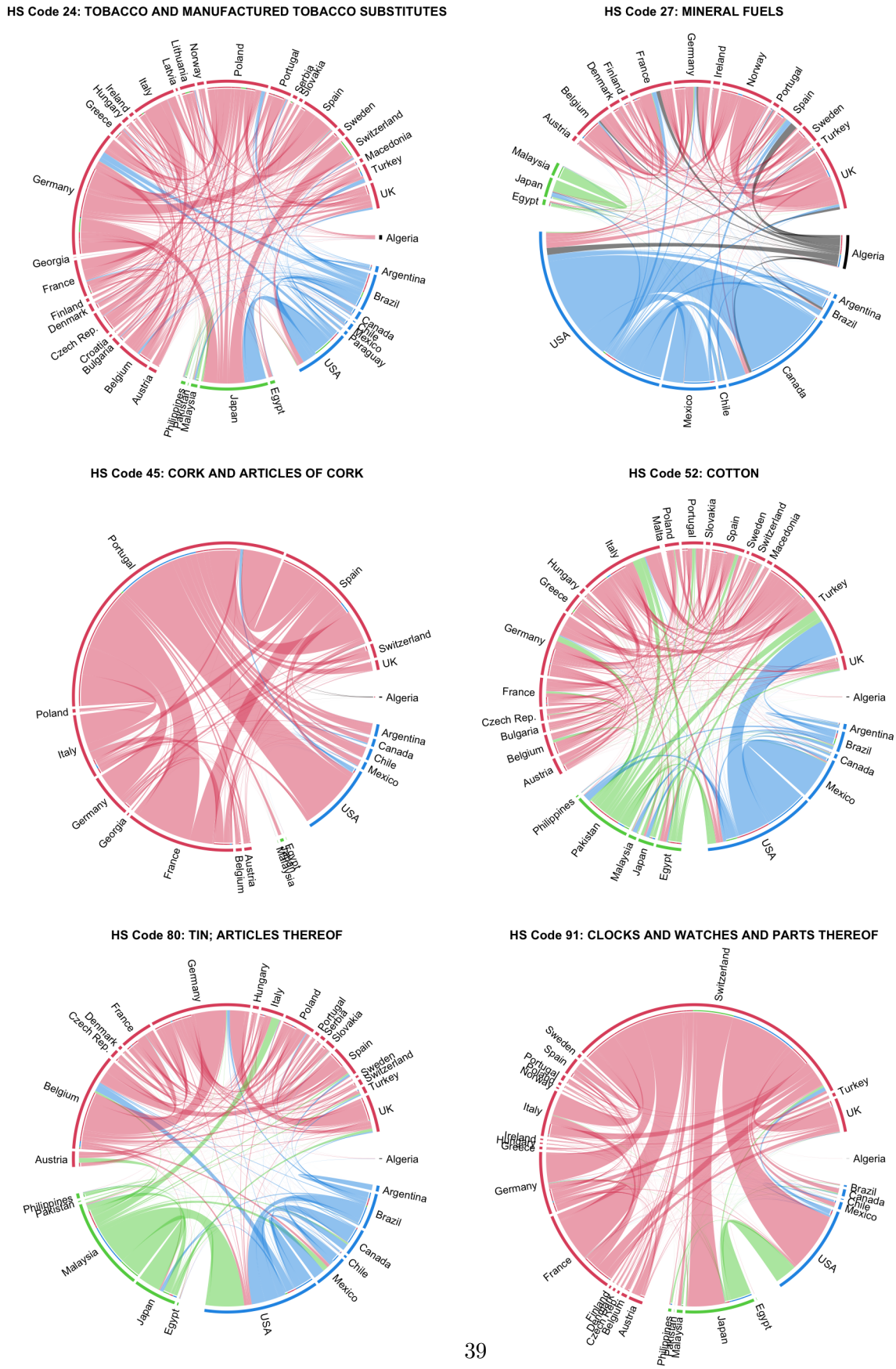


Figure 5: International Commodity Trade Flow. Import and export flow of some commodities.

countries split into two clusters as α increases. Moreover, one cluster comprising those wealthy and advanced Western European countries move closer to the group of US, Japan and Canada. These countries have close ties in trading high end products and components, although they belong to distinct continents. The other cluster includes mostly the Central and Eastern European countries, among which regional trade flows are particularly intense. Interestingly, there are two outlier countries Ireland (north-western Europe) and Antigua and Barbuda (a small island country in middle America). They do not merge into any clusters. The magnitudes of coordinates of these two points suggest that their international trade is not active. While the pattern shift in Figure 6 is intriguing, we are not clear why and how it is related to the sparsity ratio.

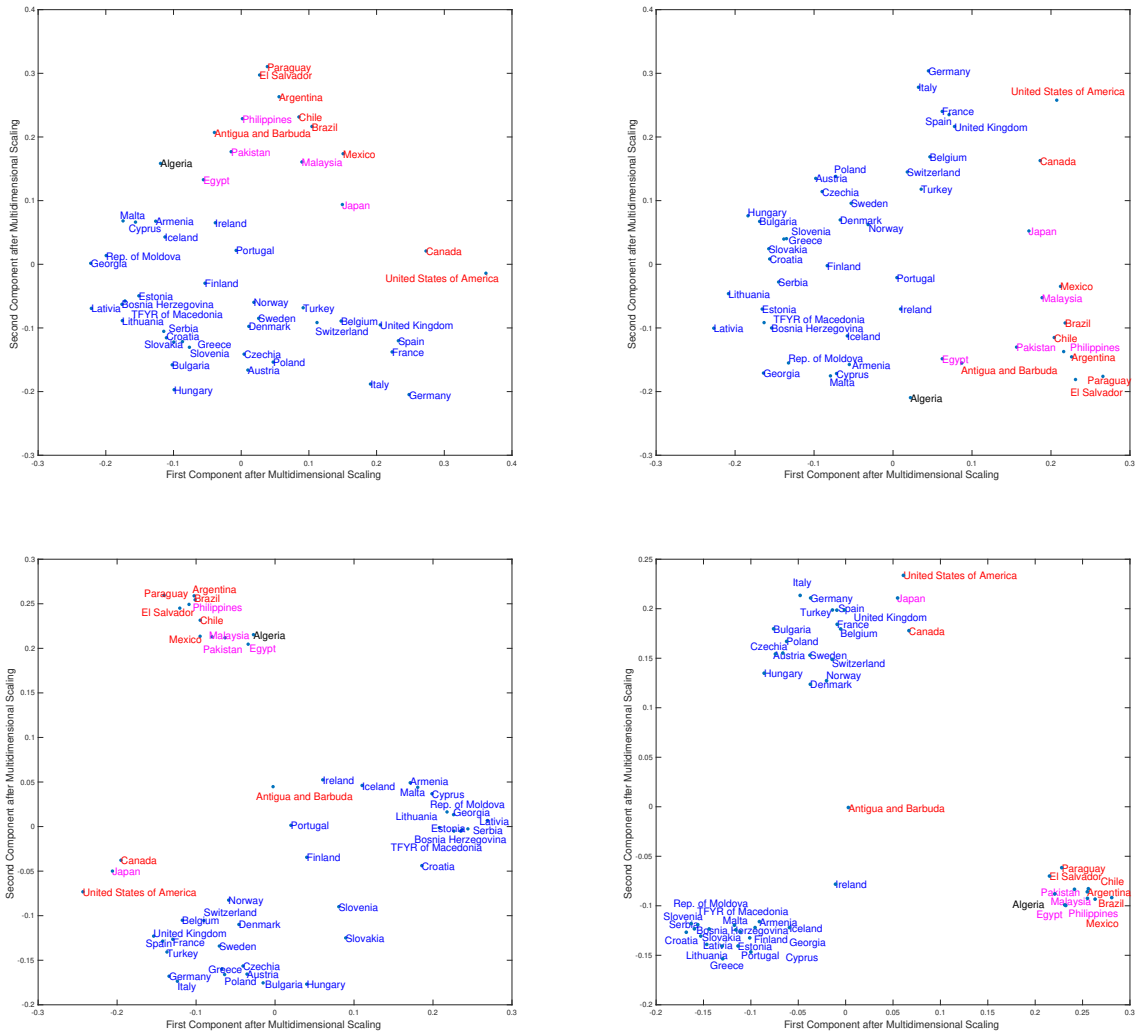
We now look into the slices of the sparse estimate $\widehat{\boldsymbol{\mathcal{S}}}$ and investigate the distinctive trading patterns of certain commodities. As the sparsity ratio α grows, the slices of $\widehat{\boldsymbol{\mathcal{S}}}$ become denser whose patterns are more difficult to examine. For simplicity, we mainly focus on small values of α like 0.03. The results are presented in Figure 7. The top-left plot shows that Algeria exports exceptionally large volumes of mineral fuels that can not be explained by the low-rank tensor estimate. Similarly, based on the top-right plot, we observe that the exact low-rank tensor PCA fails to explain the trading export of cork by Portugal. Fortunately, these interesting and significant trading patterns can be easily captured by the additional sparse tensor estimate. The bottom-left and bottom-right plots of Figure 7 showcase the unusual exports of cotton by Pakistan and exports of Tin by Malaysia, respectively. These findings echo some trading patterns displayed in Figure 5, e.g., Portugal is the single largest exporter of cork and Malaysia is the largest exporter of Tin.

8.2 Statisticians Hypergraph Co-authorship Network

This dataset (Ji and Jin, 2016) contains the co-authorship relations of 3607 statisticians based on 3248 papers published in four prestigious statistics journals during 2003-2012. The co-authorship network thus has 3607 nodes and two nodes are connected by an edge if they collaborated on at least one paper. A giant connected component of this network consisting of 236 nodes is seen to be the "High-Dimensional Data Analysis" community. They also carried out community detection analysis to discover substructures in this giant component. See more details in (Ji and Jin, 2016).

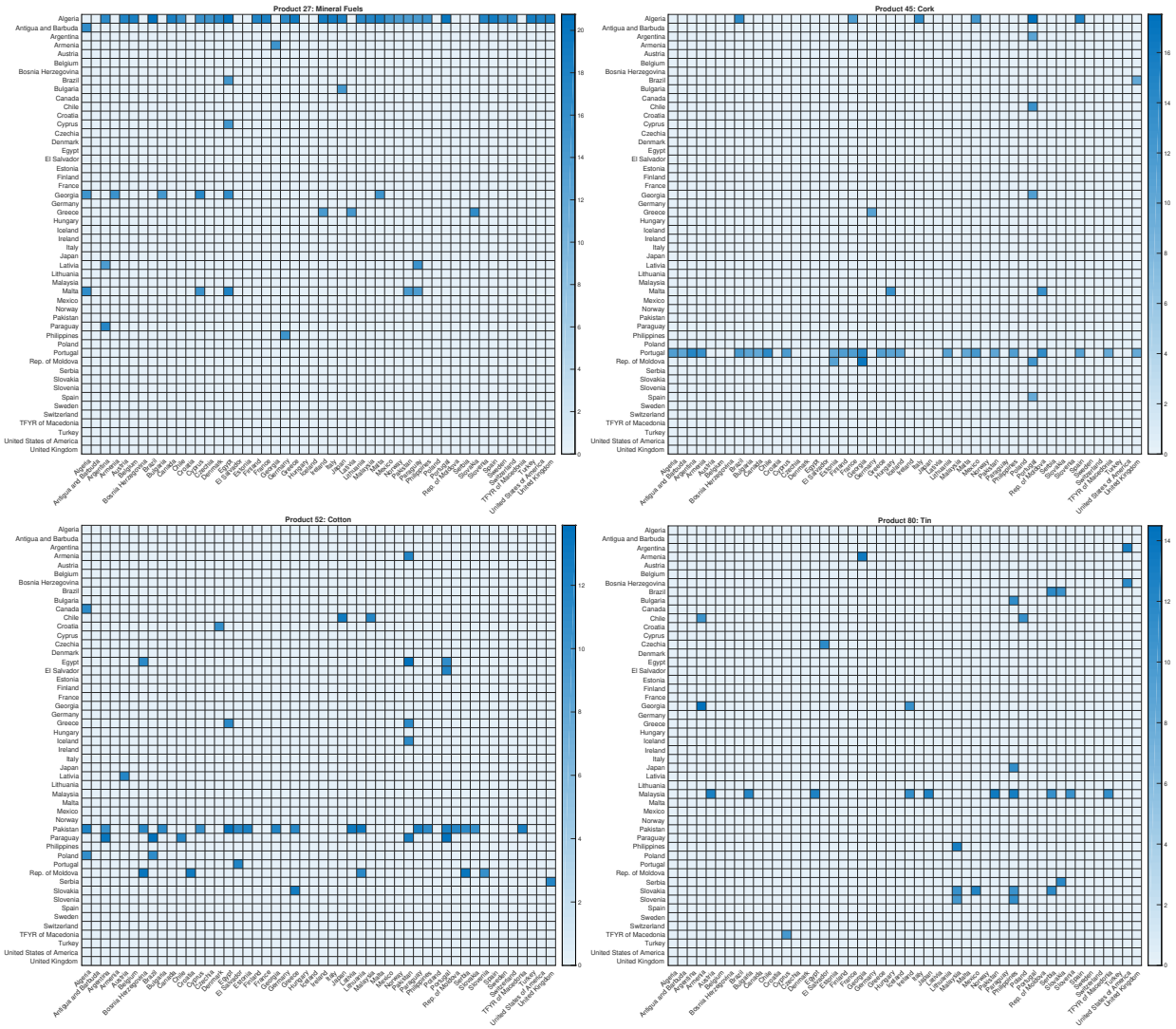
We analyze the substructures of the giant component by treating it as a hypergraph co-authorship network. These 236 statisticians co-authored 542 papers[4], among which 356 papers have two co-authors, 162 papers have three co-authors and 24 papers have four co-authors. A 3-uniform hypergraph co-authorship network is constructed by, for $i \neq j \neq k$, adding the hyperedge (i, j, k) if the authors i, j, k co-authored at least one paper, and adding the hyperedges (i, i, j) and

[4]There are 328 single-authored papers. They provide no information to co-authorship relations, and are left out in our analysis.



(a) Sparsity level: top-left, $\alpha = 0.03$; top-right, $\alpha = 0.1$; bottom-left, $\alpha = 0.2$; bottom-right, $\alpha = 0.3$.

Figure 6: Visualize the relations between countries by spectral estimates. Tucker ranks are set as $(3, 3, 3)$ and the low-rank tensor is estimated by robust PCA as in Section 5.1 with HOSVD initialization. Blue coloured countries are from Europe, red from America, magenta from Asia and black from Africa. When α is small, clusters have strong implications on the geographical closeness between nations; As α becomes large, economic structures become the major factor in that large and advanced economies tend to merge, and so do low end economies and natural resource reliant economies.



(a) Commodity: top-left, Mineral Fuels; top-right, Cork; bottom-left, Cotton; bottom-right, Tin.

Figure 7: Heatmaps of the slices of $\widehat{\boldsymbol{\mathcal{S}}}$. They reveal distinctive trading patterns of certain commodities. The sparsity ratio is set at $\alpha = 0.03$.

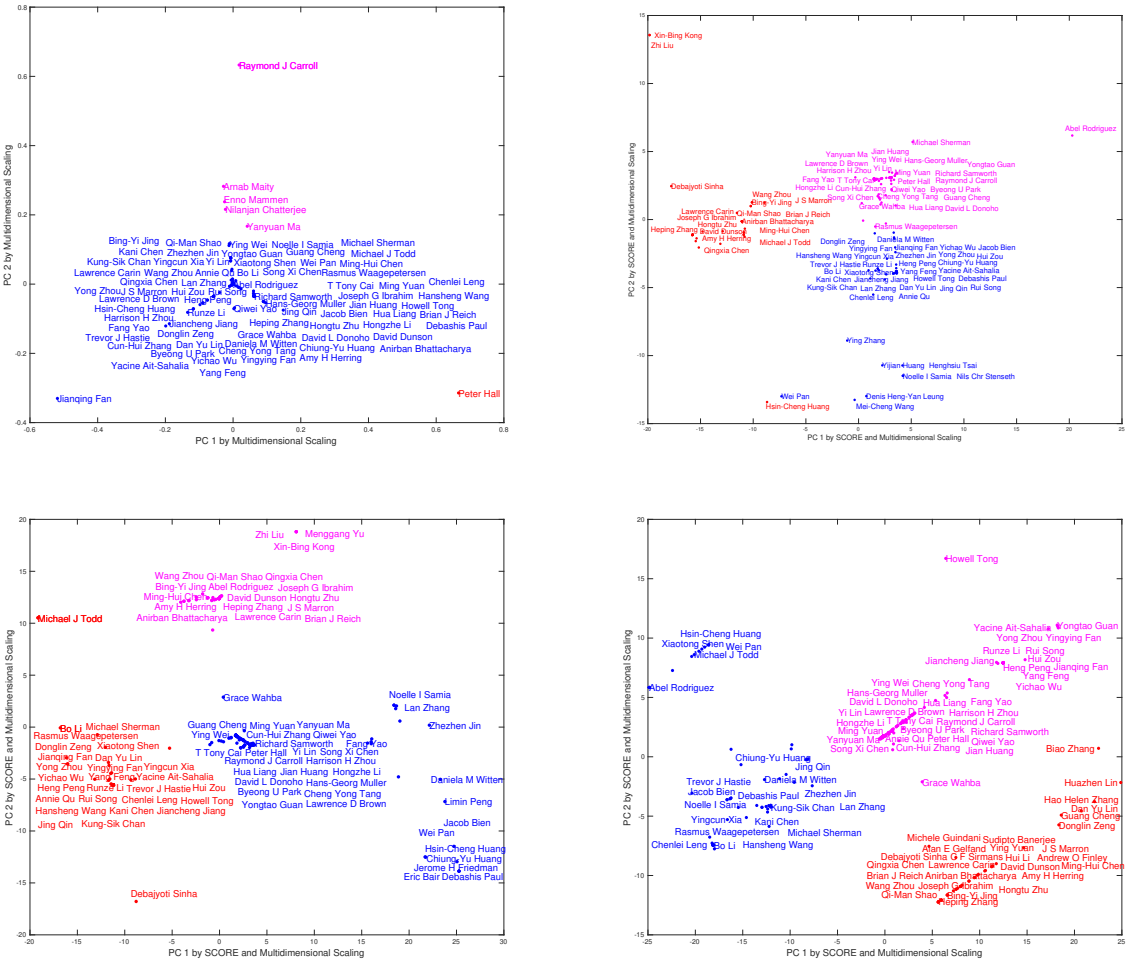
(i, j, j) if the authors i, j co-authored at least one paper. The hyperedges are *undirected* resulting into a symmetric adjacency tensor $\boldsymbol{\mathcal{A}}$. We adopt the framework from Section 5.4 to learn the latent low-rank tensor $\widehat{\boldsymbol{\mathcal{T}}}$ in $\boldsymbol{\mathcal{A}}$, which is used to detect communities in the giant component. We emphasize that our primary goal is to present the new findings by taking into consideration of higher-order interactions among co-authors and applying novel robust tensor methods. It is not our intention to label an author with a certain community.

The Tucker ranks are set as $(4, 4, 4)$ and sparsity ratio α is varied at $\{0, 10^{-4}, 5 \times 10^{-4}\}$. The number of communities is set at $K = 3$ and the algorithm is initialized by the HOSVD of $\boldsymbol{\mathcal{A}}$. To uncover community structures, we apply spectral clustering to the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$. The node degrees are severely heterogeneous with Peter Hall, Jianqing Fan and Raymond Carroll being the top-3 statisticians in terms of # of co-authors. The naive spectral clustering often performs poorly in the existence of heterogeneity, skewing to the high-degree nodes. Indeed, the top-left plot in Figure 8 shows that the naive spectral clustering identifies these three statisticians as the corners in a triangle, and puts Peter Hall in a single community. To mitigate the influence of node heterogeneity, we apply SCORE (Jin, 2015) for community detection, which uses the leading singular vector of $\widehat{\boldsymbol{\mathcal{T}}}$ as normalization.

The community structures found by SCORE are displayed in Figure 8. The top-right plot shows the three clusters identified by SCORE when the sparsity ratio is zero. The three communities are: 1). "North Carolina" group including researchers from Duke University, University of North Carolina and North Carolina State University, together with their close collaborators such as Debajyoti Sinha, Qi-Man Shao, Bing-Yi Jing, Michael J Todd and etc.; 2). "Carroll-Hall" group including researchers in non-parametric and semi-parametric statistics, functional estimation and high-dimensional statistics, together with collaborators; 3). "Fan and Others" group[5] including *primarily* the researchers collaborating closely with Jianqing Fan or his co-authors, and other researchers who do not *obviously* belong to the first two groups. We note that the fields of researchers in "Fan and Others" group are quite diverse, some of which overlap with those in "Carroll-Hall" group and "North Carolina" group. However, unlike the results in (Ji and Jin, 2016), the top-right plot in Figure 8 does not cluster the "Fan and Others" group into either the "North Carolina" group or "Carroll-Hall" group.

We then set the sparsity ratio of $\widehat{\boldsymbol{\mathcal{S}}}$ by $\alpha = 10^{-4}$. The communities identified by SCORE based on the singular vectors of $\widehat{\boldsymbol{\mathcal{T}}}$ are illustrated in the bottom-left plot of Figure 8. Compared with the top-right plot ($\alpha = 0$), the three communities displayed in the bottom-left plot largely remain the same. But the group memberships of some authors do change. Notably, Debajyoti

[5]We name it the "Fan and Others" group simply because many researchers in this group are the co-authors of Jianqing Fan. It is not our intention to rank/label the authors.



(a) Top-left: $\alpha = 10^{-4}$ and naive spectral clustering; top-right: $\alpha = 0$ and SCORE; bottom-left: $\alpha = 10^{-4}$ and SCORE; bottom-right: $\alpha = 5 \times 10^{-4}$ and SCORE.

Figure 8: Sub-structures detected in the "High-Dimensional Data Analysis" community based on the hypergraph co-authorship network. The Tucker ranks are set as $(4, 4, 4)$ with varied sparsity ratio at $\{0, 10^{-4}, 5 \times 10^{-4}\}$ and the algorithm is initialized by the HOSVD of adjacency tensor $\boldsymbol{\mathcal{A}}$.

Sinha and Michael J Todd move from the "North Carolina" group to "Fan and Others" group; Abel Rodriguez moves from the "Carroll-Hall" group to "North-Carolina" group; several authors (e.g. Daniela M Witten, Jacob Bien, Pan Wei, Chiung-Yu Huang, Debashis Paul, Zhezhen Jin, Lan Zhang and etc.) move from the "Fan and Others" group to "Carroll-Hall" group; Hsin-Cheng Huang moves from the "North Carolina" group to "Carroll-Hall" group; Rasmus Waggepetersen moves from the "Carroll-Hall" group to "Fan and Others" group. These changes of memberships suggest that these authors may not have strong ties to the "North Carolina", "Carroll-Hall" group or be the co-authors of Jianqing Fan. It may be more reasonable that these authors constitute a separate group.

This indeed happens when the sparsity ratio α increases to a certain level. The bottom-right plot of Figure 8 shows the clustering result of SCORE when $\alpha = 5 \times 10^{-4}$. Compared with the top-right ($\alpha = 0$) and bottom-left ($\alpha = 10^{-4}$) plots, the community structure has a significant change. Indeed, the "Fan and Others" group now splits into a "Fan" group including Jianqing Fan and his co-authors, and an "Others" group including the researchers who do not have obvious ties with "Fan" group. Moreover, the "Fan" group merges into the "Carroll-Hall" group, which coincides with the clustering result of SCORE when applied onto the graph co-authorship network (Fig. 6 in (Ji and Jin, 2016)). Consequently, we name the three communities in the top-right plot by the "North Carolina", "Carroll-Fan-Hall" and "Others" group. Interestingly, many of the authors in the "Others" group are those whose memberships change when the sparsity ratio α increases from 0 to 10^{-4}. See the top-right and bottom-left plots of Figure 8. In addition, we observe that, as α increases from 10^{-4} to 5×10^{-4}, Donglin Zeng and Dan Yu Lin in the "Fan and Others" group moves to "North Carolina" group. This might be more reasonable since they both work at the University of North Carolina.

9 Discussions

In this paper, we investigate the local convergence of Riemannian gradient descent algorithm in a generalized framework, assuming that a warm initialization of $\boldsymbol{\mathcal{T}}^*$ is readily available. It worth to point out that obtaining a warm initialization under weak signal-to-noise ratio (SNR) condition is generally hard. Indeed, the minimal requirements on SNR (often refereed to as the *computational limits*) for guaranteeing the existence of a polynomial-time initialization algorithm are generally unknown for most low-rank tensor models, even for the simple tensor PCA model. Interested readers are suggested to refer (Zhang and Xia, 2018; Luo and Zhang, 2020; Hopkins et al., 2015; Dudeja and Hsu, 2020; Arous et al., 2018) for the discussions on the computational hardness of tensor PCA. The computational limits are probably more difficult to study under generalized

low-rank tensor models. Nevertheless, under *strong* SNR conditions, there exist polynomial-time algorithms to produce warm initializations for different models, such as the second-order moment method for tensor completion (Xia et al., 2021; Xia and Yuan, 2019), HOSVD for tensor PCA and hypergraph/multi-layer networks (Richard and Montanari, 2014; Zhang and Xia, 2018; Ke et al., 2019; Jing et al., 2020), spectral methods for generalized low-rank tensor models (Han et al., 2020). Undoubtedly, it can be even more difficult to investigate the computational limits under our generalized low-rank plus sparse tensor models, and it is beyond the scope of this paper. Interestingly, we observe that the initialization by a simple HOSVD works satisfactorily on the two real-data examples as well as on the simulations when SNR is reasonably strong.

Note that Algorithm 2 requires the knowledge of *true* Tucker ranks $\mathbf{r}$, which is usually unavailable in practice. Fortunately, for applications where the true ranks are often small, we can simply run Algorithm 2 for multiple times with distinct choices of these ranks and decide the best ones according to certain criterion, e.g., interpretability if no ground truth (Jing et al., 2020; Fan et al., 2021) or the mis-clustering rate if ground truth is available (Ke et al., 2019; Zhou et al., 2013; Wang and Li, 2020). Sometimes, it suffices to take the singular values of the matricizations and decide the cut-off point by the famous *scree plot* (Cattell, 1966). Similarly, the parameters $\alpha, \gamma, \mu_1, \mathsf{k_{pr}}$ in Algorithm 2 are all tuning parameters. In practice, it suffices to set $\gamma = 1$ and $\mathsf{k_{pr}}$ large enough, and Algorithm 2 is often not sensitive to the spikiness-related parameter μ_1. Thus, it is often the sparsity α that most significantly impacts the algorithmic output.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *Annals of Statistics*, 47(4):2117–2144, 2019.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15: 2773–2832, 2014.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *arXiv preprint arXiv:1808.00921*, 2018.

Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019.

Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29:3873–3881, 2016.

Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *Annals of Statistics*, 46(6B):3308–3333, 2018.

Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu. Tensors in statistics. *Annual Review of Statistics and Its Application*, 8, 2020.

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874, 2019a.

Changxiao Cai, H Vincent Poor, and Yuxin Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR, 2020a.

HanQin Cai, Jian-Feng Cai, and Ke Wei. Accelerated alternating projections for robust principal component analysis. *The Journal of Machine Learning Research*, 20(1):685–717, 2019b.

HanQin Cai, Jian-Feng Cai, Tianming Wang, and Guojian Yin. Accelerated structured alternating projections for robust spectrally sparse signal recovery. *IEEE Transactions on Signal Processing*, 2020b.

Jian-Feng Cai, Lizhang Miao, Yang Wang, and Yin Xian. Provable near-optimal low-multilinear-rank tensor recovery. *arXiv preprint arXiv:2007.08904*, 2020c.

T Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1 (2):245–276, 1966.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Elynn Y Chen, Dong Xia, Chencheng Cai, and Jianqing Fan. Semiparametric tensor factor analysis by iteratively projected svd. *arXiv preprint arXiv:2007.02404*, 2020a.

Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208, 2019.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020b.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly optimal robust matrix completion. In *International Conference on Machine Learning*, pages 797–805. PMLR, 2017.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

Rishabh Dudeja and Daniel Hsu. Statistical query lower bounds for tensor pca. *arXiv preprint arXiv:2008.04101*, 2020.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *Annals of Statistics*, 46(6B):3481–3509, 2018.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

Xing Fan, Marianna Pensky, Feng Yu, and Teng Zhang. Alma: Alternating minimization algorithm for clustering mixture multilayer network. *arXiv preprint arXiv:2102.10226*, 2021.

Arvind Ganesh, John Wright, Xiaodong Li, Emmanuel J Candes, and Yi Ma. Dense error correction for low-rank matrices via principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1513–1517. IEEE, 2010.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016.

Rungang Han, Rebecca Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020.

Botao Hao, Anru Zhang, and Guang Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*, 2020.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

Jiaoyang Huang, Daniel Z Huang, Qing Yang, and Guang Cheng. Power iteration for tensor pca. *arXiv preprint arXiv:2012.13669*, 2020.

Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

Jiashun Jin. Fast community detection by score. *Annals of Statistics*, 43(1):57–89, 2015.

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multi-layer networks via regularized tensor decomposition. *arXiv preprint arXiv:2002.04457*, 2020.

Zheng Tracy Ke, Feng Shi, and Dong Xia. Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*, 2019.

Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128. IEEE, 2017.

Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, 2017.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51 (3):455–500, 2009.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, 2015.

Qiuwei Li and Gongguo Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239. IEEE, 2017.

Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2012.

Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. *arXiv preprint arXiv:1702.07449*, 2017.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5249–5257, 2016.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):925–938, 2019.

Yuetian Luo and Anru R Zhang. Tensor clustering with planted structures: Statistical optimality and computational limits. *arXiv preprint arXiv:2005.10743*, 2020.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society Series B*, 79(Part 4):1119–1141, 2016.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Annals of Statistics*, 46(6A):2871–2903, 2018.

Andrea Montanari and Nike Sun. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018.

Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81, 2014.

Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. *arXiv preprint arXiv:1410.7660*, 2014.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 809–816, 2011.

Yuqing Pan, Qing Mai, and Xin Zhang. Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association*, 2018.

Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250, 2020.

Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, 2019.

Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. *arXiv preprint arXiv:1702.06237*, 2017.

Garvesh Raskutti, Ming Yuan, and Han Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

Holger Rauhut, Reinhold Schneider, and Željka Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

Emile Richard and Andrea Montanari. A statistical model for tensor pca. *Advances in Neural Information Processing Systems*, 27:2897–2905, 2014.

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, and Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

Will Wei Sun and Lexin Li. Dynamic tensor clustering. *Journal of the American Statistical Association*, 114(528):1894–1907, 2019.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

Min Tao and Xiaoming Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Nick Vannieuwenhoven, Raf Vandebril, and Karl Meerbergen. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing*, 34(2):A1027–A1052, 2012.

Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR, 2017.

Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of brain networks. *Ann. Appl. Stat.*, 13(1):85–112, 03 2019. doi: 10.1214/18-AOAS1193. URL https://doi.org/10.1214/18-AOAS1193.

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

Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. *arXiv preprint arXiv:1906.03807*, 2019.

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

Raymond KW Wong and Thomas CM Lee. Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research*, 18(1):5404–5428, 2017.

Dong Xia. Normal approximation and confidence region of singular subspaces, 2019.

Dong Xia and Ming Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, 19(6):1265–1313, 2019.

Dong Xia and Fan Zhou. The sup-norm perturbation of hosvd and low rank tensor denoising. *J. Mach. Learn. Res.*, 20:61–1, 2019.

Dong Xia, Anru R Zhang, and Yuchen Zhou. Inference for low-rank tensors–no need to debias. *arXiv preprint arXiv:2012.14844*, 2020.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *Annals of Statistics*, 49(1):76–99, 2021.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *IEEE transactions on information theory*, 58(5):3047–3064, 2012.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381, 2016.

Ming Yuan and Cun-Hui Zhang. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.

Mingao Yuan, Ruiqi Liu, Yang Feng, and Zuofeng Shang. Testing community structures for hypergraphs. *arXiv preprint arXiv:1810.04617*, 2018.

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

Anru R Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. Islet: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science*, 2 (2):444–479, 2020a.

Chenyu Zhang, Rungang Han, Anru R Zhang, and Paul M Voyles. Denoising atomic resolution 4d scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy*, 219:113123, 2020b.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 28:559–567, 2015.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1518–1522. IEEE, 2010.

A Proofs of theorems

A.1 Proof of Theorem 4.1

We prove the theorem by induction on $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and $\|\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ alternatively.

Step 0: Base case First we have $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le c_{1,m} \min\{\frac{\delta^2}{\sqrt{\bar{r}}}, (\kappa_0^{2m}\sqrt{\bar{r}})^{-1}\} \cdot \underline{\lambda}$ from the assumption on initialization error, where the small constant $c_{1,m} > 0$ depends only on m.

Now we estimate $\|\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$. Denote $\Omega_0 = \mathrm{supp}(\widehat{\boldsymbol{\mathcal{S}}}_0)$ and $\Omega^* = \mathrm{supp}(\boldsymbol{\mathcal{S}}^*)$.

For $\forall \omega \in \Omega_0$, from the construction of $\widehat{\boldsymbol{\mathcal{S}}}_0$ in Algorithm 1, we have by the definition of Err_∞,

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0)]_\omega| \le \min_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty} \le \mathsf{k_{pr}}} \|\nabla \mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty} \le \mathsf{Err}_\infty \tag{A.1}$$

From Assumption 3, we get

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0)]_\omega - [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)]_\omega| \ge b_l |[\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*]_\omega|. \tag{A.2}$$

Note that to use (A.2), we shall verify the neighborhood condition. From the initialization error we have $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, as a result of Lemma B.6, we have $\mathsf{Trim}_{\zeta,\mathbf{r}}(\widehat{\boldsymbol{\mathcal{T}}}_0)$ is $(2\mu_1\kappa_0)^2$-incoherent if we choose $\zeta_0 = \frac{16}{7}\mu_1 \frac{\|\widehat{\boldsymbol{\mathcal{T}}}_0\|_{\mathrm{F}}}{\sqrt{d^*}}$. Indeed, from Lemma B.7, we have:

$$|[\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \le C_{1,m} \bar{r}^m \underline{d}^{-(m-1)} (\mu_1\kappa_0)^{4m} \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2.$$

So we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*]_\omega| \le C_{1,m}\sqrt{\frac{\bar{r}^m}{\underline{d}^{m-1}}}(\mu_1\kappa_0)^{2m}\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le C_{1,m}\mu_1^{2m}\sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}}\underline{\lambda},$$

where the last inequality is from the initialization error of $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. As a result, we have

$$\big|[\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*]_\omega\big| \le \big|[\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*]_\omega\big| + |[\widehat{\boldsymbol{\mathcal{S}}}_0]_\omega| + |[\boldsymbol{\mathcal{S}}^*]_\omega| \le C_{1,m}\mu_1^{2m}\sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}}\underline{\lambda} + \mathsf{k_{pr}} + \|\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty}.$$

Thus, both $\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0$ and $\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*$ belong to the ball $\mathbb{B}^*_\infty$ and thus (A.2) holds.

As a result of (A.1) and (A.2), we get for any $\omega \in \Omega_0$

$$b_l \big|[\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*]_\omega\big| \le \big|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)]_\omega\big| + \mathsf{Err}_\infty.$$

Therefore,

$$\begin{aligned}
\|\mathcal{P}_{\Omega_0}(\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 &\le \frac{2}{b_l^2}\|\mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_0|}{b_l^2}\mathsf{Err}_\infty^2 \\
&= \frac{2}{b_l^2}\|\mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)) - \mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)) + \mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_0|}{b_l^2}\mathsf{Err}_\infty^2 \\
&\le \frac{4}{b_l^2}\|\mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)) - \mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{4}{b_l^2}\|\mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 + \frac{2|\Omega_0|}{b_l^2}\mathsf{Err}_\infty^2 \\
&\le \frac{4b_u^2}{b_l^2}\|\mathcal{P}_{\Omega_0}(\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 + \frac{6|\Omega_0|}{b_l^2}\mathsf{Err}_\infty^2, \qquad \text{(A.3)}
\end{aligned}$$

where the last inequality is due to $\|\mathcal{P}_{\Omega_0}(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*))\|_{\mathrm{F}}^2 \le |\Omega_0|\mathsf{Err}_\infty^2$ and Assumption 3 since $\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^* \in \mathbb{B}^*_\infty$.

From (A.3), Lemma B.8, we have

$$\|\mathcal{P}_{\Omega_0}(\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \le \frac{C_{2,m} b_u^2}{b_l^2}(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{6|\Omega_0|}{b_l^2}\mathsf{Err}_\infty^2 \qquad \text{(A.4)}$$

here $C_{2,m} > 0$ is an absolute constant depending only on m.

For $\forall \omega = (\omega_1, \ldots, \omega_m) \in \Omega^* \backslash \Omega_0$, we have $|[\widehat{\boldsymbol{\mathcal{S}}}_l - \boldsymbol{\mathcal{S}}^*]_\omega| = |[\boldsymbol{\mathcal{S}}^*]_\omega|$. Since the loss function is entry-wise by Assumption 3, we have $[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega = [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0)]_\omega$. Clearly, $\widehat{\boldsymbol{\mathcal{T}}}_0$ and $\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*$ both belong to $\mathbb{B}^*_\infty$, by Assumption 3 we get

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega - [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)]_\omega| \ge b_l |[\boldsymbol{\mathcal{S}}^*]_\omega|.$$

Now we bound $|[\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*]_\omega|$ as follows. For any $\omega \in \Omega^* \backslash \Omega_0$,

$$\begin{aligned}
|[\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*]_\omega| = |[\boldsymbol{\mathcal{S}}^*]_\omega| &\le \frac{1}{b_l}|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega - [\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)]_\omega| \\
&\le \frac{1}{b_l}\Big(|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega| + |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)]_\omega|\Big) \\
&\le \frac{1}{b_l}\Big(|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega| + |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*) - \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega| + |[\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega|\Big) \\
&\le \frac{1}{b_l}|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega| + \frac{b_u}{b_l}|[\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*]_\omega| + \frac{1}{b_l}\mathsf{Err}_\infty,
\end{aligned}$$

where the last inequality is again due to Assumption 3 since $\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^* \in \mathbb{B}^*_\infty$. Therefore we have

$$\|\mathcal{P}_{\Omega^* \backslash \Omega_0}(\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \le \frac{3}{b_l^2}\|\mathcal{P}_{\Omega^* \backslash \Omega_0}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0))\|_{\mathrm{F}}^2 + \frac{3b_u^2}{b_l^2}\|\mathcal{P}_{\Omega^* \backslash \Omega_0}(\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 + \frac{3}{b_l^2}|\Omega^* \backslash \Omega_0|\mathsf{Err}_\infty^2 \qquad \text{(A.5)}$$

Since $\omega \in \Omega^* \backslash \Omega_0$, we have

$$|[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega| \leq \max_{i=1}^m |\mathbf{e}_{\omega_i}^\top \mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0))|^{(\gamma \alpha d_i^-)} \tag{A.6}$$

Now since we have $\boldsymbol{\mathcal{S}}^* \in \mathbb{S}_\alpha$, we have

$$\begin{aligned} |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega| &\leq \max_{i=1}^m |\mathbf{e}_{\omega_i}^T \mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*))|^{((\gamma-1)\alpha d_i^-)} \\ &\leq \max_{i=1}^m \left| \mathbf{e}_{\omega_i}^\top \left(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)) \right) \right|^{((\gamma-1)\alpha d_i^-)} + \mathsf{Err}_\infty \end{aligned} \tag{A.7}$$

Using AM-GM inequality, we have:

$$\begin{aligned} |[\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0)]_\omega|^2 &\leq 2 \max_{i=1}^m \frac{\left\| \mathbf{e}_{\omega_i}^\top \left(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)) \right) \right\|_{\mathrm{F}}^2}{(\gamma-1)\alpha d_i^-} + 2\mathsf{Err}_\infty^2 \\ &\leq 2 \sum_{i=1}^m \frac{\left\| \mathbf{e}_{\omega_i}^\top \left(\mathcal{M}_i(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*)) - \mathcal{M}_i(\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)) \right) \right\|_{\mathrm{F}}^2}{(\gamma-1)\alpha d_i^-} + 2\mathsf{Err}_\infty^2 \end{aligned} \tag{A.8}$$

Now for all fixed $i \in [m]$, for all $\omega_i \in [d_i]$, ω_i appears at most αd_i^- times since $\Omega^* \backslash \Omega_0$ is an α-fraction set. This observation together with (A.8) lead to the following:

$$\begin{aligned} \|\mathcal{P}_{\Omega^* \backslash \Omega_0}(\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0))\|_{\mathrm{F}}^2 &\leq 2 \sum_{i=1}^m \frac{\|\nabla \mathfrak{L}(\widehat{\boldsymbol{\mathcal{T}}}_0 + \boldsymbol{\mathcal{S}}^*) - \nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2}{\gamma - 1} + 2|\Omega^* \backslash \Omega_0| \mathsf{Err}_\infty^2 \\ &\leq \frac{2mb_u^2}{\gamma - 1} \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + 2|\Omega^* \backslash \Omega_0| \mathsf{Err}_\infty^2. \end{aligned} \tag{A.9}$$

Therefore together with (A.5) and (A.9) and Lemma B.8, we have

$$\|\mathcal{P}_{\Omega^* \backslash \Omega_0}(\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \leq \left(C_{3,m} \frac{b_u^2}{b_l^2} \frac{1}{\gamma - 1} + C_{4,m} \frac{b_u^2}{b_l^2} (\mu_1 \kappa_0)^{4m} \bar{r}^m \alpha \right) \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{9}{b_l^2} |\Omega^* \backslash \Omega_0| \mathsf{Err}_\infty^2 \tag{A.10}$$

where $C_{3,m}, C_{4,m} > 0$ are constants depending only on m. Now we combine (A.4) and (A.10) and we get

$$\|\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \leq \left(C_{3,m} \frac{b_u^2}{b_l^2} \frac{1}{\gamma - 1} + C_{5,m} (\mu_1 \kappa_0)^{4m} \bar{r}^m \frac{b_u^2}{b_l^2} \alpha \right) \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_0| \mathsf{Err}_\infty^2 \tag{A.11}$$

where $C_{5,m} > 0$ depending only on m and $C_1 > 0$ an absolute constant.

Now if we choose $\alpha \leq (C_{5,m} \kappa_0^{4m} \mu_0^{4m} \bar{r}^m \frac{b_u^4}{b_l^4})^{-1}$ and $\gamma - 1 \geq C_{3,m} \frac{b_u^4}{b_l^4}$ for some sufficient large constants $C_{3,m}, C_{5,m} > 0$ depending only on m, then we have

$$\|\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \leq 0.0001 \frac{b_l^2}{b_u^2} \|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_0| \mathsf{Err}_\infty^2 \tag{A.12}$$

and

$$\|\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \leq 0.01\frac{b_l}{b_u}\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_1}{b_l}\sqrt{|\Omega^* \cup \Omega_0|}\mathsf{Err}_\infty \tag{A.13}$$

In addition, by the signal-to-noise ratio condition in Theorem 4.1, (A.13) implies that $\|\widehat{\boldsymbol{\mathcal{S}}}_0 - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \leq c_0\underline{\lambda}$ for a small $c_0 > 0$. This fact is helpful later since it implies that $\widehat{\boldsymbol{\mathcal{T}}}_0 + \widehat{\boldsymbol{\mathcal{S}}}_0$ belongs to the ball $\mathbb{B}_2^*$ and thus activates the conditions in Assumption 2.

Step 1: bounding $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$ for all $l \geq 1$. To use Lemma B.6, we need to check the conditions required by the lemma. First from the previous step, we have

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \leq 0.0001\frac{b_l^2}{b_u^2}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2}|\Omega^* \cup \Omega_{l-1}|\mathsf{Err}_\infty^2. \tag{A.14}$$

This also implies $\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \leq c_0\underline{\lambda}$ for a small $c_0 > 0$.

Step 1.1: bounding $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. From the Algorithm 2, we have for arbitrary $1 \geq \delta > 0$,

$$\begin{aligned}\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*) - \beta\mathcal{P}_{\mathbb{T}_{l-1}}\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}^2 \\ &\leq (1 + \frac{\delta}{2})\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta})\beta^2\|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \end{aligned} \tag{A.15}$$

Now we consider the bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2$,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta\langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\rangle \\ &\quad + \beta^2\|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \end{aligned} \tag{A.16}$$

The upper bound of $\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ ensures that $\widehat{\boldsymbol{\mathcal{T}}}_{l-1} + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} \in \mathbb{B}_2^*$. Using the smoothness condition in Assumption 2, we get

$$\beta^2\|\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \leq \beta^2 b_u^2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \tag{A.17}$$

Now we consider the bound for $|\langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\rangle|$. First we have:

$$\langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\rangle = \langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle - \langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}^{\perp}(\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\rangle.$$

The estimation of $\langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle$ is as follows:

$$\begin{aligned}\langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle &= \langle\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle - \langle\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle \\ &\geq b_l\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* + \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 - \langle\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*\rangle, \end{aligned} \tag{A.18}$$

where the last inequality follows from Assumption 2. And the estimation of $\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*, \mathcal{P}^{\perp}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*)\rangle$ is as follows:

$$|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*, \mathcal{P}^{\perp}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*)\rangle| \le \|\mathcal{P}^{\perp}_{\mathbb{T}_{l-1}}(\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}\|\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}} \le \frac{C_{1,m}b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \tag{A.19}$$

where the last inequality follows from Lemma B.1. Together with (A.18) and (A.19), we get,

$$\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*)\rangle \ge b_l\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 - \langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\rangle - \frac{C_{1,m}b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \tag{A.20}$$

Together with (A.17) and (A.20), we get

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*-\beta\mathcal{P}_{\mathbb{T}_{l-1}}(\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \le{}& \left(1+2\beta b_u\frac{C_{1,m}}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\right)\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \\ &+(\beta^2b_u^2-2\beta b_l)\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \\ &+2\beta|\langle \widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*, \boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\rangle|\end{aligned} \tag{A.21}$$

In order to bound (A.21), we derive separately the bound for each terms.

Bounding $\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2$. From the bound for $\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ in (A.14), we get,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \le 2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+2\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 \le (2.0002\frac{b_l^2}{b_u^2})\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{C_1}{b_l^2}|\Omega^*\cup\Omega_{l-1}|\mathsf{Err}_\infty^2 \tag{A.22}$$

Thus,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \le 2\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}+\frac{C_1}{b_l}\sqrt{|\Omega^*\cup\Omega_{l-1}|}\mathsf{Err}_\infty \tag{A.23}$$

Bounding $|\langle \boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\rangle|$. We first bound $\|\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}$ by (A.23):

$$\|\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}} \le b_u\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*+\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \le 2b_u\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}+\frac{C_1b_u}{b_l}\sqrt{|\Omega^*\cup\Omega_{l-1}|}\mathsf{Err}_\infty \tag{A.24}$$

Now we estimate $|\langle \boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\rangle|$ from (A.14) and (A.24) as follows,

$$|\langle \boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\rangle| \le \|\boldsymbol{\mathcal{G}}_{l-1}-\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}\|\widehat{\boldsymbol{\mathcal{S}}}_{l-1}-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \le (0.02b_l+0.01\beta b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{1}{\beta}\frac{C_1}{b_l^2}|\Omega^*\cup\Omega_{l-1}|\mathsf{Err}_\infty^2+\frac{C_1b_u}{b_l^2}|\Omega^*\cup\Omega_{l-1}|\mathsf{Err}_\infty^2 \tag{A.25}$$

Bounding $|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^* \rangle|$. From (A.14), we have:

$$\begin{aligned}
|\langle \widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*, \widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^* \rangle| &\le \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \|\widehat{\boldsymbol{\mathcal{S}}}_{l-1} - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\
&\le (0.01 \frac{b_l}{b_u} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_1}{b_l} \sqrt{|\Omega^* \cup \Omega_{l-1}|} \mathsf{Err}_\infty) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \\
&\le 0.02 \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{C_1}{b_l^2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \qquad \text{(A.26)}
\end{aligned}$$

Now we go back to (A.21) and from (A.22) - (A.26), we get:

$$\begin{aligned}
&\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\
&\le \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + C_1 (1 + b_u + b_u^2) b_l^{-2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \qquad \text{(A.27)}
\end{aligned}$$

where the condition $\underline{\lambda} \ge \frac{C_{1,m} b_u}{b_l} \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ is used in the last step.

By combining (A.16) and (A.27), we get

$$\begin{aligned}
\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}} \boldsymbol{\mathcal{G}}_{l-1}\|_{\mathrm{F}}^2 \\
&\le (1 + \frac{\delta}{2}) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^* - \beta \mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{G}}_{l-1} - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \|\mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\
&\le (1 + \frac{\delta}{2}) \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \mathsf{Err}_{2\mathbf{r}}^2 \\
&\quad + C_1 \left(1 + \beta b_u + \beta^2 b_u^2\right) b_u^{-2} |\Omega^* \cup \Omega_{l-1}| \mathsf{Err}_\infty^2 \\
&\le (1 + \frac{\delta}{2}) \left(1 - 1.84\beta b_l + 5\beta^2 b_u^2\right) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1 + \frac{2}{\delta}) \beta^2 \mathsf{Err}_{2\mathbf{r}}^2 \\
&\quad + C_1 \left(1 + \beta b_u + \beta^2 b_u^2\right) \frac{1}{b_u^2} \left(|\Omega^*| + \gamma \alpha d^*\right) \mathsf{Err}_\infty^2 \qquad \text{(A.28)}
\end{aligned}$$

where in the second inequality we used

$$\|\mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}} = \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}=1} \langle \mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{G}}^*), \boldsymbol{\mathcal{Y}} \rangle = \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}=1} \langle \boldsymbol{\mathcal{G}}^*, \mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{Y}}) \rangle \le \mathsf{Err}_{2\mathbf{r}} \qquad \text{(A.29)}$$

since $\mathcal{P}_{\mathbb{T}_{l-1}} (\boldsymbol{\mathcal{Y}}) \in \mathbb{M}_{2\mathbf{r}}$ and in the last inequality we use $|\Omega^* \cup \Omega_{l-1}| \le |\Omega^*| + |\Omega_{l-1}| \le |\Omega^*| + \gamma \alpha d^*$.

Now we choose proper $\beta \in [0.005 b_l / (b_u^2), 0.36 b_l / (b_u^2)]$ so $1 - 1.84\beta b_l + 5\beta^2 b_u^2 \le 1 - \delta$, and we get

$$\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le (1 - \delta)(1 + \delta/2) \|\widehat{\boldsymbol{\mathcal{T}}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + 3\delta^{-1} \mathsf{Err}_{2\mathbf{r}} + C_1 (b_u + 1) b_l^{-1} \sqrt{|\Omega^*| + \alpha \gamma d^*} \mathsf{Err}_\infty \qquad \text{(A.30)}$$

where we use the fact that $\beta \le 1$. From the signal-to-noise ratio condition, we have $3\delta^{-1} \mathsf{Err}_{2\mathbf{r}} + C_1 (b_u + 1) b_l^{-1} \sqrt{|\Omega^*| + \alpha \gamma d^*} \mathsf{Err}_\infty \le \frac{\delta}{4} \frac{\underline{\lambda}}{C_m \sqrt{r}}$. This implies that $\|\boldsymbol{\mathcal{W}}_{l-1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$ holds.

Step 1.2: bounding $\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. From the Algorithm 2, $\widehat{\boldsymbol{\mathcal{T}}}_l=\mathsf{Trim}_{\zeta_l,\mathbf{r}}(\boldsymbol{\mathcal{W}}_{l-1})$. From the bound (A.30), we have $\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le\underline{\lambda}/8$. Now from Lemma B.6, we get,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2&=\|\mathsf{Trim}_{\zeta_l,\mathbf{r}}(\boldsymbol{\mathcal{W}}_{l-1})-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\\&\le\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+C_m\frac{\sqrt{\bar r}}{\underline{\lambda}}\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3\\&\le(1+\frac{\delta}{4})\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\\&\le(1-\delta^2)\|\widehat{\boldsymbol{\mathcal{T}}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+6\delta^{-1}\mathsf{Err}_{2\mathbf{r}}+C_1\left(1+b_u+b_u^2\right)b_l^{-2}\left(|\Omega^*|+\gamma\alpha d^*\right)\mathsf{Err}_\infty^2\end{aligned}\tag{A.31}$$

Also, from (A.31) and the signal-to-noise ration condition, we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le c_1\min\{\delta^2\bar r^{-1/2},\kappa_0^{-2m}\bar r^{-1/2}\}\cdot\underline{\lambda}.$$

This finishes the induction for the error $\|\widehat{\boldsymbol{\mathcal{T}}}_l-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$.

Step 2: bounding $\|\widehat{\boldsymbol{\mathcal{S}}}_l-\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}$ for all $l\ge1$. First we denote $\Omega_l=\mathrm{supp}(\widehat{\boldsymbol{\mathcal{S}}}_l)$ and $\Omega^*=\mathrm{supp}(\boldsymbol{\mathcal{S}}^*)$. The proof for the case when $l\ge1$ is almost the same as the case when $l=0$. The only difference is that now we are updating $\widehat{\boldsymbol{\mathcal{T}}}_l$ by $\widehat{\boldsymbol{\mathcal{T}}}_l=\mathsf{Trim}_{\zeta_l,\mathbf{r}}(\boldsymbol{\mathcal{W}}_{l-1})$. So we need to ensure $\|\boldsymbol{\mathcal{W}}_{l-1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le\frac{\underline{\lambda}}{8}$, but this is guaranteed by (A.30) in Step 1.1. Now replace all the subscripts 0 by l in Step 0 and we finish the proof.

A.2 Proof of Theorem 4.3

Let $\widehat{\Omega}$ and Ω^* denote the support of $\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}$ and $\boldsymbol{\mathcal{S}}^*$, respectively. By the proof of Theorem 4.1, we have

$$\left|[\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}-\boldsymbol{\mathcal{S}}^*]_\omega\right|\le\begin{cases}\frac{b_u}{b_l}\left|[\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}-\boldsymbol{\mathcal{T}}^*]_\omega\right|+\frac{2\mathsf{Err}_\infty}{b_l} & ,\text{ if }\omega\in\widehat{\Omega}\\ \frac{2b_u}{b_l}\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}-\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty}+\frac{2\mathsf{Err}_\infty}{b_l} & ,\text{ if }\omega\in\Omega^*\setminus\widehat{\Omega}\end{cases}$$

Therefore, we conclude that

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}-\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty}\le\frac{2b_u}{b_l}\left\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}-\boldsymbol{\mathcal{T}}^*\right\|_{\ell_\infty}+\frac{2\mathsf{Err}_\infty}{b_l}.\tag{A.32}$$

Now, we can apply Lemma B.7 and we obtain

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}-\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty}\le C_{1,m}\bar r^{m/2}\underline{d}^{-(m-1)/2}\mu_1^{2m}\kappa_0^{2m}\|\widehat{\boldsymbol{\mathcal{T}}}_{l_{\max}}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\tag{A.33}$$

Now, by putting together (A.32), (A.33) and (4.8), we get

$$\|\widehat{\boldsymbol{\mathcal{S}}}_{l_{\max}}-\boldsymbol{\mathcal{S}}^*\|_{\ell_\infty}\le C_{2,m}\kappa_0^{2m}\mu_1^{2m}\Big(\frac{\bar r^m}{\underline{d}^{m-1}}\Big)^{1/2}\cdot\left(\mathsf{Err}_{2\mathbf{r}}+(|\Omega^*|+\gamma\alpha d^*)^{1/2}\mathsf{Err}_\infty\right)+\frac{2\mathsf{Err}_\infty}{b_l},$$

where $C_{1,m}$ and $C_{2,m}$ are constants depending only on m. Now since we assume $b_l,b_u=O(1)$, we finish the proof of Theorem 4.3.

A.3 Proof of Theorem 5.1

We first estimate the probability of the following two events.

$$\mathsf{Err}_{2\mathbf{r}} \le C_{0,m}\sigma_z \cdot (\bar{d}\bar{r} + r^*)^{1/2} \tag{A.34}$$

$$\mathsf{Err}_{\infty} \le C'_{0,m}\sigma_z \log^{1/2} \bar{d} \tag{A.35}$$

for some constants $C_{0,m}, C'_{0,m} > 0$ depending only on m. Notice here the first event (A.34) holds with probability at least $1-\exp(-c_m \bar{r}\bar{d})$ by Lemma B.3. And for the second event (A.35), we have from the definition,

$$\mathsf{Err}_{\infty} = \max\Big\{\|\nabla \mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)\|_{\ell_\infty}, \min_{\|\boldsymbol{\mathcal{X}}\|_{\ell_\infty}\le\infty} \|\nabla \mathfrak{L}(\boldsymbol{\mathcal{X}})\|_{\ell_\infty}\Big\} = \|\boldsymbol{\mathcal{Z}}\|_{\ell_\infty} \tag{A.36}$$

So we have (A.35) holds with probability at least $1-0.5\bar{d}^{-2}$ from Lemma B.4. Taking union bounds and we get both (A.35) and (A.34) hold with probability at least $1-\bar{d}^{-2}$. And finally applying Theorem 4.1 and Theorem 4.3 gives the desired result.

A.4 Proof of Lemma 5.2

For each $j \in [m]$ and $i \in [d_j]$, we have

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{S}}_\alpha)\|_{\ell_0} = \sum\nolimits_{\omega:\omega_j=i} \mathbb{1}\big(|[\boldsymbol{\mathcal{Z}}]_\omega| > \alpha\sigma_z\big) = \sum\nolimits_{\omega:\omega_j=i} [\boldsymbol{\mathcal{Y}}]_\omega$$

where $\boldsymbol{\mathcal{Y}} \in \{0,1\}^{d_1\times\cdots\times d_m}$ having $i.i.d.$ Bernoulli entries and $q := \mathbb{P}([\boldsymbol{\mathcal{Y}}]_\omega = 1) = \mathbb{P}(|[\boldsymbol{\mathcal{Z}}]_\omega| > \alpha\sigma_z) \le \alpha^{-\theta}$.

Denote $X_{ij} = \sum_{\omega:\omega_j=i}[\boldsymbol{\mathcal{Y}}]_\omega$. By Chernoff bound, if $d_j^- q \ge 3\log(m\bar{d}^3)$, we get

$$\mathbb{P}\Big(X_{ij} - d_j^- q \ge d_j^- q\Big) \le \exp\big\{-d_j^{-1} q/3\big\} \le (m\bar{d}^3)^{-1}$$

implying that

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j}\big\{X_{ij} \le 2d_j^- q\big\}\Big) \ge 1 - m\bar{d}(m\bar{d}^3)^{-1} = 1-\bar{d}^{-2}. \tag{A.37}$$

On the other hand, if $d_j^- q \le 3\log(m\bar{d}^3)$, by Chernoff bound, we get

$$\mathbb{P}\Big(X_{ij} \ge 10\log(m\bar{d}^3)\Big) \le (m\bar{d}^3)^{-1}$$

implying that

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j}\big\{X_{ij} \le 10\log(m\bar{d}^3)\big\}\Big) \ge 1 - m\bar{d}(m\bar{d}^3)^{-1} = 1-\bar{d}^{-2}. \tag{A.38}$$

Putting (A.37) and (A.38), since $q \le \alpha^{-\theta}$, we get

$$\mathbb{P}\Big(\bigcap\nolimits_{i,j}\Big\{X_{ij} \le \max\big\{10\log(m\bar{d}^3), 2d_j^-\alpha^{-\theta}\big\}\Big\}\Big) \ge 1-\bar{d}^{-2},$$

which completes the proof.

A.5 Proof of Theorem 5.3

Conditioned on $\mathfrak{E}_1$ defined in Lemma 5.2, Theorem 5.3 is a special case of Theorem 5.1. Indeed, in Theorem 5.1, we replace σ_z with $\alpha\sigma_z$, and $|\Omega^*|\log\bar{d}$ with $\alpha' d^* \asymp \bar{d}\log(m\bar{d})$, then we get Theorem 5.3.

A.6 Proof of Lemma 5.4

From Lemma B.6, we have $\mathsf{Trim}_{\eta,\mathbf{r}}(\boldsymbol{\mathcal{W}})$ is $2\mu_1\kappa_0$-incoherent. Now for all $j \in [m]$,

$$\|\mathcal{M}_j(\mathscr{H}^{\mathrm{HO}}_{\mathbf{r}}(\widetilde{\boldsymbol{\mathcal{W}}}))\| \le \|\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\| \le \|\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\| + \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \frac{9}{8}\bar{\lambda}.$$

So we conclude

$$\|\mathsf{Trim}_{\eta,\mathbf{r}}(\boldsymbol{\mathcal{W}})\|_{\ell_\infty} \le \frac{9}{8}\bar{\lambda}\prod_{i=1}^{m}(2\mu_1\kappa_0)\sqrt{\frac{r_j}{d_j}} \le (9\zeta/16)\cdot(\mu_1\kappa_0)^m.$$

where the last inequality follows from the upper bound for $\bar{\lambda}$. This finishes the proof of the lemma.

A.7 Proof of Theorem 5.5

We first estimate Err_∞ and $\mathsf{Err}_{2\mathbf{r}}$. From (5.9), we have $\mathsf{Err}_\infty \le L_\zeta$. Now we estimate $\mathsf{Err}_{2\mathbf{r}}$. In fact, from the definition of $\mathsf{Err}_{2\mathbf{r}}$, we have

$$\mathsf{Err}_{2\mathbf{r}} = \sup_{\boldsymbol{\mathcal{M}}\in\mathbb{M}_{2\mathbf{r}},\|\boldsymbol{\mathcal{M}}\|_{\mathrm{F}}\le 1}\langle\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*), \boldsymbol{\mathcal{M}}\rangle.$$

Since for all $\omega \in [d_1]\times\ldots\times[d_m]$, we have $[\nabla\mathfrak{L}(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{S}}^*)]_\omega$ is bounded random variable with the upper bound given by L_ζ. So apply Lemma B.3, we have $\mathsf{Err}_{2\mathbf{r}} \le CL_\zeta\cdot(\bar{d}\bar{r} + r^*)^{1/2}$ with with probability at least $1-\bar{d}^{-2}$. Now we plug in the bounds for Err_∞ and $\mathsf{Err}_{2\mathbf{r}}$ to Theorem 4.1 and we get the first part of the theorem. For the ℓ_∞ bound, we apply Theorem 4.3 and Lemma B.7. And we finish the proof of the theorem.

A.8 Proof of Theorem 5.6

As illustrated in the proof of Theorem 5.5, the convergence of Algorithm has been proven in Theorem 4.1. Thus, we can set, in Theorem 4.1, $b_l = b_u = 1, \delta = 0.15, \bar{r} = K$ and $\kappa_0 = O(1)$. It suffices to check that the conditions of Theorem 4.1 hold.

The initialization condition clearly holds. For the signal-to-noise ratio condition, we have $\mathsf{Err}_\infty = 1$ by definition. For the term $\mathsf{Err}_{2\mathbf{r}}$ (we abuse the notation here since we know $\mathbf{r} = (K,\cdots,K)$), we emphasize that, although the definition (4.5) is used, eq. (A.29) in the proof of Theorem 4.1 implies that the key quantity is

$$\widetilde{\mathsf{Err}_{2\mathbf{r}}} := \max_{1\le l\le l_{\max}}\sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\le 1}\left\langle\boldsymbol{\mathcal{A}} - \mathbb{E}\boldsymbol{\mathcal{A}}, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}})\right\rangle \tag{A.39}$$

where we used the fact $\boldsymbol{\mathcal{G}}^* = \boldsymbol{\mathcal{A}} - \mathbb{E}\boldsymbol{\mathcal{A}}$. It therefore suffices to bound (A.39). By the definition of hGSBM, the supports of $\boldsymbol{\mathcal{T}}^*$ and $\boldsymbol{\mathcal{S}}^*$ are disjoint. Thus, we can write $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}_1 + \boldsymbol{\mathcal{A}}_2$ where

$$[\boldsymbol{\mathcal{A}}_1]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}(\nu_n[\boldsymbol{\mathcal{T}}^*]_\omega) \quad \text{and} \quad [\boldsymbol{\mathcal{A}}_2]_\omega \overset{\text{ind.}}{\sim} \text{Bernoulli}([\boldsymbol{\mathcal{S}}^*]_\omega).$$

Then, we get

$$\begin{aligned}\widetilde{\mathsf{Err}_{2\mathbf{r}}} &= \max_{1\le l\le l_{\max}} \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\le 1} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}})\big\rangle + \|\boldsymbol{\mathcal{A}}_2 - \mathbb{E}\boldsymbol{\mathcal{A}}_2\|_{\mathrm{F}} \\ &\le \max_{1\le l\le l_{\max}} \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\le 1} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}})\big\rangle + (\alpha m n_o n^{m-1})^{1/2}, \end{aligned} \tag{A.40}$$

where the last inequality is due to (2.8). To bound the first term in RHS of (A.40), we take advantage of the incoherence of $\widehat{\boldsymbol{\mathcal{T}}}_l$.

Without loss of generality, we denote $\widehat{\boldsymbol{\mathcal{T}}}_l = \widehat{\boldsymbol{\mathcal{C}}}_l \cdot (\widehat{\mathbf{U}}_{l,1}, \cdots, \widehat{\mathbf{U}}_{l,m})$ the Tucker decomposition of $\widehat{\boldsymbol{\mathcal{T}}}_l$. By property (3.1), for any $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n\times\cdots\times n}$ with $\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}} \le 1$, there exist $\boldsymbol{\mathcal{D}}_{l,0} \in \mathbb{R}^{K\times\cdots\times K}$, $\mathbf{W}_i \in \mathbb{R}^{n\times K}$ and $\boldsymbol{\mathcal{D}}_{l,i} \in \mathbb{R}^{K\times\cdots\times K}$ for all $i \in [m]$ satisfying

$$\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}}) = \boldsymbol{\mathcal{D}}_{l,0} \times_{i\in[m]} \widehat{\mathbf{U}}_{l,i} + \sum_{i=1}^m \boldsymbol{\mathcal{D}}_{l,i} \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i,$$

where $\|\boldsymbol{\mathcal{D}}_{l,0}\|_{\mathrm{F}} \le 1$, $\boldsymbol{\mathcal{W}}_i$ has orthogonal columns, $\mathbf{W}_i^\top \widehat{\mathbf{U}}_{l,i} = \mathbf{0}$ and $\|\mathbf{W}_i \mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}} \le 1$ for all $i \in [m]$. The last requirements are due to $\mathbf{W}_i^\top \widehat{\mathbf{U}}_{l,i} = \mathbf{0}$.

Then, we write

$$\begin{aligned}\max_{1\le l\le l_{\max}} \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\le 1} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}})\big\rangle &\le \max_{1\le l\le l_{\max}} \sup_{\|\boldsymbol{\mathcal{D}}_{l,0}\|_{\mathrm{F}}\le 1} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{D}}_{l,0} \times_{i\in[m]} \widehat{\mathbf{U}}_{l,i}\big\rangle \\ &+ \sum_{i=1}^m \max_{1\le l\le l_{\max}} \sup_{\substack{\|\mathbf{W}_i\mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}}\le 1 \\ \mathbf{W}_i \text{ has orthogonal columns}}} \Big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{D}}_{l,i} \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i\Big\rangle. \end{aligned} \tag{A.41}$$

We bound each terms in RHS of (A.41). Write

$$\begin{aligned}&\max_{1\le l\le l_{\max}} \sup_{\substack{\|\mathbf{W}_i\mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}}\le 1 \\ \mathbf{W}_i \text{ has orthogonal columns}}} \Big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{D}}_{l,i} \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i\Big\rangle \\ &= \max_{1\le l\le l_{\max}} \sup_{\substack{\|\mathbf{W}_i\mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}}\le 1 \\ \mathbf{W}_i \text{ has orthogonal columns}}} \sum_{\omega=(\omega_1,\cdots,\omega_m)\in[K]^m} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, [\boldsymbol{\mathcal{D}}_{l,i}]_\omega \otimes_{j\in[m]\setminus i} [\widehat{\mathbf{U}}_{l,j}]_{:\omega_j} \otimes_i [\mathbf{W}_i]_{:\omega_i}\big\rangle, \end{aligned}$$

where $\otimes$ denotes tensor product so that $\mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \mathbf{u}_3 \in \mathbb{R}^{d_1\times d_2\times d_3}$ if $\mathbf{u}_j \in \mathbb{R}^{d_j}$ for $j = 1, 2, 3$. For each $\omega = (\omega_1, \cdots, \omega_m) \in [K]^m$, $[\boldsymbol{\mathcal{D}}_{l,i}]_\omega \otimes_{j\in[m]\setminus i} [\widehat{\mathbf{U}}_{l,j}]_{:\omega_j} \otimes_i [\mathbf{W}_i]_{:\omega_i}$ is a rank-one tensor with Frobenius norm upper bounded by 1. Meanwhile, due to Lemma 5.4, $\widehat{\boldsymbol{\mathcal{T}}}_l$ is incoherent implying that

$$\big\|[\widehat{\mathbf{U}}_{l,j}]_{:\omega_j}\big\|_{\ell_\infty} = O\big(\sqrt{K/n}\big) \quad \forall j \in [m] \setminus i.$$

Therefore, we get

$$\max_{1\le l\le l_{\max}} \sup_{\substack{\|\mathbf{W}_i\mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}}\le 1\\ \mathbf{W}_i \text{ has orthogonal columns}}} \Big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{D}}_{l,i} \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i \Big\rangle \le K^m \cdot \max_{j=1,\cdots,m} \big\|\boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1\big\|_{j,O(\sqrt{K/n})},$$

where the tensor incoherent norm $\|\cdot\|_{j,\delta}$, for $j\in[m]$ and $\delta\in(0,1)$, is defined by

$$\|\boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1\|_{j,\delta} := \sup_{\substack{\|\mathbf{u}_i\|_{\ell_2}\le 1,\forall i\in[m]\\ \|\mathbf{u}_i\|_{\ell_\infty}\le\delta,\forall i\in[m]\setminus j}} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \mathbf{u}_1\otimes\cdots\otimes\mathbf{u}_m\big\rangle.$$

The tensor incoherent norm was initially proposed in (Yuan and Zhang, 2017), and later extended in network analysis by (Ke et al., 2019; Jing et al., 2020). Since $\boldsymbol{\mathcal{A}}_1$ is symmetric, by (Ke et al., 2019, Theorem 5, by setting $\theta_i = 1$ there), under the condition $\nu_n n^{m-1} \ge C_{2,m}\log n$ for a large constant $C_{2,m}>0$ depending only on m, we get with probability at least $1-n^{-2}$ that

$$\max_{j=1,\cdots,m} \big\|\boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1\big\|_{j,O\left(\sqrt{K/n}\right)} \le C_{3,m}\sqrt{\nu_n n}\cdot K^{(m-1)/2}(\log n)^{(m+8)/2},$$

where $C_{3,m}>0$ depends on m only. Therefore, with the same probability, we have

$$\max_{1\le l\le l_{\max}} \sup_{\substack{\|\mathbf{W}_i\mathcal{M}_i(\boldsymbol{\mathcal{D}}_{l,i})\|_{\mathrm{F}}\le 1\\ \mathbf{W}_i \text{ has orthogonal columns}}} \Big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{D}}_{l,i} \times_{j\in[m]\setminus i} \widehat{\mathbf{U}}_{l,j} \times_i \mathbf{W}_i \Big\rangle \le C_{3,m}\sqrt{\nu_n n}\cdot K^{3m/2}(\log n)^{(m+8)/2}.$$

Continuing from (A.41), we get with probability at least $1-n^{-2}$ that

$$\max_{1\le l\le l_{\max}} \sup_{\|\boldsymbol{\mathcal{Y}}\|_{\mathrm{F}}\le 1} \big\langle \boldsymbol{\mathcal{A}}_1 - \mathbb{E}\boldsymbol{\mathcal{A}}_1, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{Y}})\big\rangle \le C_{4,m}\sqrt{\nu_n n}\cdot K^{(3m-1)/2}(\log n)^{(m+8)/2}$$

implying that

$$\widetilde{\mathsf{Err}_{2\mathbf{r}}} \le C_{4,m}\sqrt{\nu_n n}\cdot K^{(3m-1)/2}(\log n)^{(m+8)/2} + (\alpha m n_o n^{m-1})^{1/2}.$$

Together with the fact $\underline{\lambda}\ge c'_{0,m}\nu_n(n/K)^{m/2}$, the initialization condition of Theorem 4.1 becomes

$$\nu_n(n/K)^{m/2} \ge C_{5,m}\sqrt{\nu_n n}K^{3m/2}(\log n)^{(m+8)/2} + C_{6,m}(Km\alpha n_o n^{m-1} + K\gamma\alpha n^m)^{1/2},$$

which also implies $\nu_n n^{m-1}\ge C_{2,m}\log n$. The rest of proof is identical to that of Theorem 4.1.

A.9 Proof of Theorem 6.1

We use induction to prove this theorem.

Step 0: Base case. From the initialization, we have $\|\widehat{\boldsymbol{\mathcal{T}}}_0 - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\delta\bar{r}^{-1/2} \cdot \underline{\lambda}$.

Step 1: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. We prove this case assuming

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}\delta\bar{r}^{-1/2} \cdot \underline{\lambda}. \tag{A.42}$$

We point out that this also implies $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq c_{1,m}b_l b_u^{-1}\bar{r}^{-1/2} \cdot \underline{\lambda}$ since $\delta \lesssim b_l^2 b_u^{-2}$. In order to use Lemma B.2, we need to derive an upper bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}$.

Step 1.1: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}$. For arbitrary $1 \geq \delta > 0$, we have,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 \leq (1+\delta/2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 + (1+2/\delta)\beta^2\|\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}^*\|_{\mathrm{F}}^2 \tag{A.43}$$

Now we consider the bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2$.

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &= \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle + \beta^2\|\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 \\ &\leq (1+\beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - 2\beta\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle \end{aligned} \tag{A.44}$$

where the last inequality holds from the Assumption 2 since $\widehat{\boldsymbol{\mathcal{T}}}_l \in \mathbb{B}_2^*$ from (A.42). Also,

$$\begin{aligned}\langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\rangle &= \langle\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*, \boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*\rangle - \langle\mathcal{P}_{\mathbb{T}_l}^{\perp}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*), \boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*\rangle \\ &\geq b_l\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 - \frac{C_{1,m}b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3 \end{aligned} \tag{A.45}$$

where the last inequality is from Assumption 2, Lemma B.1 and Cauchy-Schwartz inequality and $C_{1,m} = 2^m - 1$. Together with (A.44) and (A.45), and since we have $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{0.1b_l}{2b_u C_{1,m}} \cdot \underline{\lambda}$, we get,

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}(\boldsymbol{\mathcal{G}}_l - \boldsymbol{\mathcal{G}}^*)\|_{\mathrm{F}}^2 &\leq (1 - 2\beta b_l + \beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + \frac{2\beta C_{1,m}b_u}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^3 \\ &\leq (1 - 1.9\beta b_l + \beta^2 b_u^2)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2. \end{aligned} \tag{A.46}$$

Since we have $0.75 b_l b_u^{-1} \geq \delta^{1/2}$, if we choose $\beta \in [0.4 b_l b_u^{-2}, 1.5 b_l b_u^{-2}]$, we have $1 - 1.9\beta b_l + \beta^2 b_u^2 \leq 1 - \delta$.

So from (A.43) and (A.46), we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2 \leq (1+\frac{\delta}{2})(1-\delta)\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 + (1+\frac{2}{\delta})\mathsf{Err}_{2\mathbf{r}}^2 \tag{A.47}$$

where in the inequality we use the definition of $\mathsf{Err}_{2\mathbf{r}}$ and that $\beta \leq 1$. Now from the upper bound for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and the signal-to-noise ratio, we verified that $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}} \leq \underline{\lambda}/8$ and thus $\sigma_{\max}(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* - \beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l) \leq \underline{\lambda}/8$.

Step 1.2: Estimating $\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$. Now that we verified the condition of Lemma B.2, from the Algorithm 3, we have,

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \le \|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*-\beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2+C_m\frac{\sqrt{\bar{r}}}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*-\beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^3 \tag{A.48}$$

where $C_m>0$ is the constant depending only on m as in Lemma B.2. From (A.47) and the assumption that $\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\lesssim_m \frac{\delta}{\sqrt{\bar{r}}}\cdot\underline{\lambda}$ and $\mathsf{Err}_{2\mathbf{r}}\lesssim_m \frac{\delta^2}{\sqrt{\bar{r}}}\cdot\underline{\lambda}$, we get

$$C_m\frac{\sqrt{\bar{r}}}{\underline{\lambda}}\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*-\beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}\le\frac{\delta}{4} \tag{A.49}$$

From (A.48), (A.47) and (A.49), we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2\le(1+\frac{\delta}{4})\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*-\beta\mathcal{P}_{\mathbb{T}_l}\boldsymbol{\mathcal{G}}_l\|_{\mathrm{F}}^2\le(1-\delta^2)\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2+\frac{4}{\delta}\mathsf{Err}_{2\mathbf{r}}^2 \tag{A.50}$$

Together with the assumption $\|\widehat{\boldsymbol{\mathcal{T}}}_{l}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\lesssim_m \frac{\delta}{\sqrt{\bar{r}}}\cdot\underline{\lambda}$ and $\mathsf{Err}_{2\mathbf{r}}\lesssim_m \frac{\delta^2}{\sqrt{\bar{r}}}\cdot\underline{\lambda}$, we get

$$\|\widehat{\boldsymbol{\mathcal{T}}}_{l+1}-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le c_{1,m}\frac{\delta}{\sqrt{\bar{r}}}\cdot\underline{\lambda}, \tag{A.51}$$

which completes the induction and completes the proof.

B Technical Lemmas

Lemma B.1. *Suppose $\mathbb{T}_l$ is the tangent space at the point $\widehat{\boldsymbol{\mathcal{T}}}_l$, then we have*

$$\|\mathcal{P}_{\mathbb{T}_l}^{\perp}\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le\frac{2^m-1}{\underline{\lambda}}\|\boldsymbol{\mathcal{T}}^*-\widehat{\boldsymbol{\mathcal{T}}}_l\|_{\mathrm{F}}^2.$$

Proof. See (Cai et al. (2020c), Lemma 5.2). □

Lemma B.2. *Let $\boldsymbol{\mathcal{T}}^*=\boldsymbol{\mathcal{S}}^*\cdot(\mathbf{V}_1^*,\ldots,\mathbf{V}_m^*)$ be the tensor with Tucker rank $\mathbf{r}=(r_1,\ldots,r_m)$. Let $\boldsymbol{\mathcal{D}}\in\mathbb{R}^{d_1\times\ldots\times d_m}$ be a perturbation tensor such that $\underline{\lambda}\ge 8\sigma_{\max}(\boldsymbol{\mathcal{D}})$, where $\sigma_{\max}(\boldsymbol{\mathcal{D}})=\max_{i=1}^m\|\mathcal{M}_i(\boldsymbol{\mathcal{D}})\|$. Then we have*

$$\|\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\boldsymbol{\mathcal{T}}^*+\boldsymbol{\mathcal{D}})-\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\le\|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}}+C_m\frac{\sqrt{\bar{r}}\|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}}^2}{\underline{\lambda}}$$

where $C_m>0$ is an absolute constant depending only on m.

Proof. Without loss of generality, we only prove the Lemma in the case $m=3$. First notice that

$$\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\boldsymbol{\mathcal{T}}^*+\boldsymbol{\mathcal{D}})=(\boldsymbol{\mathcal{T}}^*+\boldsymbol{\mathcal{D}})\cdot[\![\mathcal{P}_{\mathbf{U}_1},\mathcal{P}_{\mathbf{U}_2},\mathcal{P}_{\mathbf{U}_3}]\!],$$

where $\mathbf{U}_i$ are leading r_i left singular vectors of $\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*+\boldsymbol{\mathcal{D}})$ and $\mathcal{P}_{\mathbf{U}_i}=\mathbf{U}_i\mathbf{U}_i^\top$.

First from (Xia (2019), Theorem 1), we have for all $i \in [m]$

$$\mathcal{P}_{\mathbf{U}_i} - \mathcal{P}_{\mathbf{V}_i^*} = \mathcal{S}_{i,1} + \sum_{j\geq 2} \mathcal{S}_{i,j},$$

where $\mathcal{S}_{i,j} = \mathcal{S}_{\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*),j}(\mathcal{M}_i(\boldsymbol{\mathcal{D}}))$ and specially $\mathcal{S}_{i,1} = (\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*)^\top)^\dagger(\mathcal{M}_i(\boldsymbol{\mathcal{D}}))^\top \mathcal{P}^\perp_{\mathbf{V}_i^*} + \mathcal{P}^\perp_{\mathbf{V}_i^*}\mathcal{M}_i(\boldsymbol{\mathcal{D}})(\mathcal{M}_i(\boldsymbol{\mathcal{T}}^*))^\dagger$. The explicit form of $S_{i,j}$ can be found in (Xia, 2019, Theorem 1). Here, we denote $\mathbf{A}^\dagger$ the pseudo-inverse of $\mathbf{A}$, i.e., $\mathbf{A}^\dagger = \mathbf{R}\boldsymbol{\Sigma}^{-1}\mathbf{L}^\top$ if $\mathbf{A}$ has a thin-SVD as $\mathbf{A} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{R}^\top$. With a little abuse of notations, we write $(\mathbf{A}^\dagger)^k = \mathbf{R}\boldsymbol{\Sigma}^{-k}\mathbf{L}^\top$ for any positive integer $k \geq 1$.

For the sake of brevity, we denote $\mathbf{S}_i = \sum_{j\geq 1} \mathcal{S}_{i,j}$. By the definition of $\mathcal{S}_{i,j}$, we have the bound $\|\mathcal{S}_{i,j}\| \leq \left(\frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}}\right)^j$. We get the upper bound for $\|\mathbf{S}_i\|$ as follows,

$$\|\mathbf{S}_i\| = \|\sum_{j\geq 1} \mathcal{S}_{i,j}\| \leq \frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda} - 4\sigma_{\max}(\boldsymbol{\mathcal{D}})} \leq \frac{8\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}} \tag{B.1}$$

So we have,

$$\begin{aligned}
\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!] &= \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*} + \mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*} + \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*} + \mathbf{S}_3]\!] \\
=& \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] \qquad \text{(B.2)} \\
&+ \boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&+ \boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathbf{S}_3]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&+ \boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3]\!] \qquad \text{(B.3)}
\end{aligned}$$

We now bound each of $\|\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}$, $\|\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathbf{S}_3]\!]\|_{\mathrm{F}}$ and $\|\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!]\|_{\mathrm{F}}$. Without loss of generality, we only prove the bound of the first term.

$$\mathcal{M}_1\big(\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\big) = \mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\left(\mathcal{P}^*_{\mathbf{V}_3} \otimes \mathbf{S}_2\right)^\top \tag{B.4}$$

Write

$$\begin{aligned}
\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) &= \left(\mathcal{S}_{1,1} + \sum_{j\geq 2} \mathcal{S}_{1,j}\right)\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \\
&= \mathcal{P}^\perp_{\mathbf{V}_1^*}\mathcal{M}_1(\boldsymbol{\mathcal{D}})\left(\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\right)^\dagger \mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) + \sum_{j\geq 2} \mathcal{S}_{1,j}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \\
&= \mathcal{M}_1\big(\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}^\perp_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]\big) + \sum_{j\geq 2} \mathcal{S}_{1,j}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \qquad \text{(B.5)}
\end{aligned}$$

where we used the fact $\mathcal{P}^\perp_{\mathbf{V}_1^*}\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) = \mathbf{0}$.

Thus we obtain an upper bound for $\|\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\|$ as follows

$$\|\mathbf{S}_1\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\| \leq \sigma_{\max}(\boldsymbol{\mathcal{D}}) + \underline{\lambda}\sum_{j\geq 2}\left(\frac{4\sigma_{\max}(\boldsymbol{\mathcal{D}})}{\underline{\lambda}}\right)^j \leq 4\sigma_{\max}(\boldsymbol{\mathcal{D}}), \tag{B.6}$$

where the first inequality is due to the explicit form of $\mathcal{S}_{1,j}$. See (Xia, 2019, Theorem 1).

So from (B.4) and (B.6), we get

$$\|\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}} \leq \|\mathbf{S}_1 \mathcal{M}_1(\boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}} \cdot \|\mathcal{P}^*_{\mathbf{V}_3} \otimes \mathbf{S}_2\| \leq C_1 \sqrt{\bar{r}} \frac{\sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \tag{B.7}$$

where $C_1 > 0$ is an absolute constant.

Now we consider the linear terms $\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]$, $\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]$ and $\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!]$. Clearly, we have

$$\begin{aligned}
\mathcal{M}_1\left(\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]\right) &= \mathbf{S}_1 \mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \\
\mathcal{M}_2\left(\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\right) &= \mathbf{S}_2 \mathcal{M}_2(\boldsymbol{\mathcal{T}}^*) \\
\mathcal{M}_3\left(\boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!]\right) &= \mathbf{S}_3 \mathcal{M}_3(\boldsymbol{\mathcal{T}}^*),
\end{aligned} \tag{B.8}$$

whose explicit representations are already studied in eq. (B.5). As a result, we can write

$$\begin{aligned}
&\boldsymbol{\mathcal{T}}^* \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{T}}^* \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad = \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}^{\perp}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}^{\perp}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}^{\perp}_{\mathbf{V}_3^*}]\!] \\
&\qquad + \sum_{j \geq 2} \Big(\mathcal{M}_1(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{S}_{1,j}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \mathcal{M}_2(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{S}_{2,j}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \mathcal{M}_3(\boldsymbol{\mathcal{T}}^*) \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{S}_{3,j}]\!] \Big).
\end{aligned} \tag{B.9}$$

Now we bound $\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!]$ as follows

$$\begin{aligned}
\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!] &= \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*} + \mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*} + \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*} + \mathbf{S}_3]\!] \\
&= \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathbf{S}_2, \mathbf{S}_3]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathbf{S}_3]\!] \\
&\quad + \boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3]\!]
\end{aligned} \tag{B.10}$$

Similarly as proving the bound (B.7), we can show

$$\max\left\{\|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}, \|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathcal{P}_{\mathbf{V}_3^*}]\!]\|_{\mathrm{F}}, \|\boldsymbol{\mathcal{D}} \cdot [\![\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3]\!]\|_{\mathrm{F}}\right\} \leq C_1 \sqrt{\bar{r}} \frac{\sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \tag{B.11}$$

where $C_1 > 0$ is an absolute constant.

Finally, by (B.5), (B.7), (B.9) and (B.11), we have

$$\begin{aligned}
&\left\|(\boldsymbol{\mathcal{T}}^* + \boldsymbol{\mathcal{D}}) \cdot [\![\mathcal{P}_{\mathbf{U}_1}, \mathcal{P}_{\mathbf{U}_2}, \mathcal{P}_{\mathbf{U}_3}]\!] - \boldsymbol{\mathcal{T}}^*\right\|_{\mathrm{F}} \\
&\quad \leq \left\|\boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}^{\perp}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}^{\perp}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}^{\perp}_{\mathbf{V}_3^*}]\!] + \boldsymbol{\mathcal{D}} \cdot [\![\mathcal{P}_{\mathbf{V}_1^*}, \mathcal{P}_{\mathbf{V}_2^*}, \mathcal{P}_{\mathbf{V}_3^*}]\!]\right\|_{\mathrm{F}} \\
&\qquad + C_1 \frac{\sqrt{\bar{r}} \sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}} \\
&\quad \leq \|\boldsymbol{\mathcal{D}}\|_{\mathrm{F}} + C_2 \frac{\sqrt{\bar{r}} \sigma_{\max}(\boldsymbol{\mathcal{D}})^2}{\underline{\lambda}}
\end{aligned} \tag{B.12}$$

where $C_1, C_2 > 0$ are absolute constants ($C_{2,m} = 16m + 2^{m+1}$ in the case of general m). This finishes the proof of the lemma. □

Lemma B.3. *Assume all the entries of $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{d_1 \times \ldots \times d_m}$ are independent mean-zero random variables with bounded Orlicz-ψ_2 norm:*

$$\|[\boldsymbol{\mathcal{Z}}]_\omega\|_{\psi_2} = \sup_{q \geq 1} (\mathbb{E}|[\boldsymbol{\mathcal{Z}}]_\omega|^q)^{1/q}/q^{1/2} \leq \sigma_z$$

Then there exists some constants $C_m, c_m > 0$ depending only on m such that

$$\sup_{\boldsymbol{\mathcal{M}} \in \mathbb{M}_{2\mathbf{r}}, \|\boldsymbol{\mathcal{M}}\|_F \leq 1} \langle \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{M}} \rangle \leq C_m \sigma_z \left(r^* + \sum_{i=1}^m d_i r_i \right)^{1/2}$$

with probability at least $1 - \exp(-c_m \sum_{i=1}^m d_i r_i)$, where $r^* = r_1 \ldots r_m$.

Proof. See the proof of (Han et al. (2020), Lemma D.5). □

Lemma B.4 (Maximum of sub-Gaussian)**.** *Let $Z_1, \ldots, Z_N$ be N random variables such that $\mathbb{E} \exp\{tZ_i\} \leq \exp\{t^2\sigma_z^2/2\}$ for all $i \in [N]$. Then*

$$\mathbb{P}(\max_{1 \leq i \leq N} |Z_i| > t) \leq 2N \exp(-\frac{t^2}{2\sigma_z^2}).$$

Proof. The claim follows from the following two facts:

$$\mathbb{P}(\max_{1 \leq i \leq N} Z_i > t) \leq \mathbb{P}(\cup_{1 \leq i \leq N} \{Z_i > t\}) \leq N\mathbb{P}(Z_i > t) \leq N \exp(-\frac{t^2}{2\sigma_z^2}),$$

and

$$\max_{1 \leq i \leq N} |Z_i| = \max_{1 \leq i \leq 2N} Z_i$$

with $Z_{N+i} = -Z_i$ for $i \in [N]$. □

Lemma B.5 (Spikiness implies incoherence)**.** *Let $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ satisfies Assumption 1 with parameter μ_1. Then we have:*

$$\mu(\boldsymbol{\mathcal{T}}^*) \leq \mu_1 \kappa_0.$$

where $\mu(\boldsymbol{\mathcal{T}}^*)$ is the incoherence parameter of $\boldsymbol{\mathcal{T}}^*$ and κ_0 is the condition number of $\boldsymbol{\mathcal{T}}^*$.

Proof. Denote $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot [\![\mathbf{U}_1, \ldots, \mathbf{U}_m]\!]$. Now we check the incoherence condition of $\boldsymbol{\mathcal{T}}^*$. For all $i \in [d_j]$ and $j \in [m]$,

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\|_{\ell_2} = \|\mathbf{e}_i \mathbf{U}_j \mathcal{M}_j(\boldsymbol{\mathcal{C}}^*)\|_{\ell_2} \geq \|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \cdot \underline{\lambda} \geq \|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{r_j}\kappa_0}.$$

On the other hand, we have

$$\|\mathbf{e}_i^\top \mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)\|_{\ell_2} \le \sqrt{d_j^-}\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \mu_1\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}\frac{1}{\sqrt{d_j}},$$

where the last inequality is due to the spikiness condition $\boldsymbol{\mathcal{T}}^*$ satisfies. Together with these two inequalities, we have

$$\|\mathbf{e}_i^\top \mathbf{U}_j\|_{\ell_2} \le \sqrt{\frac{r_j}{d_j}}\mu_1\kappa_0.$$

And this finishes the proof of the lemma. □

Lemma B.6. *Let $\boldsymbol{\mathcal{T}}^* \in \mathbb{M}_{\mathbf{r}}$ satisfies Assumption 1 with parameter μ_1. Suppose that $\boldsymbol{\mathcal{W}}$ satisfies $\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \frac{\underline{\lambda}}{8}$, then we have $\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}})$ is $(2\mu_1\kappa_0)^2$-incoherent if we choose $\zeta = \frac{16}{7}\mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}}$. Also, it satisfies*

$$\|\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \frac{C_m\sqrt{\bar{r}}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}},$$

where $C_m > 0$ depends only on m.

Proof. Notice $\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) = \mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\widetilde{\boldsymbol{\mathcal{W}}})$, where $\widetilde{\boldsymbol{\mathcal{W}}}$ is the entrywise truncation of $\boldsymbol{\mathcal{W}}$ with the thresholding $\zeta/2$. To check the incoherence of $\mathscr{H}_{\mathbf{r}}^{\mathrm{HO}}(\widetilde{\boldsymbol{\mathcal{W}}})$, denote $\widetilde{\mathbf{U}}_j$ the top-r_j left singular vectors of $\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})$, and $\widetilde{\boldsymbol{\Lambda}}_j$ the $r_j \times r_j$ diagonal matrix containing the top-r_j singular values of $\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})$. Then, there exist a $\widetilde{\mathbf{V}}_j \in \mathbb{R}^{d_j^- \times r_j}$ satisfying $\widetilde{\mathbf{V}}_j^\top\widetilde{\mathbf{V}}_j = \mathbf{I}_{r_j}$ such that

$$\widetilde{\mathbf{U}}_j\widetilde{\boldsymbol{\Lambda}}_j = \mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\widetilde{\mathbf{V}}_j.$$

Now we can also bound the ℓ_∞-norm of $\boldsymbol{\mathcal{T}}^*$:

$$\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \mu_1\frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \le \mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{T}}^* - \boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}} \le \mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}/8}{\sqrt{d^*}}.$$

This together with the definition of ζ, we have:

$$\mu_1\frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \le 8/7\cdot\mu_1\frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}} = \zeta/2.$$

And thus $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \zeta/2$. Then for all $i \in [d_j]$,

$$\|\mathbf{e}_i^\top\widetilde{\mathbf{U}}_j\|_{\ell_2} = \|\mathbf{e}_i^\top\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\widetilde{\mathbf{V}}_j\widetilde{\boldsymbol{\Lambda}}_j^{-1}\|_{\ell_2} \le \frac{\|\mathbf{e}_i^\top\mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{W}}})\|_{\ell_2}}{\lambda_{r_j}(\widetilde{\boldsymbol{\Lambda}}_j)} \le \frac{\zeta/2\cdot(d_j^-)^{1/2}}{7/8\cdot\lambda_{r_j}(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*))}.$$

where the last inequality is due to $\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$ since $\|\boldsymbol{\mathcal{T}}^*\|_{\ell_\infty} \le \zeta/2$ and $\|\widetilde{\boldsymbol{\mathcal{W}}}\|_{\ell_\infty} \le \zeta/2$. Meanwhile,

$$\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \sqrt{r_j}\kappa_0\lambda_{r_j}(\mathcal{M}_j(\boldsymbol{\mathcal{T}}^*)).$$

There for the $\zeta = \frac{16}{7}\mu_1 \frac{\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}}}{\sqrt{d^*}}$, we have for all $j \in [m]$

$$\max_{i\in[d_j]} \|\mathbf{e}_i \widetilde{\mathbf{U}}_j\|_{\ell_2} \le \frac{64}{49}\mu_1\kappa_0 \frac{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \underline{\lambda}/8}{\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}\sqrt{\frac{r_j}{d_j}} \le 2\mu_1\kappa_0\sqrt{\frac{r_j}{d_j}}.$$

where the second last inequality is from $\|\boldsymbol{\mathcal{W}}\|_{\mathrm{F}} \le \|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}$ and the last inequality is from $\|\boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \ge \underline{\lambda}$.

The second claim follows from the fact that $\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \underline{\lambda}/8$, and from Lemma B.2,

$$\begin{aligned}\|\mathsf{Trim}_{\zeta,\mathbf{r}}(\boldsymbol{\mathcal{W}}) - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} = \|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} &\le \|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + C_m\frac{\sqrt{\bar{r}}\|\widetilde{\boldsymbol{\mathcal{W}}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}} \\ &\le \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} + C_m\frac{\sqrt{\bar{r}}\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2}{\underline{\lambda}}\end{aligned}$$

This finishes the proof of the lemma. □

We introduce some notations for the following lemmas. Denote by $\widehat{\boldsymbol{\mathcal{T}}}_l = \boldsymbol{\mathcal{C}}_l \cdot (\mathbf{U}_1, \ldots, \mathbf{U}_m)$, $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{C}}^* \cdot (\mathbf{U}_1^*, \ldots, \mathbf{U}_m^*)$.

$$\mathbf{R}_i = \arg\min_{\mathbf{R}\in\mathbb{O}_{r_i}} \|\mathbf{U}_i - \mathbf{U}_i^*\mathbf{R}\|_{\mathrm{F}}, i \in [m] \tag{B.13}$$

If we let $\mathbf{U}_i^{*T}\mathbf{U}_i = \mathbf{L}_i\mathbf{S}_i\mathbf{W}_i^\top$ be the SVD of $\mathbf{U}_i^{*T}\mathbf{U}_i$, then the closed form of $\mathbf{R}_i$ is given by $\mathbf{R}_i = \mathbf{L}_i\mathbf{W}_i^\top$. And we rewrite

$$\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \cdots, \mathbf{V}_m^*)$$

where $\boldsymbol{\mathcal{S}}^* = \boldsymbol{\mathcal{C}}^* \cdot (\mathbf{R}_1^\top, \cdots, \mathbf{R}_m^\top)$ and $\mathbf{V}_i^* = \mathbf{U}_i^*\mathbf{R}_i, i \in [m]$. So $\mathbf{V}_i^*$ is also μ_0-incoherent.

Lemma B.7 (Entry-wise estimation of $|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|$)**.** *Suppose $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1. Under the assumptions that $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$, then we have*

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \le C_m\bar{r}^m\underline{d}^{-(m-1)}(\mu_1\kappa_0)^{4m}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

where $C_m = 2^{4m+1}(m+1)$.

Proof. First we have

$$\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* = (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m) + \sum_{i=1}^m \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m) \tag{B.14}$$

From Lemma B.5, we get $\boldsymbol{\mathcal{T}}^*$ is $\mu_1^2\kappa_0^2$-incoherent. So we have for all $\omega = (\omega_1, \ldots, \omega_m) \in [d_1] \times \ldots \times [d_m]$

$$\begin{aligned}|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega| &\le \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\prod_{i=1}^m\|(\mathbf{U}_i)_{\omega_i:}\| + \sum_{i=1}^m\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|\prod_{k=1}^{i-1}\|(\mathbf{V}_k^*)_{\omega_k:}\|\prod_{k=i+1}^m\|(\mathbf{U}_k)_{\omega_k:}\| \\ &\le \sqrt{\frac{r^*}{d^*}}(2\mu_1\kappa_0)^{2m}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} + (2\mu_1\kappa_0)^{2m-2}\sqrt{\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}\sum_{i=1}^m\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|\end{aligned}$$

where $r^* = \prod_{i=1}^m r_i, d^* = \prod_{i=1}^m d_i$ and $\bar{r} = \max_{i=1}^m r_i, \underline{d} = \min_{i=1}^m d_i$. From AG–GM inequality, we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \le (m+1)(2\mu_1\kappa_0)^{4m}\frac{r^*}{d^*}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|^2 \tag{B.15}$$

$$\le (m+1)\bar{r}^m\underline{d}^{-(m-1)}(2\mu_1\kappa_0)^{4m}\left(\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \underline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2\right)$$

$$\le 2(m+1)\bar{r}^m\underline{d}^{-(m-1)}(2\mu_1\kappa_0)^{4m}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$$

where the last inequality is from Lemma B.9, and this finishes the proof of the lemma. □

Lemma B.8 (Estimation of $\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2$)**.** *Let Ω be the α-fraction set. Suppose $\boldsymbol{\mathcal{T}}^*$ satisfies Assumption 1. Under the assumptions that $\widehat{\boldsymbol{\mathcal{T}}}_l$ is $(2\mu_1\kappa_0)^2$-incoherent and $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \le \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$, we have*

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 \le C_m(\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

where $C_m = 2^{4m+1}(m+1)$.

Proof. First from (B.15) in Lemma B.7, we have

$$|[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega|^2 \le (m+1)(2\mu_1\kappa_0)^{4m}\frac{r^*}{d^*}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\frac{\bar{r}^{m-1}}{\underline{d}^{m-1}}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|(\mathbf{U}_i - \mathbf{V}_i^*)_{\omega_i:}\|^2.$$

Since Ω is an α-fraction set, we have

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 = \sum_{\omega\in\Omega}[\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*]_\omega^2$$

$$\le (m+1)(2\mu_1\kappa_0)^{4m}\alpha r^*\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\alpha\bar{r}^{m-1}\|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

$$\le (m+1)(2\mu_1\kappa_0)^{4m}\alpha r^*\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + (m+1)(2\mu_1\kappa_0)^{4m-4}\alpha\bar{r}^m\underline{\lambda}^{-2}\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

$$\le (m+1)(2\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\left(\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \underline{\lambda}^2\sum_{i=1}^m\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2\right) \tag{B.16}$$

Now we invoke Lemma B.9, and we get

$$\|\mathcal{P}_\Omega(\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*)\|_{\mathrm{F}}^2 \le 2(m+1)(2\mu_1\kappa_0)^{4m}\bar{r}^m\alpha\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2,$$

which finishes the proof of the lemma. □

Lemma B.9 (Estimation of $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$)**.** *Let $\widehat{\boldsymbol{\mathcal{T}}}_l = \boldsymbol{\mathcal{C}}_l \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m)$ be the l-th step value in Algorithm 2 and let $\boldsymbol{\mathcal{T}}^* = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \cdots, \mathbf{V}_m^*)$. Suppose $\widehat{\boldsymbol{\mathcal{T}}}_l$ satisfies $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$. Then we have the following estimation for $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2$:*

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \geq 0.5\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + 0.5\underline{\lambda}^2 \sum_{i=1}^m \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2.$$

Proof. First we have

$$\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^* = (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m) + \sum_{i=1}^m \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m) \tag{B.17}$$

Notice that we have

$$\|\boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m)\|_{\mathrm{F}}^2 = \|(\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 \tag{B.18}$$

Denote $\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{S}}^* \cdot (\mathbf{V}_1^*, \ldots, \mathbf{V}_{i-1}^*, \mathbf{U}_i - \mathbf{V}_i^*, \mathbf{U}_{i+1}, \ldots, \mathbf{U}_m)$, then we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 &= \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \sum_{i=1}^m \|(\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\|_{\mathrm{F}}^2 + 2\sum_{i<j}\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle + 2\sum_{i=1}^m \langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle \\
&\geq \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \sum_{i=1}^m \underline{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 + 2\sum_{i<j}\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j\rangle + 2\sum_{i=1}^m \langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle
\end{aligned} \tag{B.19}$$

Notice that $\mathcal{M}_i(\boldsymbol{\mathcal{X}}_i) = (\mathbf{U}_i - \mathbf{V}_i^*)\mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\big(\mathbf{U}_m \otimes \mathbf{U}_{i+1} \otimes \mathbf{V}_{i-1} \otimes \mathbf{V}_1\big)^\top$. So we have the estimation of $|\langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle|$ is as follows:

$$\begin{aligned}
|\langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle| &= |\langle \mathcal{M}_i\big((\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m)\big), \mathcal{M}_i(\boldsymbol{\mathcal{X}}_i)\rangle| \\
&\leq \|(\mathbf{U}_i - \mathbf{V}_i^*)^\top \mathbf{U}_i\| \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \\
&\leq \sqrt{\bar{r}}\bar{\lambda} \|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}} \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}
\end{aligned} \tag{B.20}$$

Now we estimate $\|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}}$ by plugging in the closed form of $\mathbf{V}_i^*$ as in (B.13)

$$\|\mathbf{U}_i^\top (\mathbf{U}_i - \mathbf{V}_i^*)\|_{\mathrm{F}} = \|\mathbf{I} - \mathbf{S}_i\|_{\mathrm{F}} \leq \|\mathbf{I} - \mathbf{S}_i^2\|_{\mathrm{F}} = \|\mathbf{U}_{i\perp}^{*T}\mathbf{U}_i\|_{\mathrm{F}}^2 \leq \|\mathbf{U}_i - \mathbf{U}_i^*\mathbf{R}_i\|_{\mathrm{F}}^2 \tag{B.21}$$

From Wedin' sinΘ Theorem, we have for $i \in [m]$

$$\|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}} \leq \|\mathbf{U}_i - \mathbf{U}_i^*\|_{\mathrm{F}} \leq \frac{\sqrt{2}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\underline{\lambda} - \|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|} \leq \frac{2\sqrt{2}\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}}{\underline{\lambda}} \leq \frac{1}{4m\bar{r}^{1/2}\kappa_0} \tag{B.22}$$

where the second last inequality is from $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\| \leq \underline{\lambda}/2$ and the last inequality is from $\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}} \leq \frac{\underline{\lambda}}{16m\bar{r}^{1/2}\kappa_0}$. Then from (B.20) and (B.22), we have

$$|\langle (\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*) \cdot (\mathbf{U}_1, \cdots, \mathbf{U}_m), \boldsymbol{\mathcal{X}}_i\rangle| \leq \frac{1}{8m^2}\|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + \frac{1}{8}\underline{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 \tag{B.23}$$

The estimation of $|\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j \rangle| (i < j)$ is as follows. From (B.22), we have

$$\begin{aligned}
|\langle \boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j \rangle| &= |\langle \mathcal{M}_i(\boldsymbol{\mathcal{S}}^*)\mathbf{M}_{i,j}, (\mathbf{U}_i - \mathbf{V}_i^*)^\top \mathbf{V}_i^* \mathcal{M}_i(\boldsymbol{\mathcal{S}}^*) \rangle| \\
&\leq \bar{\lambda} \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \|\mathbf{M}_{i,j}\| \|(\mathbf{U}_i - \mathbf{V}_i^*)^\top \mathbf{V}_i^*\|_{\mathrm{F}} \\
&\leq \bar{\lambda} \|\boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}} \|(\mathbf{U}_i - \mathbf{V}_i^*)^\top \mathbf{V}_i^*\|_{\mathrm{F}} \|(\mathbf{U}_j - \mathbf{V}_j^*)^\top \mathbf{V}_j^*\|_{\mathrm{F}} \\
&\overset{(a)}{\leq} \sqrt{\bar{r}} \bar{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 \|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}}^2 \\
&\overset{(b)}{\leq} \frac{1}{16m^2} \underline{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}} \|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}} \\
&\leq \frac{1}{32m^2} \underline{\lambda}^2 \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2 + \frac{1}{32m^2} \underline{\lambda}^2 \|\mathbf{U}_j - \mathbf{V}_j^*\|_{\mathrm{F}}^2
\end{aligned} \tag{B.24}$$

where $\mathbf{M}_{i,j} = \mathbf{I} \otimes \ldots \otimes \mathbf{I} \otimes \mathbf{U}_j^\top (\mathbf{U}_j - \mathbf{V}_j^*) \otimes \mathbf{U}_{j-1}^\top \mathbf{V}_{j-1}^* \otimes \ldots \mathbf{U}_{i+1}^\top \mathbf{V}_{i+1}^* \otimes \mathbf{I} \otimes \ldots \otimes \mathbf{I}$, (a) holds because of (B.21), (b) holds because of (B.22).

As a result of (B.19), (B.23) and (B.24), we have

$$\|\widehat{\boldsymbol{\mathcal{T}}}_l - \boldsymbol{\mathcal{T}}^*\|_{\mathrm{F}}^2 \geq 0.5 \|\boldsymbol{\mathcal{C}}_l - \boldsymbol{\mathcal{S}}^*\|_{\mathrm{F}}^2 + 0.5 \underline{\lambda}^2 \sum_{i=1}^{m} \|\mathbf{U}_i - \mathbf{V}_i^*\|_{\mathrm{F}}^2$$

which finishes the proof of the lemma. □