Non-Convex Tensor Recovery from Tube-Wise Sensing

Tongle Wu
The Pennsylvania State University
tfw5381@psu.edu

Ying Sun[*]
The Pennsylvania State University
ybs5190@psu.edu

Abstract

In this paper, we propose a novel tube-wise local tensor compressed sensing (CS) model under the tensor product framework, where sensing operators are independently applied to each tube of a third-order tensor. To recover the low-rank ground truth tensor, we minimize a non-convex objective via Burer–Monteiro factorization and solve it using gradient descent (GD) with spectral initialization. We prove that this approach achieves exact recovery with a linear convergence rate. Notably, our method attains provably lower sample complexity than existing TCS methods if the low tubal rank ground truth tensor satisfies the defined incoherence condition. Our proof leverages the leave-one-out technique to show that gradient descent generates iterates implicitly biased towards solutions with bounded incoherence, which ensures contraction of optimization error in consecutive iterates. Empirical results validate the effectiveness of GD in solving the proposed local TCS model.

1 Introduction

Tensor compressed sensing (TCS) that recovers a low-rank tensor from limited measurements is a fundamental problem in signal processing and machine learning. Specifically, the classical TCS problem aims to recover a ground truth $\boldsymbol{\mathcal{X}}^\star \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ tensor from a few linear measurements generated as

$$\boldsymbol{y} := [y_1, \cdots, y_m]^\top = \boldsymbol{\mathcal{A}}(\boldsymbol{\mathcal{X}}^\star) := [\langle \boldsymbol{\mathcal{A}}_1, \boldsymbol{\mathcal{X}}^\star \rangle, \cdots, \langle \boldsymbol{\mathcal{A}}_m, \boldsymbol{\mathcal{X}}^\star \rangle]^\top, \tag{1}$$

with $\boldsymbol{\mathcal{A}} : \mathbb{R}^{n_1 \times n_2 \times n_3} \rightarrow \mathbb{R}^m$ being a linear sensing map. Without any structure assumption on $\boldsymbol{\mathcal{X}}^\star$, the problem is ill-posed and it is impossible to recover it from measurements less than its dimension. In this paper, we consider $\boldsymbol{\mathcal{X}}^\star$ exhibiting a structure that can be well-approximated by a low-rank model. However, unlike a matrix that has a unique notion of rank, a tensor has multiple definitions of rank [19] induced by different tensor decompositions such as Canonical Polyadic (CP) [16], Tucker [54], tensor singular value decomposition (t-SVD) [17], and others [39, 65, 70]. Algorithms and their theoretical guarantees to recover $\boldsymbol{\mathcal{X}}^\star$ under different notions of low-rankness have been studied in a series of existing works [3, 11, 18, 26, 28, 30, 36, 43, 44].

Despite the long-standing literature, works mentioned above share a common limitation: the TCS model (1) is a global CS model where the sensing operator $\boldsymbol{\mathcal{A}}$ is applied to the entire tensor $\boldsymbol{\mathcal{X}}^\star$. In certain applications, however, fulfilling such a requirement is impossible or inefficient. For example, when dealing with massive tensor datasets that exceed available memory, one common strategy is to split the data into smaller sub-tensors and allocate them across a distributed network of storage nodes [49, 55]. In this case, it is more practical to take measurements of the sub-tensors locally at each node out of implementation efficiency and privacy considerations. Another example is real-time applications such as video streaming, where the data is modeled as a tensor whose dimension is increasing with time. For these applications, algorithms that can process incoming measurements

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

incrementally and generate the tensor on the fly are favored over those that repeatedly measure and recover the full tensor whenever new data arrives [37, 64].

To address the above limitations in the global TCS model, recent work [64] firstly proposed a local TCS model where measurements are taken with respect to each lateral slice matrix of $\boldsymbol{\mathcal{X}}^\star$. Methods proposed therein provably recover the ground truth tensor $\boldsymbol{\mathcal{X}}^\star$ of low tubal rank with a lower sample complexity than the global TCS model. Inspired by this result, this work takes a further step by proposing a more flexible local TCS model that senses each tube of $\boldsymbol{\mathcal{X}}^\star$. Our proposed tube-wise TCS model only requires access to individual tubes (mode-3 fiber vectors) of the third-order tensor, whereas slice-wise [64] and global sensing methods [20, 26, 28, 60] must measure the entire slice (full matrix) and the whole tensor, respectively. We refer to our framework as the tube-wise TCS model and define it precisely below:

Definition 1 (Tube-Wise TCS)**.** *Given third-order tensor $\boldsymbol{\mathcal{X}}^\star \in \mathbb{R}^{n_1\times n_2\times n_3}$, tube-wise TCS measures $\boldsymbol{\mathcal{X}}^\star$ by the sensing map $\boldsymbol{\mathcal{Z}} : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}^{n_1\times n_2\times m}$ that consists of $n_1 \times n_2$ sensing matrices $\boldsymbol{Z}_{ij} \in \mathbb{R}^{m\times n_3},\ i = 1,\dots,n_1,\ j = 1,\dots,n_2$ and generates the measurement*

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{X}}^\star), \quad \textit{where} \quad \boldsymbol{\mathcal{Y}}(i,j,:) := \boldsymbol{Z}_{ij}\boldsymbol{\mathcal{X}}^\star(i,j,:),\ \forall (i,j) \in [n_1]\times[n_2]. \tag{2}$$

Here, n_1, n_2, and n_3 are respectively the height, width, and depth of $\boldsymbol{\mathcal{X}}^\star$; and n_1 and n_2 are collectively referred to as the spatial dimension. Each $\boldsymbol{\mathcal{X}}^\star(i,j,:)$ denotes (i,j)-th tube vector of $\boldsymbol{\mathcal{X}}^\star$, which is measured by $\boldsymbol{Z}_{ij}$ with an output vector $\boldsymbol{\mathcal{Y}}(i,j,:) \in \mathbb{R}^m$, where m denotes the tube-wise sample size. We can observe that given samples and measurements obtained via the tube-wise TCS model, we can turn it into an equivalent slice-wise or global TCS model, but not the other way around. This reflects the flexibility of the proposed tube-wise TCS model.

This work considers recovering $\boldsymbol{\mathcal{X}}^\star$ with low tubal rank (cf. Definition 8) from $\boldsymbol{\mathcal{Y}}$ under the tube-wise TCS model. The low tubal rank assumption is well-motivated in applications where multi-dimensional data exhibit strong correlations along the third mode [24, 68], such as time or spectrum. Under the t-SVD framework, this assumption enables a compact representation of the tensor and has shown empirical success in video, hyperspectral, and multichannel signal recovery tasks [22, 41, 57, 58, 62]. The TCS model under the t-SVD framework has also been widely studied in existing works [15, 25, 26, 28, 63, 66, 67]. Although we consider the Fourier transform-based t-SVD framework, our results in this paper can be extended to general t-SVD framework whose transform matrix satisfies $\boldsymbol{L}^\top\boldsymbol{L} = \boldsymbol{L}\boldsymbol{L}^\top = n_3\boldsymbol{I}$, which is standard in the transformed t-SVD framework [27, 29] that includes the DFT matrix as a special case.

Given the proposed tube-wise TCS model, a central question is how to design a method that can exactly recover the ground truth tensor $\boldsymbol{\mathcal{X}}^\star$ that is both sample and computationally efficient. This paper addresses this question and our main contributions are summarized as follows:

- **Computational complexity.** We formulate the tensor recovery problem by minimizing a quadratic loss and encode the low rank structure utilizing tensor Burer-Monteiro (BM) factorization [5]. By reparameterizing the tensor to be learned as the product of two smaller factors based on low tubal rank assumption, the BM factorization enables more computationally and memory-efficient numerical solutions than existing t-SVD-based TCS methods [20, 25, 28, 60, 67]. Specifically, we prove that despite being nonconvex, with a sufficient amount of measurements under the defined tensor incoherence condition (cf. Assumption 2), applying the GD with proper spectral initialization finds the ground truth $\boldsymbol{\mathcal{X}}^\star$ at a linear rate.
- **Tube-wise sample complexity.** We provide tube-wise sample complexity to ensure the linear rate of GD and exactly recover the ground truth, given by (12). The expression decreases linearly with the spatial dimension n, implying that for fixed tubal rank and tube size, increasing spatial dimension can reduce the sample complexity significantly. Such a property is unique to our local TCS model and has not appeared in any previous results, revealing that the local TCS model is more sample-efficient than existing methods, as summarized in Table 1. Simulation result shown in Fig. 2 corroborates our findings. Furthermore, our sample complexity is essentially tight since the total sample complexity of all tubes nearly matches the information-theoretical lower bound.
- **Technical analysis.** To prove the linear rate of GD under the provided tube-wise sample complexity, our analysis hinges on two key ingredients. One is showing the existence of a benign landscape near the optimum of non-convex loss (6), and the other is that GD iterates stay in this region. To this end, we leverage the leave-one-out (LOO) technique and judiciously construct two auxiliary sequences to show that GD with spectral initialization is implicitly regularized with all iterates having bounded

Table 1: Comparison of existing works under different TCS models.

Method	**TCS Model**	**Convexity**	**Convergence Rate**	**Normalized sample complexity**
TNN [28]	Global	Convex	✗	$O(r(2n-r)n_3)$
IR-tTNN [60]	Global	Non-convex	Sub-linear	✗
Schatten-p TNN [20]	Global	Non-convex	✗	✗
FGD [26]	Global	Non-convex	Linear	$\tilde{\mathcal{O}}(r^2 n n_3)$
Alt-PGD-Min [64]	Slice-wise	Non-convex	Linear	$\tilde{\mathcal{O}}(r^2 n_3)$
Ours	Tube-wise	Non-convex	Linear	$\tilde{\mathcal{O}}\left(\max\left\{\frac{r^5 n_3}{n}, 1\right\}\right)$

[†] Here we assume for simplicity and without loss the generality that $n_1 = n_2 = n$. r denotes the low tubal rank of the ground truth tensor. Normalized sample complexity is defined as the total number of entries in the sensing map divided by the tensor size $n_1 \times n_2 \times n_3$.

incoherence. Based on this implicit regularization phenomenon, new concentration results are established showing that the empirical Hessian of non-convex loss is close to its population counterpart under reduced sample complexity than state-of-the-art results.

Related works. Based on our proposed model and technique, we organize relevant works as follows.

- **T-SVD based global TCS.** Recovering low-tubal-rank tensors under the t-SVD framework has been considered in [28], wherein a convex tensor nuclear norm (TNN) minimization formulation is proposed with order-optimal sample complexity guarantees. The results were extended to the robust recovery settings in [67]. Under the TNN minimization framework, nonconvex surrogates for tubal rank have been explored in [20] and [58] to improve performance, yet no sample complexity has been provided. In addition to low-rankness, the local smoothness structure has been incorporated in [25] for visual data to improve sample complexity. Algorithmically, all these methods require performing multiple SVDs per iteration, which is computationally intense. More recently, a nonconvex formulation based on the BM factorization for symmetric tensors was proposed in [26], wherein the factorized gradient descent (FGD) algorithm is developed to solve the problem with reduced computational costs and storage requirements. All these works focus on the global TCS model rather than the local sensing setting considered in this paper.
- **Local CS model.** Compared to the global CS model, methods for local CS have remained relatively underexplored. Recent works [21, 38, 48, 49, 55] have studied local CS for low-rank matrix recovery, yet much less work has been devoted to local CS for tensors. A lateral column-wise sketching method under the Tucker decomposition was proposed in [4], and gradient-based algorithms were introduced, but without theoretical analysis. Recent work [64] has studied slice-wise local TCS under the t-SVD framework, wherein both sample complexity and algorithmic convergence are provided. In contrast, this work considers a more fine-grained tube-wise TCS model where the sensing map is applied to each tube independently. Our model achieves improved sample complexity compared to [64], as shown in Table 1.
- **Implicit regularization in tensor learning.** Under the BM factorization, one line of research has studied overparameterized tensor models and has shown that some optimization algorithms implicitly promote low-rank structures in CP [10, 12, 45], Tucker [13, 31, 46], and t-SVD decompositions [15]. Another line of work has studied the exact parameterization setting and has proved that GD or its variants generate balanced factors [51, 52] or enforce incoherence conditions [2, 6, 7, 59]. There are also works that have investigated other types of implicit regularization in tensor learning, such as [33, 34]. Our work reveals the implicit regularization effect of gradient descent in the proposed tube-wise local TCS model, which allows the tube-wise sample complexity required for exact recovery to be reduced linearly as the spatial dimension increases, as shown in (12). This leads to a significant improvement in sample efficiency.

2 Preliminaries

Notations. Scalars, vectors, matrices, and tensors are denoted by x, $\boldsymbol{x}$, $\boldsymbol{X}$, and $\boldsymbol{\mathcal{X}}$, respectively. For third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote its (i, j, k)-th entry as $\boldsymbol{\mathcal{X}}_{ijk}$. The (i, j)-th tube, denoted as $\boldsymbol{\mathcal{X}}(i, j, :)$ or $\boldsymbol{\mathcal{X}}_{ij:}$, is considered as a tensor in $\mathbb{R}^{1 \times 1 \times n_3}$ or a vector in $\mathbb{R}^{n_3}$, depending on the context. The i-th horizontal, lateral, and frontal slice of $\boldsymbol{\mathcal{X}}$ are denoted as $\boldsymbol{\mathcal{X}}(i,:,:)$ or $\boldsymbol{\mathcal{X}}_{i::} \in \mathbb{R}^{1\times n_2\times n_3}$, $\boldsymbol{\mathcal{X}}(:,i,:) \in \mathbb{R}^{n_1\times 1\times n_3}$, and $\boldsymbol{\mathcal{X}}(:,:,i) \in \mathbb{R}^{n_1\times n_2\times 1}$ respectively. We also refer to the frontal slice $\boldsymbol{\mathcal{X}}(:,:,i)$ as $\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_1\times n_2}$. $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{ijk} \boldsymbol{\mathcal{X}}_{ijk}\boldsymbol{\mathcal{Y}}_{ijk}$ is the tensor inner product. $\overline{\boldsymbol{\mathcal{X}}}$ denotes the result of performing DFT on all the tubes of $\boldsymbol{\mathcal{X}}$, i.e., $\overline{\boldsymbol{\mathcal{X}}} := \text{fft}(\boldsymbol{\mathcal{X}}, [\,], 3)$; and the inverse DFT is $\boldsymbol{\mathcal{X}} := \text{ifft}(\overline{\boldsymbol{\mathcal{X}}}, [\,], 3)$. $\sigma_{\max}(\boldsymbol{X})$ ($\sigma_{\min}(\boldsymbol{X})$) is the largest (smallest) nonzero singular value of $\boldsymbol{X}$.

Basics of t-SVD algebra. The three basic operations, unfold, fold and block circulant, for $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ are defined as

$$\begin{aligned} &\text{Unfold}\,(\boldsymbol{\mathcal{X}}) := \left[\boldsymbol{X}^{(1)}; \cdots; \boldsymbol{X}^{(n_3)}\right]; \\ &\text{Fold}\,(\text{Unfold}\,(\boldsymbol{\mathcal{X}})) := \boldsymbol{\mathcal{X}}; \end{aligned} \qquad \text{bcirc}(\boldsymbol{\mathcal{X}}) := \begin{bmatrix} \boldsymbol{X}^{(1)} & \boldsymbol{X}^{(n_3)} & \cdots & \boldsymbol{X}^{(2)} \\ \boldsymbol{X}^{(2)} & \boldsymbol{X}^{(1)} & \cdots & \boldsymbol{X}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{X}^{(n_3)} & \boldsymbol{X}^{(n_3-1)} & \cdots & \boldsymbol{X}^{(1)} \end{bmatrix}. \tag{3}$$

With the above operations, we introduce the following tensor algebraic basics [17, 28].

Definition 2. *The tensor product of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_2\times n_4\times n_3}$ is defined as:*

$$\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}} = \mathit{Fold}\big(\mathit{bicrc}(\boldsymbol{\mathcal{X}})\mathit{Unfold}\big(\boldsymbol{\mathcal{Y}}\big)\big) \in \mathbb{R}^{n_1\times n_4\times n_3}. \tag{4}$$

Definition 3. *The conjugate transpose of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ is defined as the tensor $\boldsymbol{\mathcal{X}}^\top \in \mathbb{R}^{n_2\times n_1\times n_3}$, obtained by taking the conjugate transpose of each frontal slice of $\boldsymbol{\mathcal{X}}$, then reversing the order of transposed frontal slices 2 through n_3.*

Definition 4. *If $\boldsymbol{\mathcal{I}}(:,:,1) = \boldsymbol{I}_n$ and $\boldsymbol{\mathcal{I}}(:,:,i) = \boldsymbol{0}_n$ for $2 \le i \le n_3$, then $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{n\times n\times n_3}$ is defined as the identity tensor.*

Definition 5. *$\boldsymbol{\mathcal{Q}}^{n\times n\times n_3}$ is orthogonal if and only if it satisfies $\boldsymbol{\mathcal{Q}}^\top * \boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{Q}} * \boldsymbol{\mathcal{Q}}^\top = \boldsymbol{\mathcal{I}}$.*

Definition 6. *A tensor is f-diagonal if each of its frontal slices is diagonal .*

Definition 7. *$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times n\times n_3}$ has an inverse $\boldsymbol{\mathcal{Y}}$ if $\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{Y}} * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{I}}$. If $\boldsymbol{\mathcal{X}}$ is invertible, we use $\boldsymbol{\mathcal{X}}^{-1}$ to denote its inverse.*

Definition 8. *The tubal and average ranks are respectively defined as $\mathit{rank}_t(\boldsymbol{\mathcal{X}}) := \max_{i\in[n_3]} \mathit{rank}\big(\overline{\boldsymbol{X}}^{(i)}\big)$, and $\mathit{rank}_a(\boldsymbol{\mathcal{X}}) := (1/n_3) \cdot \sum_{i=1}^{n_3} \mathit{rank}\big(\overline{\boldsymbol{X}}^{(i)}\big)$.*

Theorem 1 (t-SVD [17])**.** *$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ can be decomposed as $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^\top$ with $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1\times n_1\times n_3}$, $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2\times n_2\times n_3}$ being orthogonal and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ f-diagonal.*

Definition 9. *The spectral norm of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ is defined as $\|\boldsymbol{\mathcal{X}}\| := \sigma_{\max}\left(\mathit{bcirc}\left(\boldsymbol{\mathcal{X}}\right)\right)$.*

Given the sensing map $\boldsymbol{\mathcal{Z}}$ in (2), its adjoint map is defined as follows:

Definition 10. *The adjoint map of $\boldsymbol{\mathcal{Z}} : \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}^{n_1\times n_2\times m}$, denoted as $\boldsymbol{\mathcal{Z}}^c : \mathbb{R}^{n_1\times n_2\times m} \to \mathbb{R}^{n_1\times n_2\times n_3}$, is defined such that for $\forall \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1\times n_2\times m}$, it outputs $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Z}}^c(\boldsymbol{\mathcal{Y}}) \in \mathbb{R}^{n_1\times n_2\times n_3}$ with*

$$\boldsymbol{\mathcal{X}}(i,j,:) = \boldsymbol{Z}_{ij}^\top \boldsymbol{\mathcal{Y}}(i,j,:), \quad \forall\, (i,j) \in [n_1] \times [n_2]. \tag{5}$$

3 Local TCS via Burer-Monteiro factorization

We employ the Burer-Monteiro factorization [5] to parameterize the low rank tensor to be learned as $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}$, where $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1\times r\times n_3}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_2\times r\times n_3}$. Recovering $\boldsymbol{\mathcal{X}}$ under the tube-wise local TCS model (2) can then be formulated as minimizing the following non-convex loss function:

$$\mathcal{L}\left(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\right) := \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \left\|\boldsymbol{Z}_{ij}\left(\boldsymbol{\mathcal{A}}(i,:,:) * (\boldsymbol{\mathcal{B}}(j,:,:))^\top\right) - \boldsymbol{y}_{ij}\right\|^2 + \frac{m}{8}\left\|\boldsymbol{\mathcal{A}}^\top * \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}^\top * \boldsymbol{\mathcal{B}}\right\|_F^2, \tag{6}$$

where the first term fits the model to the measurements and $\boldsymbol{y}_{ij} \in \mathbb{R}^m$ is the measurement vector for (i,j)-th tube. Since the BM factorization is not unique, factors $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ cannot be identified. To address this ambiguity, we generalize a technique widely used in asymmetric matrix recovery problems [8, 53, 73] to tensors and introduce in (6) the second term balances the size of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$. If there is no such term, the loss has an identifiability issue since for any invertible $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{r\times r\times n_3}$, factors $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})$ and $(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{R}}^{-1}, \boldsymbol{\mathcal{B}} * \boldsymbol{\mathcal{R}}^\top)$ has the same loss value. Such ambiguity may result in the norms of the two factors being highly unbalanced, leading to even divergence of the algorithm.

To minimize $\mathcal{L}$, we substitute the expression of measurement $\boldsymbol{\mathcal{Y}}$ generated in Eq. (2) into the above loss function and then apply the GD method with step size η taking the form:

$$\boldsymbol{\mathcal{A}}^{t+1} = \boldsymbol{\mathcal{A}}^t - \eta \boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{A}}^t * \left(\boldsymbol{\mathcal{B}}^t\right)^\top - \boldsymbol{\mathcal{X}}^\star\right) * \boldsymbol{\mathcal{B}}^t - \frac{m\eta}{2} \boldsymbol{\mathcal{A}}^t * \left(\left(\boldsymbol{\mathcal{A}}^t\right)^\top * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^\top * \boldsymbol{\mathcal{B}}^t\right) \tag{7}$$

$$\boldsymbol{\mathcal{B}}^{t+1} = \boldsymbol{\mathcal{B}}^t - \eta \left(\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{A}}^t * \left(\boldsymbol{\mathcal{B}}^t\right)^\top - \boldsymbol{\mathcal{X}}^\star\right)\right)^\top * \boldsymbol{\mathcal{A}}^t - \frac{m\eta}{2} \boldsymbol{\mathcal{B}}^t * \left(\left(\boldsymbol{\mathcal{B}}^t\right)^\top * \boldsymbol{\mathcal{B}}^t - \left(\boldsymbol{\mathcal{A}}^t\right)^\top * \boldsymbol{\mathcal{A}}^t\right), \tag{8}$$

where the adjoint map $\boldsymbol{\mathcal{Z}}^c$ is introduced in Definition 10. The estimated recovery result at each iteration t is therefore given by $\boldsymbol{\mathcal{X}}^t := \boldsymbol{\mathcal{A}}^t * \left(\boldsymbol{\mathcal{B}}^t\right)^\top$.

To overcome the challenge of getting stuck at local solutions due to the nonconvexity of (6) and enabling exact recovery of $\boldsymbol{\mathcal{X}}^\star$, we further provide a spectral initialization scheme that starts GD at a point sufficiently close to the ground truth. The complete algorithm is given by Algorithm 1 and 2.

Algorithm 1 Gradient descent with spectral initialization for minimizing (6).

```
Input: Operator $\boldsymbol{\mathcal{Z}}$ and measurements $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1 \times n_2 \times m}$, rank $r$, step size $\eta$, total iteration $T_{\max}$.
    ▷ Spectral Initialization:
    $\boldsymbol{\mathcal{X}}^0 := \frac{\boldsymbol{\mathcal{Z}}^c(\boldsymbol{\mathcal{Y}})}{m}$
    $[\boldsymbol{\mathcal{A}}^0, \boldsymbol{\mathcal{B}}^0] = \texttt{T-SVD-Spec}(\boldsymbol{\mathcal{X}}^0, r)$
    ▷ Gradient Descent:
    for $t = 0, \dots, T_{\max} - 1$ do
        Update $\boldsymbol{\mathcal{A}}^{t+1} = \boldsymbol{\mathcal{A}}^t - \eta \nabla_{\boldsymbol{\mathcal{A}}} \mathcal{L}(\boldsymbol{\mathcal{A}}^t, \boldsymbol{\mathcal{B}}^t), \boldsymbol{\mathcal{B}}^{t+1} = \boldsymbol{\mathcal{B}}^t - \eta \nabla_{\boldsymbol{\mathcal{B}}} \mathcal{L}(\boldsymbol{\mathcal{A}}^t, \boldsymbol{\mathcal{B}}^t)$ based on (7), (8).
    end for
Output: Recover tensor $\boldsymbol{\mathcal{X}}^{T_{\max}} = \boldsymbol{\mathcal{A}}^{T_{\max}} * (\boldsymbol{\mathcal{B}}^{T_{\max}})^\top$.
```

Algorithm 2 Truncated t-SVD based spectral initialization: $[\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}]$ = $\texttt{T-SVD-Spec}(\boldsymbol{\mathcal{X}}, r)$

```
Input: $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $r$.
    $\overline{\boldsymbol{\mathcal{X}}} = \text{fft}(\boldsymbol{\mathcal{X}}, [\,], 3)$
    for $i = 1, \dots, n_3$ do
        $[\overline{\boldsymbol{U}}^{(i)}, \overline{\boldsymbol{S}}^{(i)}, \overline{\boldsymbol{V}}^{(i)}] = \text{SVD}(\overline{\boldsymbol{X}}^{(i)})$
        $\overline{\boldsymbol{A}}^{(i)} = \overline{\boldsymbol{U}}^{(i)}(:, 1:r)(\overline{\boldsymbol{S}}^{(i)}(1:r, 1:r))^{\frac{1}{2}}; \overline{\boldsymbol{B}}^{(i)} = \overline{\boldsymbol{V}}^{(i)}(:, 1:r)(\overline{\boldsymbol{S}}^{(i)}(1:r, 1:r))^{\frac{1}{2}}$
        $\boldsymbol{\mathcal{A}} = \text{ifft}(\overline{\boldsymbol{\mathcal{A}}}, [\,], 3), \boldsymbol{\mathcal{B}} = \text{ifft}(\overline{\boldsymbol{\mathcal{B}}}, [\,], 3)$
    end for
Output: Tensor factors $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$.
```

To establish the theoretical results, we first impose the following low-rank structure on $\boldsymbol{\mathcal{X}}^\star$.

Assumption 1. *Each front slice of* $\overline{\boldsymbol{\mathcal{X}}^\star}$ *has rank* $r \ll \min\{n_1, n_2, n_3\}$.

Assumption 1 is the same as the low multi-rank assumption in [69] and implies $\text{rank}_t(\boldsymbol{\mathcal{X}}^\star) = r$. Thus, Theorem 1 implies that $\boldsymbol{\mathcal{X}}^\star$ has exact skinny t-SVD [68] given by

$$\boldsymbol{\mathcal{X}}^\star = \boldsymbol{\mathcal{U}}^\star * \boldsymbol{\mathcal{S}}^\star * (\boldsymbol{\mathcal{V}}^\star)^\top, \tag{9}$$

where $\boldsymbol{\mathcal{U}}^\star \in \mathbb{R}^{n_1 \times r \times n_3}, \boldsymbol{\mathcal{S}}^\star \in \mathbb{R}^{r \times r \times n_3}$ and $\boldsymbol{\mathcal{V}}^\star \in \mathbb{R}^{n_2 \times r \times n_3}$. Based on (9), we introduce the two balanced factors $\boldsymbol{\mathcal{A}}^\star \in \mathbb{R}^{n_1 \times r \times n_3}, \boldsymbol{\mathcal{B}}^\star \in \mathbb{R}^{n_2 \times r \times n_3}$ defined as

$$\overline{\boldsymbol{A}^\star}^{(i)} := \overline{\boldsymbol{U}^\star}^{(i)} \left(\overline{\boldsymbol{S}^\star}^{(i)}\right)^{\frac{1}{2}}, \overline{\boldsymbol{B}^\star}^{(i)} := \overline{\boldsymbol{V}^\star}^{(i)} \left(\overline{\boldsymbol{S}^\star}^{(i)}\right)^{\frac{1}{2}}; \boldsymbol{\mathcal{A}}^\star := \text{ifft}(\overline{\boldsymbol{\mathcal{A}}^\star}, [\,], 3), \boldsymbol{\mathcal{B}}^\star := \text{ifft}(\overline{\boldsymbol{\mathcal{B}}^\star}, [\,], 3).$$

In addition to Assumption 1, we assume $\boldsymbol{\mathcal{X}}^\star$ satisfies the following tensor incoherence condition, which ensures information is spread out across all tubes of the tensor, and no small subset of tubes carries dominant information.

Assumption 2 (Tensor incoherence condition)**.** *Let* $\boldsymbol{\mathcal{X}}^\star$ *satisfy Assumption 1 with skinny t-SVD given by* (9)*. There exists constant* μ *that*

$$\max_{i \in [n_1]} \|\boldsymbol{\mathcal{U}}^\star(i, :, :)\| \le \sqrt{\frac{\mu r}{n_1 n_3}} \quad \textit{and} \quad \max_{j \in [n_2]} \|\boldsymbol{\mathcal{V}}^\star(j, :, :)\| \le \sqrt{\frac{\mu r}{n_2 n_3}}, \tag{10}$$

where $\|\cdot\|$ is the tensor spectral norm (cf. Definition 9). Each horizontal slice matrix in the above is actually a degenerated third-order tensor.

Remark 1. *Assumption 2 imposes uniform bounds on the tensor spectral norm of horizontal slices of tensor factors. This assumption is equivalent to the matrix weak incoherence condition [68, Eq. (31)], which is assumed in the tensor completion problem under random tubal sampling. This condition is slightly stronger than the existing tensor incoherence condition that uses the Frobenius norm [24, 28, 68], due to the fact that our measurements of $\boldsymbol{\mathcal{X}}^\star$ are taken tube-wise rather than globally.*

For simplicity and without loss of generality, we assume $n_1 = n_2 = n$ for the spatial dimension of $\boldsymbol{\mathcal{X}}^\star$. We denote $\sigma_{\max} := \sigma_{\max}(\boldsymbol{\mathcal{X}}^\star)$, $\sigma_{\min} := \sigma_{\min}(\boldsymbol{\mathcal{X}}^\star)$ and the tensor condition number $\kappa := \frac{\sigma_{\max}}{\sigma_{\min}}$. Since the factors are identifiable up to orthogonal transformations, we introduce the following alignment tensor $\boldsymbol{\mathcal{R}}^t$ to remove the ambiguity:

$$\boldsymbol{\mathcal{R}}^t := \mathop{\arg\min}_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r} \left\| \begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix} * \boldsymbol{\mathcal{R}} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\|_F, \tag{11}$$

where the set $\boldsymbol{\mathcal{O}}_r$ includes all orthogonal tensors with size $r \times r \times n_3$. Recall that m is the number of measurements taken for each tube (cf. Definition 1). With Assumption 1 and 2, we are now ready to state the main result.

Theorem 2. *Let $\boldsymbol{\mathcal{X}}^\star$ satisfy Assumption 1 and 2, and the entries of local sensing matrices $\{\boldsymbol{Z}_{ij}\}$ be generated from $i.i.d$ Gaussian distribution $\mathcal{N}(0,1)$. If the tube-wise measurements satisfy*

$$m \geq \frac{(5000C)^2(\mu r)^5\kappa^9 n_3 \log^2(n\vee n_3)}{n} \vee 1. \tag{12}$$

Then under the step size condition $\eta \leq \frac{\sigma_{\min}}{100m\sigma_{\max}^2}$, with probability at least $1 - \frac{1}{(n\vee n_3)^{10}}$ the iterates generated by Algorithm 1 satisfy

$$\left\| \begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \leq (1-0.1m\eta\sigma_{\min})^t \cdot 98C\sqrt{\frac{(\mu r)^4\kappa^5 n_3\log^2(n\vee n_3)\sigma_{\max}}{mn}}, \tag{13}$$

$$\max_{l\in[2n]} \left\| \left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\| \leq (1-0.1m\eta\sigma_{\min})^t \cdot 2964C\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{mn^2}}, \tag{14}$$

where C is a universal constant.

Theorem 2 further implies the following fine-grained uniform bound on the recovery accuracy for all tubes of $\boldsymbol{\mathcal{X}}^\star$; a new result in the t-SVD based tensor recovery literature.

Corollary 1 (Tube-wise bound)**.** *Under the same setting as Theorem 2, $\forall\epsilon > 0$, the iterates generated by GD achieve*

$$\max_{ij} \left\| \left(\boldsymbol{\mathcal{A}}^t * (\boldsymbol{\mathcal{B}}^t)^\top - \boldsymbol{\mathcal{X}}^\star\right)_{ij:}\right\|_F \leq \epsilon \tag{15}$$

in at most $\mathcal{O}(\kappa^2 \log\frac{1}{\epsilon})$ iterations with high probability.

Next, we discuss the implications of Theorem 2 and Corollary 1 in detail.

Computational complexity. The above results show that even for the non-convex loss (6), GD with spectral initialization converges to $\boldsymbol{\mathcal{X}}^\star$ at a linear rate with iteration complexity $\mathcal{O}(\kappa^2\log\frac{1}{\epsilon})$. The complexity scales quadratically with κ, showing that GD converges faster if the ground truth tensor is well-conditioned. The computation cost per iteration mainly lies in performing FFT, SVD, inverse FFT, and matrix multiplication. The complexity of initialization is $\mathcal{O}\left(mn_3n^2 + n^2n_3\log n_3 + n^3n_3\right)$ and that of per GD iteration is $\mathcal{O}\left(mn_3n^2 + n^2n_3\log n_3 + rnn_3\log n_3 + n^2r + nr^2\right)$. Thus, the total computational cost of Algorithm 1 is $\tilde{\mathcal{O}}\left(\kappa^2\left(mn_3n^2 + n^2n_3\log n_3 + n^2r\right) + n^3n_3\right)$.

Sample complexity. Based on the local TCS model (1), a naive approach to reconstruct $\boldsymbol{\mathcal{X}}^\star$ is to recover each tube independently, which requires measurements m on the order of $\mathcal{O}(n_3)$. In contrast, Eq. (12) shows that collectively recovering all the tubes exploiting the low-rankness structure can significantly reduce the sample complexity. Specifically, since the right hand side of (12) decreases linearly with n, we have that for a fixed tubal rank r and tube length n_3, a larger spatial dimension n leads to fewer samples needed per tube to achieve exact recovery. A direct consequence is that the sample complexity of our model outperforms state-of-the-art results as illustrated by Table 1. The total number of samples taken over all tubes is $\tilde{\mathcal{O}}\left(r^5 n_3 n\right)$, which is nearly order-optimal when compared to the degrees of freedom $\mathcal{O}(rnn_3)$ for tensors with tubal rank r [28].

Remark 2. *The dependence of the sample complexity on the tubal rank of $\boldsymbol{\mathcal{X}}^\star$ is r^5 due to analysis technicalities. Similar high-order polynomial dependencies on the rank can also be found in the literature of non-convex matrix and tensor recovery [8, 32, 47, 50, 59].*

The implicit regularization of GD. Eq. (14) reveals that even without explicit regularization, the iterates generated by GD remain incoherent throughout the optimization process. This implicit bias is crucial since the non-convex loss (6) is strongly convex and smooth when restricted to the incoherent directions, thereby enabling the linear convergence of GD. Eq. (14) also leads to the new tube-wise convergence rate (15), showing that the recovery error is spread out across all tubes. The main technique to prove this property is the LOO analysis, which will be discussed in detail in Sec. 4.

4 Theoretical analysis

This section outlines the analysis by providing intermediate results paving the way towards Theorem 2, whose complete proof is provided in Appendix A. The starting point is proving the sensing map $\boldsymbol{\mathcal{Z}}$ in (2) satisfies the tensor Restricted Isometry Property (T-RIP).

Proposition 1. *If the sensing map $\boldsymbol{\mathcal{Z}}$ and ground truth $\boldsymbol{\mathcal{X}}^\star$ satisfy the conditions in Theorem 2, and the tube-wise sample size has*

$$m \gtrsim \frac{r^2(n-r)n_3\kappa^2\mu^2}{n^2\delta_r^2}, \tag{16}$$

then tensor RIP holds with parameter $\delta_r \in (0,1)$ given as

$$m(1-\delta_r)\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2 \le \left\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2 \le m(1+\delta_r)\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2 \tag{17}$$

with probability at least $1-\exp\{-cr(n-r)n_3\}$.

For the global TCS model, existing proofs used a Bernstein inequality that is uniformly applicable to all rank-r tensors to prove the RIP condition. If we use the same inequality, the sample size required to establish the RIP condition given in Proposition 1 would be much larger than our current result due to ignoring the structures of the tube-wise sensing model.

Thus, compared to RIP result in Lemma 5 of [26], Proposition 1 shows that under the local TCS model, a smaller sample size m suffices to guarantee $\left\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2$ concentrates on its expected value $m\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2$. The key property that leads to this result is that $\boldsymbol{\mathcal{X}}^\star$ satisfies the tensor incoherent condition (10). Motivated by this observation, we prove next that the GD iterates also satisfy the incoherence condition, which will allow us to use similar arguments to show the empirical Hessian of the non-convex loss (6) concentrates on the population Hessian under the same order of sample complexity. The linear rate of GD then follows from the fact that the population Hessian is restricted strongly convex and smooth.

To prove that GD iterates satisfy the incoherence condition implicitly, we leverage the LOO technique and define the LOO map for the tube-wise sensing map $\boldsymbol{\mathcal{Z}}$, which is motivated by the matrix completion setting [8, 32].

Definition 11. *Given a tube-wise TCS map $\boldsymbol{\mathcal{Z}}$, the LOO map $\boldsymbol{\mathcal{Z}}_{-i,\cdot} : \mathbb{R}^{n\times n\times n_3} \to \mathbb{R}^{n\times n\times m}$ is identical to $\boldsymbol{\mathcal{Z}}$ but excludes the measurements of all tubes belonging to the i-th horizontal slice of $\boldsymbol{\mathcal{X}}^\star$. Correspondingly, the projection map $\boldsymbol{\mathcal{P}}_{i,\cdot} : \mathbb{R}^{n\times n\times n_3} \to \mathbb{R}^{n\times n\times m}$ for $\boldsymbol{\mathcal{X}}^\star$ sets its i-th horizontal slice to zero. That is,*

$$\left[\boldsymbol{\mathcal{Z}}_{-i,\cdot}\left(\boldsymbol{\mathcal{X}}\right)\right]_{kj:} := \begin{cases}\boldsymbol{Z}_{kj}\boldsymbol{\mathcal{X}}_{kj:}, & k \neq i,\\ \mathbf{0}_m, & k = i.\end{cases}; \quad \left[\boldsymbol{\mathcal{P}}_{i,\cdot}\left(\boldsymbol{\mathcal{X}}\right)\right]_{kj:} := \begin{cases}\boldsymbol{\mathcal{X}}_{kj:}, & k = i,\\ \mathbf{0}_{n_3}, & k \neq i.\end{cases} \tag{18}$$

In analogy to Definition 11, we additionally introduce the map $\boldsymbol{\mathcal{Z}}_{\cdot,-j}$ that leaves out measuring the j-th horizontal slice and projection map $\boldsymbol{\mathcal{P}}_{\cdot,j}$.

Next, we construct the auxiliary sequences as the GD iterates generated by minimizing loss functions defined based on the LOO maps. Specifically, for $1 \leq l \leq n$, we define the loss function corresponding to leaving out measuring the i-th horizontal slice as

$$\mathcal{L}^{(l)}\left(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\right) := \left\|\boldsymbol{\mathcal{Z}}_{-l,\cdot}\left(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^\top - \boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2 + \left\|\boldsymbol{\mathcal{P}}_{l,\cdot}\left(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^\top - \boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2 + \frac{m}{8}\left\|\boldsymbol{\mathcal{A}}^\top * \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}^\top * \boldsymbol{\mathcal{B}}\right\|_F^2 \tag{19}$$

with variables $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times r \times n_3}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_2 \times r \times n_3}$. The LOO sequence is given by

$$\boldsymbol{\mathcal{A}}^{t+1,l} = \boldsymbol{\mathcal{A}}^{t,l} - \nabla_{\boldsymbol{\mathcal{A}}}\mathcal{L}^{(l)}(\boldsymbol{\mathcal{A}}^{t,l}, \boldsymbol{\mathcal{B}}^{t,l}); \quad \boldsymbol{\mathcal{B}}^{t+1,l} = \boldsymbol{\mathcal{B}}^{t,l} - \nabla_{\boldsymbol{\mathcal{B}}}\mathcal{L}^{(l)}(\boldsymbol{\mathcal{A}}^{t,l}, \boldsymbol{\mathcal{B}}^{t,l}). \tag{20}$$

Similarly, for $n+1 \leq l \leq 2n$, we can also define the loss based on leaving out $(l-n)$-th lateral slice map $\boldsymbol{\mathcal{Z}}_{\cdot,-(l-n)}$ and corresponding generated GD iterates as (19) and (20), respectively.

Furthermore, we define the following two orthogonal alignment tensors

$$\boldsymbol{\mathcal{R}}^{t,l} := \underset{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r}{\arg\min} \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix} * \boldsymbol{\mathcal{R}} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\|_F ; \boldsymbol{\mathcal{T}}^{t,l} := \underset{\boldsymbol{\mathcal{T}} \in \boldsymbol{\mathcal{O}}_r}{\arg\min} \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix} * \boldsymbol{\mathcal{T}} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix} * \boldsymbol{\mathcal{R}}^t\right\|. \tag{21}$$

Theorem 3. *In the same setting as Theorem 2, the LOO sequences satisfy*

$$\max_{l \in [2n]} \left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t,l} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\| \leq (1 - 0.1 m\eta\sigma_{\min})^t \cdot 2496C\sqrt{\frac{(\mu r)^5\kappa^9 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn^2}} \tag{22}$$

$$\max_{l \in [2n]} \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix} * \boldsymbol{\mathcal{R}}^t\right\| \leq (1 - 0.1 m\eta\sigma_{\min})^t \cdot 52C\sqrt{\frac{(\mu r)^4\kappa^2 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn^2}} \tag{23}$$

with probability at least $1 - \frac{1}{(n \vee n_3)^{10}}$.

Assumption 2 and (22) together show that the LOO sequences satisfy the bounded incoherence condition, while (23) shows that the GD iterates are sufficiently close to the LOO sequences. Eq. (22) and (23) thereby ensure that the incoherence condition (14) holds along the GD iterates.

The rest of the proof for Theorem 2 is devoted to establishing that the non-convex loss (6) under sample complexity (12) is strongly convex and smooth along the incoherent directions satisfying (14) in a local region sufficiently close to $\boldsymbol{\mathcal{X}}^\star$ (cf. Lemma 7 in Appendix A).

Our last result shows that Algorithm 1 generates initial points sufficiently close to $\boldsymbol{\mathcal{X}}^\star$ and also satisfies the tensor incoherence condition. Similar to the GD stage, $2n$ auxiliary LOO variables are introduced as follows. For $1 \leq l \leq n$, we define the l-th LOO variables $\boldsymbol{\mathcal{A}}^{0,l}, \boldsymbol{\mathcal{B}}^{0,l}$ as

$$\boldsymbol{\mathcal{X}}^{0,l} := \frac{1}{m} \cdot \left(\boldsymbol{\mathcal{Z}}^c_{-l,\cdot}(\boldsymbol{\mathcal{Y}}) + \boldsymbol{\mathcal{P}}_{l,\cdot}(\boldsymbol{\mathcal{X}}^\star)\right) \quad [\boldsymbol{\mathcal{A}}^{0,l}, \boldsymbol{\mathcal{B}}^{0,l}] = \texttt{T-SVD-Spec}(\boldsymbol{\mathcal{X}}^{0,l}, r). \tag{24}$$

For $n+1 \leq l \leq 2n$, we generate $(l-n)$-th LOO pair of variables in analogous to (24) by leaving l-th lateral slice out.

Theorem 4. *In the same setting as Theorem 2, if the tube-wise measurements satisfy*

$$m \geq \frac{(100C)^2(\mu r)^4\kappa^7 n_3 \log^2(n \vee n_3)}{n} \vee 1, \tag{25}$$

then the initialization generated by Algorithm 1 and (24) *satisfy*

$$\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^0\\ \boldsymbol{\mathcal{B}}^0\end{bmatrix} * \boldsymbol{\mathcal{R}}^0 - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \leq 98C\sqrt{\frac{(\mu r)^4\kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn}} \tag{26}$$

$$\max_{l \in [2n]} \left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^0\\ \boldsymbol{\mathcal{B}}^0\end{bmatrix} * \boldsymbol{\mathcal{R}}^0 - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\| \leq 594C\sqrt{\frac{(\mu r)^5\kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn^2}} \tag{27}$$

$$\max_{l \in [2n]} \left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{0,l}\\ \boldsymbol{\mathcal{B}}^{0,l}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{0,l} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\| \leq 126C\sqrt{\frac{(\mu r)^5\kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn^2}} \tag{28}$$

$$\max_{l\in[2n]}\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{0,l}\\ \boldsymbol{\mathcal{B}}^{0,l}\end{bmatrix}*\boldsymbol{\mathcal{T}}^{0,l}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{0}\\ \boldsymbol{\mathcal{B}}^{0}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{0}\right\|\leq 52C\sqrt{\frac{(\mu r)^4\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2 m}} \tag{29}$$

with probability at least $1-\frac{1}{(n\vee n_3)^{10}}$.

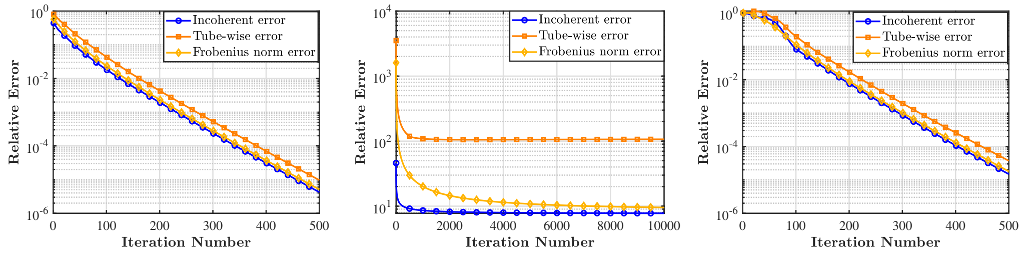
The idea of the proof for initialization is similar to that in [8, 32] and is provided in Appendix B.

5 Experimental results

We validate our theoretical findings on synthetic tensors (real-data results appear in Appendix C). First, we verify the linear convergence of Algorithm 1. Then we study its recovery performance under varying tube-wise sample complexity. Finally, we compare our method with existing works in terms of sample complexity and running time.

To validate convergence, we set $(n_1, n_2, n_3) = (100, 100, 50)$, tubal rank $r = 5$, and condition number $\kappa = 1$ for ground-truth tensor $\boldsymbol{\mathcal{X}}^\star$ that is generated from $i.i.d.\mathcal{N}(0,1)$ entries, performing an FFT along tube mode, truncating each frontal slice to rank r via SVD (setting the top r singular values to 1 and the rest to 0), and then applying the inverse FFT. We set the $\eta = 0.2$ and tube-wise sample complexity $m = 10$. Fig. 1 plots the recovery errors vs. iteration number in 20 Monte Carlo trials under different initializations. The recovery errors are plotted in three metrics: (1) the relative Frobenius norm error $\|\boldsymbol{\mathcal{X}}^t - \boldsymbol{\mathcal{X}}^\star\|_F/\|\boldsymbol{\mathcal{X}}^\star\|_F$; (1) the relative tube-wise error $\max_{ij}\left\|(\boldsymbol{\mathcal{X}}^t - \boldsymbol{\mathcal{X}}^\star)_{ij:}\right\|_F/\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F$; (3) the relative incoherence error $\max_{l\in[2n]}\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^t\\ \boldsymbol{\mathcal{B}}^t\end{bmatrix}*\boldsymbol{\mathcal{R}}^t-\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\|\Big/\max_{l\in[2n]}\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\|$.

Fig. 1(a) shows the linear convergence of our Algorithm 1 in all three relative error metrics. Fig. 1(b) and Fig. 1(c) show the results of the GD under random initialization $\boldsymbol{\mathcal{A}}^0, \boldsymbol{\mathcal{B}}^0 \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \alpha^2\boldsymbol{\mathcal{I}}_{n\times r\times n_3})$ with $\alpha = 1, 0.001$, respectively. For $\alpha = 1$, we select the largest $\eta = 0.0002$ to guarantee convergence. Fig. 1(b) shows large initialization cannot achieve exact recovery. For $\alpha = 0.001$, the $\eta = 0.2$ is the same as in the Fig. 1(a) setting. Fig. 1(c) shows that GD can achieve exact recovery under small initialization with the nearly same computational cost as Algorithm 1. Fig. 1(b) and Fig. 1(c) show that our spectral initialization is not necessary to achieve exact recovery, and GD with random initialization can also attain this in global convergence, but the small initialization is a necessary condition even under exact parameterization.



(a) (b) (c)

Figure 1: Relative recovery errors vs. iteration number. (a) Algorithm 1 with spectral initialization in Algorithm 2 . (b) Random initialization with $\alpha = 1$. (c) Random initialization with $\alpha = 0.001$.

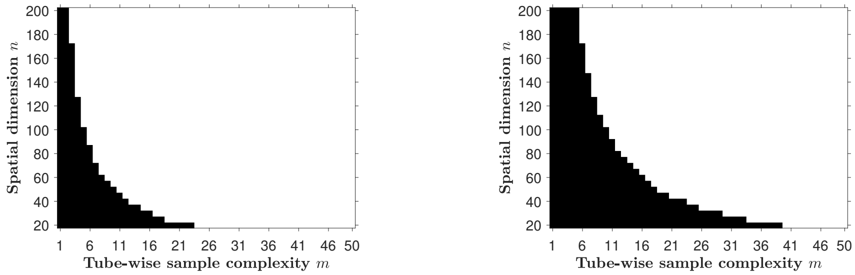
To illustrate the sample efficiency of the method, we plot phase transition diagrams to examine the recovery phenomenon with varying spatial dimension n and varying tube-wise sample complexity m. We fix the tube dimension as $n_3 = 50$ and consider the two settings: (1) $r = 5$; (2) $r = 10$. We generate ground truth $\boldsymbol{\mathcal{X}}^\star$ through the abovementioned ways. We choose n in $[20 : 5 : 200]$ and $m = 1, 2, \cdots, 50$ for both $r = 5$ and $r = 10$ cases. For each (n, m) pair, we simulate 10 test trials and assess a trial to be successful if the recovered $\hat{\boldsymbol{\mathcal{X}}}$ satisfies $\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}^\star\|_F/\|\boldsymbol{\mathcal{X}}^\star\|_F \leq 0.001$.

Fig. 2 plots the fraction of successful recovery for each (n, m) pair, and white denotes perfect recovery (100%), while black denotes failure (0%). Fig. 2(a) and Fig. 2(b) plot $r = 5$ and $r = 10$ cases, respectively. In both cases, the white successful region is large, but it is noticeably larger for

Table 2: Comparison of runtime and required measurements m (TCS). m: sample size in tensor unit; n_3: tube size; n: spatial dimension; r: tubal rank.

$n_3(r)$	n	FGD (m / time (s))	TNN (m / time (s))	Schatten-$2/3$ TNN (m / time (s))	Alt-PGD-Min (m / time (s))	Ours (m / time (s))
10 (r=2)	20	$840/7.1$	$1450/1.6$	$1020/5.1$	$43/13.8$	$3/1.1$
10 (r=2)	50	$2200/175.7$	$4600/109.4$	$3100/181.7$	$45/15.6$	$2/3.7$
20 (r=3)	20	$2400/86.9$	$3950/7.7$	$2750/20.7$	$130/13.8$	$7/4.5$
20 (r=3)	50	$6800/442.4$	$13000/717.5$	$8500/988.2$	$135/80.4$	$3/11.6$
30 (r=4)	20	$4900/181.2$	$7100/20.9$	$5250/90.0$	$250/47.3$	$12/16.1$
30 (r=4)	50	$13500/1626.1$	$24000/2395.6$	$16500/2344.9$	$268/155.9$	$6/18.9$

$r = 5$ than for $r = 10$, reflecting the greater difficulty for higher rank settings. Moreover, as the spatial dimension n increases, the required tube-wise sample size m for exact recovery decreases substantially, which corroborates our sample-efficient theoretical result.



(a) $r = 5$

(b) $r = 10$

Figure 2: Phase transitions for tube-wise local TCS model.

To validate the sample efficiency of the proposed tube-wise model and method as given in Table 1, we compare with existing t-SVD-based TCS methods, which include global TCS methods: FGD [26], TNN [28], Schatten-2/3 TNN [20], and the latest slice-wise local TCS model Alt-PGD-Min [64]. We report results for $n = 20, 50$ only, as Schatten-$2/3$ TNN and TNN encounter out-of-memory issues for larger n due to the huge sample size. For each method, we find the smallest number of random tensor samples m (each tensor has the same size as $\boldsymbol{\mathcal{X}}^\star$ and all entries generated from $i.i.d.$ standard Gaussian distribution) that can achieve exact recovery in the sense of $\frac{\|\hat{\boldsymbol{\mathcal{X}}}-\boldsymbol{\mathcal{X}}^\star\|_F}{\|\boldsymbol{\mathcal{X}}^\star\|_F} \leq 10^{-3}$. For the computational cost, we choose the largest step size for each method that can converge under each parameter setting.

The comparison results are summarized in the Table 2. We can observe that our tube-wise model and solving method can use both the smallest tensor samples and computing time to achieve exact recovery. This validates that our proposed model and method achieve both sample and computation efficiency. In addition, we can also find that for the fixed tube size n_3 and tubal rank r with increasing size of spatial dimension n, all other methods need more samples to achieve the recovery. On the contrary, our model can use fewer samples to recover, which means that increasing the spatial dimension can reduce the tube-wise sample complexity. This result also validates our theoretical sample complexity in (12).

6 Conclusion and future work

In this paper, we proposed a tube-wise TCS model under the t-SVD framework. We prove that GD with spectral initialization achieves exact recovery at a linear rate. Our analysis reveals that the tube-wise sample complexity decreases linearly with respect to the spatial dimension. But several important questions remain open. The current linear-rate constant scales quadratically with the tensor condition number, motivating preconditioned or curvature-aware gradient methods to mitigate slow convergence on ill-conditioned instances. We also assume a known tubal rank and focus on third-order tensors; extending the theory to over-parameterized factorizations and to arbitrary orders under modern high-order tensor algebra is an important next direction.

References

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

[2] Joshua Agterberg and Anru R Zhang. Estimating higher-order mixed memberships via the $\ell_{2,\infty}$ tensor perturbation bound. *Journal of the American Statistical Association*, pages 1–11, 2024.

[3] Talal Ahmed, Haroon Raja, and Waheed U Bajwa. Tensor regression using low-rank and sparse tucker decompositions. *SIAM Journal on Mathematics of Data Science*, 2(4):944–966, 2020.

[4] Silpa Babu, Selin Aviyente, and Namrata Vaswani. Tensor low rank column-wise compressive sensing for dynamic imaging. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[5] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming*, 95(2):329–357, 2003.

[6] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Advances in neural information processing systems*, 32, 2019.

[7] 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, 2020.

[8] Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841, 2020.

[9] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[10] Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34:1299–1311, 2021.

[11] Rungang Han, Rebecca Willett, and Anru R Zhang. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

[12] Kais Hariz, Hachem Kadri, Stéphane Ayache, Maher Moakher, and Thierry Artieres. Implicit regularization with polynomial growth in deep tensor factorization. In *International Conference on Machine Learning*, pages 8484–8501. PMLR, 2022.

[13] Kais Hariz, Hachem Kadri, Stéphane Ayache, Maher Moakher, and Thierry Artières. Implicit regularization in deep tucker factorization: Low-rankness via structured sparsity. In *International Conference on Artificial Intelligence and Statistics*, pages 2359–2367. PMLR, 2024.

[14] Nick Higham and Pythagoras Papadimitriou. Matrix procrustes problems. *Rapport technique, University of Manchester*, 1995.

[15] Santhosh Karnik, Anna Veselovska, Mark Iwen, and Felix Krahmer. Implicit regularization for tubal tensor factorizations via gradient descent. *arXiv preprint arXiv:2410.16247*, 2024.

[16] Henk AL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):105–122, 2000.

[17] Misha E Kilmer and Carla D Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011.

[18] Eren C Kizildag. Information-theoretic guarantees for recovering low-rank tensors from symmetric rank-one measurements. In *36th International Conference on Algorithmic Learning Theory*.

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

[20] Hao Kong, Xingyu Xie, and Zhouchen Lin. t-schatten-p norm for low-rank tensor recovery. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1405–1419, 2018.

[21] Kiryung Lee, Rakshith Srinivasa Sharma, Marius Junge, and Justin Romberg. Approximately low-rank recovery from noisy and local measurements by convex program. *Information and Inference: A Journal of the IMA*, 12(3):1612–1654, 2023.

[22] Jing Li, Quanxue Gao, Qianqian Wang, Ming Yang, and Wei Xia. Orthogonal non-negative tensor factorization based multi-view clustering. *Advances in neural information processing systems*, 36:18186–18202, 2023.

[23] Chunyan Liu, Sui Li, Dianlin Hu, Jianjun Wang, Wenjin Qin, Chen Liu, and Peng Zhang. Nonlocal tensor decomposition with joint low rankness and smoothness for spectral ct image reconstruction. *IEEE Transactions on Computational Imaging*, 10:613–627, 2024.

[24] Xiao-Yang Liu, Shuchin Aeron, Vaneet Aggarwal, and Xiaodong Wang. Low-tubal-rank tensor completion using alternating minimization. *IEEE Transactions on Information Theory*, 66(3):1714–1737, 2019.

[25] Xinling Liu, Jingyao Hou, Jiangjun Peng, Hailin Wang, Deyu Meng, and Jianjun Wang. Tensor compressive sensing fused low-rankness and local-smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8879–8887, 2023.

[26] Zhiyu Liu, Zhi Han, Yandong Tang, Xi-Le Zhao, and Yao Wang. Low-tubal-rank tensor recovery via factorized gradient descent. *IEEE Transactions on Signal Processing*, 2024.

[27] Canyi Lu. Transforms based tensor robust pca: Corrupted low-rank tensors recovery via convex optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1145–1152, 2021.

[28] Canyi Lu, Jiashi Feng, Zhouchen Lin, and Shuicheng Yan. Exact low tubal rank tensor recovery from gaussian measurements. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[29] Canyi Lu, Peng Xi, and Wei Yunchao. Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Yuetian Luo and Anru R Zhang. Low-rank tensor estimation via riemannian gauss-newton: Statistical optimality and second-order convergence. *Journal of Machine Learning Research*, 24(381):1–48, 2023.

[31] Yuetian Luo and Anru R Zhang. Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap and their interplay. *The Annals of Statistics*, 52(6):2583–2612, 2024.

[32] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[33] Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. Algorithmic regularization in tensor optimization: Towards a lifted approach in matrix sensing. *Advances in Neural Information Processing Systems*, 36:39994–40005, 2023.

[34] Ziye Ma, Igor Molybog, Javad Lavaei, and Somayeh Sojoudi. Over-parametrization via lifting for low-rank matrix sensing: Conversion of spurious solutions to strict saddle points. In *International Conference on Machine Learning*, pages 23373–23387. PMLR, 2023.

[35] Roy Mathias. Perturbation bounds for the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):588–597, 1993.

[36] 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. PMLR, 2014.

[37] Seyedehsara Nayer, Praneeth Narayanamurthy, and Namrata Vaswani. Phaseless pca: Low-rank matrix recovery from column-wise phaseless measurements. In *International Conference on Machine Learning*, pages 4762–4770. PMLR, 2019.

[38] Seyedehsara Nayer and Namrata Vaswani. Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections. *IEEE Transactions on Information Theory*, 69(2):1177–1202, 2022.

[39] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[40] Vern Paulsen. *Completely bounded maps and operator algebras*, volume 78. Cambridge University Press, 2002.

[41] Liangzu Peng, Juan Elenter, Joshua Agterberg, Alejandro Ribeiro, and Rene Vidal. TSVD: Bridging theory and practice in continual learning with pre-trained models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Wei Pu, Junjie Wu, Yulin Huang, and Jianyu Yang. Ortp: A video sar imaging algorithm based on low-tubal-rank tensor recovery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1293–1308, 2021.

[43] Zhen Qin, Michael B Wakin, and Zhihui Zhu. Guaranteed nonconvex factorization approach for tensor train recovery. *Journal of Machine Learning Research*, 25(383):1–48, 2024.

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

[45] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

[46] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, pages 18422–18462. PMLR, 2022.

[47] Mahdi Soltanolkotabi, Dominik Stöger, and Changzhi Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5140–5142. PMLR, 2023.

[48] Rakshith S Srinivasa, Seonho Kim, and Kiryung Lee. Sketching low-rank matrices with a shared column space by convex programming. *IEEE Journal on Selected Areas in Information Theory*, 4:54–60, 2023.

[49] Rakshith Sharma Srinivasa, Kiryung Lee, Marius Junge, and Justin Romberg. Decentralized sketching of low rank matrices. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Dominik Stöger and Yizhe Zhu. Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity. *arXiv preprint arXiv:2408.13276*, 2024.

[51] Tian Tong, Cong Ma, Ashley Prater-Bennette, Erin Tripp, and Yuejie Chi. Scaling and scalability: Provable nonconvex low-rank tensor completion. In *International Conference on Artificial Intelligence and Statistics*, pages 2607–2617. PMLR, 2022.

[52] Tian Tong, Cong Ma, Ashley Prater-Bennette, Erin Tripp, and Yuejie Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *Journal of Machine Learning Research*, 23(163):1–77, 2022.

[53] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning*, pages 964–973. PMLR, 2016.

[54] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[55] Namrata Vaswani. Efficient federated low rank matrix recovery via alternating gd and minimization: A simple proof. *IEEE Transactions on Information Theory*, 2024.

[56] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[57] Andong Wang, Chao Li, Mingyuan Bai, Zhong Jin, Guoxu Zhou, and Qibin Zhao. Transformed low-rank parameterization can help robust generalization for tensor neural networks. *Advances in Neural Information Processing Systems*, 36:3032–3082, 2023.

[58] Andong Wang, Yuning Qiu, Mingyuan Bai, Zhong Jin, Guoxu Zhou, and Qibin Zhao. Generalized tensor decomposition for understanding multi-output regression under combinatorial shifts. *Advances in Neural Information Processing Systems*, 37:47559–47635, 2024.

[59] Haifeng Wang, Jinchi Chen, and Ke Wei. Implicit regularization and entrywise convergence of riemannian optimization for low tucker-rank tensor completion. *Journal of Machine Learning Research*, 24(347):1–84, 2023.

[60] Hailin Wang, Feng Zhang, Jianjun Wang, Tingwen Huang, Jianwen Huang, and Xinling Liu. Generalized nonconvex approach for low-tubal-rank tensor recovery. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3305–3319, 2021.

[61] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[62] Tongle Wu and Jicong Fan. Smooth tensor product for tensor completion. *IEEE Transactions on Image Processing*, 2024.

[63] Tongle Wu, Bin Gao, Jicong Fan, Jize Xue, and Wai Lok Woo. Low-rank tensor completion based on self-adaptive learnable transforms. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):8826–8838, 2022.

[64] Tongle Wu, Ying Sun, and Jicong Fan. Non-convex tensor recovery from local measurements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21590–21598, 2025.

[65] Zhong-Cheng Wu, Ting-Zhu Huang, Liang-Jian Deng, Hong-Xia Dou, and Deyu Meng. Tensor wheel decomposition and its tensor completion application. *Advances in Neural Information Processing Systems*, 35:27008–27020, 2022.

[66] Jize Xue, Yong-Qiang Zhao, Tongle Wu, and Jonathan Cheung-Wai Chan. Tensor convolution-like low-rank dictionary for high-dimensional image representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13257–13270, 2024.

[67] Feng Zhang, Wendong Wang, Jianwen Huang, Jianjun Wang, and Yao Wang. Rip-based performance guarantee for low-tubal-rank tensor recovery. *Journal of Computational and Applied Mathematics*, 374:112767, 2020.

[68] Zemin Zhang and Shuchin Aeron. Exact tensor completion using t-svd. *IEEE Transactions on Signal Processing*, 65(6):1511–1526, 2016.

[69] Zemin Zhang, Gregory Ely, Shuchin Aeron, Ning Hao, and Misha Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-svd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3842–3849, 2014.

[70] Qibin Zhao, Guoxu Zhou, Shengli Xie, Liqing Zhang, and Andrzej Cichocki. Tensor ring decomposition. *arXiv preprint arXiv:1606.05535*, 2016.

[71] Xi-Le Zhao, Hao Zhang, Tai-Xiang Jiang, Michael K Ng, and Xiong-Jun Zhang. Fast algorithm with theoretical guarantees for constrained low-tubal-rank tensor recovery in hyperspectral images denoising. *Neurocomputing*, 413:397–409, 2020.

[72] Meng-Meng Zheng, Zheng-Hai Huang, and Yong Wang. T-positive semidefiniteness of third-order symmetric tensors and t-semidefinite programming. *Computational Optimization and Applications*, 78:239–272, 2021.

[73] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract/Intro state the tube-wise TCS model and our guarantees; Thm. 2 (linear rate) and Eq. (12) (sample complexity) establish the claims; Alg. 1–2 specify the method; Table 1 compares sample complexity; Sec. 5 validates on synthetic data.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 6 (Conclusion) lists limits: quadratic dependence on condition number, known-rank assumption, and focus on third-order tensors; it also outlines preconditioning, rank-agnostic analysis, and higher-order extensions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assump. 1–2 are stated; main results are Thm. 2 and Cor. 1; supporting tools include Prop. 1, Def. 11, and LOO sequences; full proofs in App. A–B with intermediate lemmas.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include anonymized code and run scripts with instructions.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendices

A Proof for GD Stage

We first prove the tensor RIP result of Proposition 1 in Section A.1. Section A.2, A.3, A.4 and A.5 proves the (13), (22), (23) and (14) in Theorem 2 and Theorem 3, respectively. The final section A.6, presents the technical lemmas used throughout the main analysis. The core idea of the proof is an induction argument: assuming that equations (13), (22), (23), and (14) hold at the t-th iteration, we show they continue to hold at iteration $t+1$. Therefore, establishing that they are satisfied at initialization ($t=0$) via spectral methods completes the proof.

A.1 Proof of Proposition 1

We prove this proposition by expressing $\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\|_F^2 - m\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2$ as a sum of independent, mean-zero, sub-exponential random variables. We then apply an ϵ-net argument together with the sub-exponential Bernstein inequality to derive the desired concentration bound. Crucially, independence of the sensing matrices in our tube-wise model and the tensor incoherence condition allow us to obtain a tight bound under the efficient sample size with high probability.

Proof.

$$
\begin{aligned}
\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\|_F^2 - m\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left\|\boldsymbol{Z}_{ij}\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2 - m\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left(\boldsymbol{\mathcal{X}}^\star_{ij:}\right)^T\boldsymbol{Z}_{ij}^T\boldsymbol{Z}_{ij}\boldsymbol{\mathcal{X}}^\star_{ij:} - m\left(\boldsymbol{\mathcal{X}}^\star_{ij:}\right)^T\boldsymbol{\mathcal{X}}^\star_{ij:}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{m}\underbrace{\left(\left(\boldsymbol{\mathcal{X}}^\star_{ij:}\right)^T\boldsymbol{z}_{ijs}^T\boldsymbol{z}_{ijs}\boldsymbol{\mathcal{X}}^\star_{ij:} - \left(\boldsymbol{\mathcal{X}}^\star_{ij:}\right)^T\boldsymbol{\mathcal{X}}^\star_{ij:}\right)}_{\boldsymbol{\mathcal{T}}_{ijs}} \qquad (30)
\end{aligned}
$$

It is apparent that $\{\boldsymbol{\mathcal{T}}_{ijs}\}$ are independent sub-exponential random variables with zero mean and bounded sub-exponential norm $\|K_{ijs}\|_{\psi_1} \le C\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2$. Thus, we can use sub-exponential Bernstein inequality with ϵ-net argument to bound such deviation. Firstly, $\forall t>0$ we have

$$
\begin{aligned}
\frac{t^2}{\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\|\boldsymbol{\mathcal{T}}_{ijs}\|_{\psi_1}^2} &\ge \frac{t^2}{mC^2\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^4} \\
&\ge \frac{t^2}{mC^2\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2} \\
&= \frac{t^2}{mC^2\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2} \qquad (31)
\end{aligned}
$$

Secondly, there is

$$
\frac{t}{\max_{ijk}\|\boldsymbol{\mathcal{T}}_{ijs}\|_{\psi_1}} \ge \frac{t}{C\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2}. \qquad (32)
$$

Setting $t := mC\delta_r\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2$, then there is

$$
\min\left\{\frac{t^2}{mC^2\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}, \frac{t}{C\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2}\right\} \ge \min\left\{\frac{m\delta_r^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}{\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2}, \frac{m\delta_r\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}{\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2}\right\}
$$

$$
= \frac{m\delta_r^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}{\max_{ij}\left\|\boldsymbol{\mathcal{X}}^\star_{ij:}\right\|_F^2} \overset{(i)}{\ge} \frac{m\delta_r^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}{\max_i\|\boldsymbol{\mathcal{U}}^\star_{i::}\|\left\|\boldsymbol{\mathcal{S}}^\star\right\|^2\max_j\left\|\boldsymbol{\mathcal{V}}^\star_{j::}\right\|} \overset{(ii)}{\ge} \frac{m\delta_r^2\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2}{\frac{(\mu r)^2}{n^2}\left\|\boldsymbol{\mathcal{X}}^\star\right\|^2} \overset{(iii)}{=} \frac{m\delta_r^2\frac{\left\|\overline{\boldsymbol{\mathcal{X}}^\star}\right\|_F^2}{n_3}}{\frac{(\mu r)^2}{n^2}\left\|\boldsymbol{\mathcal{X}}^\star\right\|^2}
$$

$$\geq \frac{m\delta_r^2 \frac{n_3 r \sigma_{\min}^2(\boldsymbol{\mathcal{X}}^\star)}{n_3}}{\frac{(\mu r)^2}{n^2}\left\|\boldsymbol{\mathcal{X}}^\star\right\|^2} = \frac{m\delta_r^2 n^2}{\mu^2 r \kappa^2}, \tag{33}$$

where (i) is based on the tensor norm inequalities (93), (95), and the tensor incoherence condition in (10). Based on sub-exponential Bernstein inequality for fixed $\forall \boldsymbol{\mathcal{X}}^\star$, we have

$$\mathbb{P}\left(\left|\left\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2 - m\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2\right| \geq mC\delta_r\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2\right) \leq \exp\left\{-\frac{m\delta_r^2 n^2}{\mu^2 r\kappa^2}\right\} \tag{34}$$

Thus, for all the incoherence $\boldsymbol{\mathcal{X}}^\star$ with tubal rank r, the covering number is $\mathcal{O}\left(r(2n-r)n_3\right)$, then based on the ϵ-net argument in [56], there is uniform bound as

$$\mathbb{P}\left(\left|\left\|\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{X}}^\star\right)\right\|_F^2 - m\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2\right| \geq mC\delta_r\left\|\boldsymbol{\mathcal{X}}^\star\right\|_F^2\right) \leq \exp\left\{-\frac{m\delta_r^2 n^2}{\mu^2 r\kappa^2} + r(2n-r)n_3\right\}. \tag{35}$$

Combining the tube-wise sample complexity in (16), we obtain the (17) with high probability. □

A.2 Proof of (13)

Following the proof of [8], we construct the auxiliary sequence as follows:

$$\tilde{\boldsymbol{\mathcal{A}}}^{t+1} := \boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \eta\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\mathcal{B}}^\star - \frac{m\eta}{2}\boldsymbol{\mathcal{A}}^\star * \left(\boldsymbol{\mathcal{R}}^t\right)^T\left(\left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t\right) * \boldsymbol{\mathcal{R}}^t \tag{36}$$

$$\tilde{\boldsymbol{\mathcal{B}}}^{t+1} := \boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \eta\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\mathcal{A}}^\star - \frac{m\eta}{2}\boldsymbol{\mathcal{B}}^\star * \left(\boldsymbol{\mathcal{R}}^t\right)^T\left(\left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t - \left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t\right) * \boldsymbol{\mathcal{R}}^t, \tag{37}$$

where we define the quantity $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t := \boldsymbol{\mathcal{A}}^t * \left(\boldsymbol{\mathcal{B}}^t\right)^T - \boldsymbol{\mathcal{A}}^\star * \left(\boldsymbol{\mathcal{B}}^\star\right)^T$.

Next, we show that the expected auxiliary sequence contracts toward the global optimum in one gradient step. By controlling the deviation between the auxiliary sequence and its expectation via concentration bounds, we ensure the actual auxiliary iterates also exhibit this local contraction. Finally, we prove that the true GD iterates remain sufficiently close to the auxiliary sequence, and hence inherit its local contraction toward the ground-truth factors.

It is apparent that the above defined quantities expectations have

$$\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1}\right] = \boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \eta m\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t * \boldsymbol{\mathcal{B}}^\star - \frac{m\eta}{2}\boldsymbol{\mathcal{A}}^\star * \left(\boldsymbol{\mathcal{R}}^t\right)^T\left(\left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t\right) * \boldsymbol{\mathcal{R}}^t, \tag{38}$$

$$\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1}\right] = \boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \eta m\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)^T * \boldsymbol{\mathcal{A}}^\star - \frac{m\eta}{2}\boldsymbol{\mathcal{B}}^\star * \left(\boldsymbol{\mathcal{R}}^t\right)^T\left(\left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t - \left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t\right) * \boldsymbol{\mathcal{R}}^t. \tag{39}$$

Based on triangle inequality and decomposition, there is

$$\begin{aligned}\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \leq{}& \underbrace{\left\|\begin{bmatrix}\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1}\right]\\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1}\right]\end{bmatrix} - \begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1}\\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix}\right\|}_{\Pi_1} + \underbrace{\left\|\begin{bmatrix}\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1}\right]\\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1}\right]\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\|}_{\Pi_2}\\ &+ \underbrace{\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1} - \begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1}\\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix}\right\|}_{\Pi_3}.\end{aligned} \tag{40}$$

Next, we need to bound Π_1, Π_2, Π_3, respectively.

A.2.1 Bound for Π_1

We first give the concentration inequality for auxiliary sequence as follows.

$$\Pi_1 = \eta\left\|\begin{bmatrix}\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) - m\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\mathcal{B}}^\star\\ \left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) - m\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)^T * \boldsymbol{\mathcal{A}}^\star\end{bmatrix}\right\|$$

$$
\begin{aligned}
&\leq \eta \left(\left\| \left(\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right) - m \boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right) * \boldsymbol{\mathcal{B}}^\star \right\| + \left\| \left(\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right) - m \boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right)^T * \boldsymbol{\mathcal{A}}^\star \right\| \right) \\
&\overset{(i)}{\leq} 2\eta C m r \sqrt{\frac{n n_3}{m}} \log(n \vee n_3) \max_{ij} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right)_{ij:} \right\|_F \left(\|\boldsymbol{\mathcal{B}}^\star\| + \|\boldsymbol{\mathcal{A}}^\star\| \right) \\
&\overset{(ii)}{=} 4\eta C m r \sqrt{\frac{n n_3 \sigma_{\max}}{m}} \log(n \vee n_3) \max_{ij} \Big\| \left(\boldsymbol{\mathcal{A}}_{i::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}_{i::}^\star \right) * \left(\boldsymbol{\mathcal{B}}_{j::}^\star \right)^T + \boldsymbol{\mathcal{A}}_{i::}^\star * \left(\boldsymbol{\mathcal{B}}_{j::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}_{j::}^\star \right)^T \\
&\quad + \left(\boldsymbol{\mathcal{A}}_{i::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}_{i::}^\star \right) * \left(\boldsymbol{\mathcal{B}}_{j::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}_{j::}^\star \right)^T \Big\|_F \\
&\overset{(iii)}{\leq} 4\eta C m r \sqrt{\frac{n n_3 \sigma_{\max}}{m}} \log(n \vee n_3) \Big(\max_i \left\| \left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}^\star \right)_{i::} \right\| \max_j \left\| \boldsymbol{\mathcal{B}}_{j::}^\star \right\| \\
&\quad + \max_i \left\| \boldsymbol{\mathcal{A}}_{i::}^\star \right\| \max_j \left\| \left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}^\star \right)_{j::} \right\| + \max_i \left\| \left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}^\star \right)_{i::} \right\| \max_j \left\| \left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}^\star \right)_{j::} \right\| \Big) \\
&\overset{(iv)}{\leq} 4\eta C m r \sqrt{\frac{\sigma_{\max} n n_3}{m}} \log(n \vee n_3) \Bigg(1188 \rho^t \sqrt{\frac{(\mu r)^5 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \\
&\quad + \left(594 \rho^t \sqrt{\frac{(\mu r)^5 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}} \right)^2 \Bigg) \\
&\overset{(v)}{\leq} 4\eta C m r \sqrt{\frac{\sigma_{\max} n n_3}{m}} \log(n \vee n_3) \cdot 0.3 \sqrt{\frac{\mu r}{n^2}} \sigma_{\max} \\
&\leq 0.012 \eta m \sigma_{\min} \cdot 98 C \rho^t \sqrt{\frac{(\mu r)^3 \kappa^2 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n m}}, \qquad (41)
\end{aligned}
$$

where (i) is based on Lemma 2 and (ii) is because

$$
\begin{aligned}
\left\| (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t)_{ij:} \right\|_F &= \left\| \boldsymbol{\mathcal{A}}_{i::}^t * \boldsymbol{\mathcal{R}}^t * (\boldsymbol{\mathcal{B}}_{j::}^t * \boldsymbol{\mathcal{R}}^t)^T - \boldsymbol{\mathcal{A}}_{i::}^\star * (\boldsymbol{\mathcal{B}}_{j::}^\star)^T \right\|_F \\
&= \left\| \left(\boldsymbol{\mathcal{A}}_{i::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}_{i::}^\star \right) * (\boldsymbol{\mathcal{B}}_{j::}^\star)^T + \boldsymbol{\mathcal{A}}_{i::}^\star * (\boldsymbol{\mathcal{B}}_{j::}^t - \boldsymbol{\mathcal{B}}_{j::}^\star)^T + \left(\boldsymbol{\mathcal{A}}_{i::}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}_{i::}^\star \right) * (\boldsymbol{\mathcal{B}}_{j::}^t - \boldsymbol{\mathcal{B}}_{j::}^\star)^T \right\|. \qquad (42)
\end{aligned}
$$

(iii) is due to the tensor norm inequality (95). (iv) is based on tensor incoherence condition (10) and condition in (14). The (v) is due to the tube-wise sample complexity in (12).

A.2.2 Bound for Π_2

To bound the distance between expectation of auxiliary sequence and optimum factors, we firstly defined the following error terms

$$
\begin{aligned}
\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t &:= \boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}^\star, \\
\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t &:= \boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}^\star, \\
\boldsymbol{\Delta}^t &:= \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t \end{bmatrix}, \\
\boldsymbol{\mathcal{E}}_1^t &:= -\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t \right)^T * \boldsymbol{\mathcal{B}}^\star - \frac{1}{2} \boldsymbol{\mathcal{A}}^\star * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t \right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t + \frac{1}{2} \boldsymbol{\mathcal{A}}^\star * (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t, \qquad (43) \\
\boldsymbol{\mathcal{E}}_2^t &:= -\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t \right)^T * \boldsymbol{\mathcal{A}}^\star - \frac{1}{2} \boldsymbol{\mathcal{B}}^\star * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t \right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t + \frac{1}{2} \boldsymbol{\mathcal{B}}^\star * (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t. \qquad (44)
\end{aligned}
$$

The Π_2 can be bounded as follows.

$$
\begin{aligned}
\Pi_2 &= \left\| \begin{bmatrix} \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1} \right] \\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1} \right] \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| \\
&= \left\| \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t - \eta m \boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t * \boldsymbol{\mathcal{B}}^\star - \frac{\eta m}{2} \boldsymbol{\mathcal{A}}^\star * (\boldsymbol{\mathcal{R}})^T * \left((\boldsymbol{\mathcal{A}}^t)^T * \boldsymbol{\mathcal{A}}^t - (\boldsymbol{\mathcal{B}}^t)^T * \boldsymbol{\mathcal{B}}^t \right) * \boldsymbol{\mathcal{R}}^t \\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t - \eta m \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t \right)^T * \boldsymbol{\mathcal{A}}^\star - \frac{\eta m}{2} \boldsymbol{\mathcal{B}}^\star * (\boldsymbol{\mathcal{R}})^T * \left((\boldsymbol{\mathcal{A}}^t)^T * \boldsymbol{\mathcal{A}}^t - (\boldsymbol{\mathcal{B}}^t)^T * \boldsymbol{\mathcal{B}}^t \right) * \boldsymbol{\mathcal{R}}^t \end{bmatrix} \right\|
\end{aligned}
$$

$$
\begin{aligned}
&= \left\| \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} \left(1-\eta m \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star}\right) + \frac{\eta m}{2} \boldsymbol{\mathcal{A}}^{\star} * \left(-\left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} - \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star} - \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star}\right) + \eta m \boldsymbol{\mathcal{E}}_{1}^{t} \\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} \left(1-\eta m \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star}\right) + \frac{\eta m}{2} \boldsymbol{\mathcal{B}}^{\star} * \left(-\left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} + \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} - \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} - \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star}\right) + \eta m \boldsymbol{\mathcal{E}}_{2}^{t} \end{bmatrix} \right\| \\
&\overset{(i)}{=} \left\| \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} - \eta m \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star} - \eta m \boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} + \eta m \boldsymbol{\mathcal{E}}_{1}^{t} \\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} - \eta m \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} * \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} - \eta m \boldsymbol{\mathcal{B}}^{\star} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} + \eta m \boldsymbol{\mathcal{E}}_{2}^{t} \end{bmatrix} \right\| \\
&= \left\| \frac{1}{2} \begin{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} \\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} \end{bmatrix} * \left(\boldsymbol{\mathcal{I}}_{r} - 2\eta m \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star}\right) + \frac{1}{2} \begin{bmatrix} \left(\boldsymbol{\mathcal{I}}_{n} - 2\eta m \boldsymbol{\mathcal{A}}^{\star} \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T}\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t} \\ \left(\boldsymbol{\mathcal{I}}_{n} - 2\eta m \boldsymbol{\mathcal{B}}^{\star} \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T}\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t} \end{bmatrix} + \eta m \begin{bmatrix} \boldsymbol{\mathcal{E}}_{1}^{t} \\ \boldsymbol{\mathcal{E}}_{2}^{t} \end{bmatrix} \right\| \\
&\leq \frac{1}{2} \left\| \boldsymbol{\mathcal{I}}_{r} - 2\eta m \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} \right\| \left\| \boldsymbol{\Delta}^{t} \right\| + \frac{1}{2} \left\| \boldsymbol{\mathcal{I}}_{2n} - 2\eta m \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} & \boldsymbol{\mathcal{O}} \\ \boldsymbol{\mathcal{O}} & \boldsymbol{\mathcal{B}}^{\star} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} \end{bmatrix} \right\| \left\| \boldsymbol{\Delta}^{t} \right\| \\
&\quad + \eta m \left\| \begin{bmatrix} \boldsymbol{\mathcal{E}}_{1}^{t} \\ \boldsymbol{\mathcal{E}}_{2}^{t} \end{bmatrix} \right\| \\
&\overset{(ii)}{\leq} \left(1 - \eta m \sigma_{\min}\right) \cdot \left\| \boldsymbol{\Delta}^{t} \right\| + 4\eta m \left\| \boldsymbol{\Delta}^{t} \right\|^{2} \left\| \boldsymbol{\mathcal{A}}^{\star} \right\| \\
&\overset{(iii)}{\leq} \left(1 - \eta m \sigma_{\min}\right) \cdot \left\| \boldsymbol{\Delta}^{t} \right\| + 4m\eta \sqrt{\sigma_{\max}} \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^{4} \kappa^{5} n_{3} \log^{2}(n \vee n_{3}) \sigma_{\max}}{nm}} \left\| \boldsymbol{\Delta}^{t} \right\| \\
&= \left(1 - \eta m \sigma_{\min}\right) \cdot \left\| \boldsymbol{\Delta}^{t} \right\| + 4m\eta \sigma_{\min} \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^{4} \kappa^{7} n_{3} \log^{2}(n \vee n_{3}) \sigma_{\max}}{nm}} \left\| \boldsymbol{\Delta}^{t} \right\| \\
&\leq \left(1 - 0.9\eta m \sigma_{\min}\right) \cdot \left\| \boldsymbol{\Delta}^{t} \right\|, \qquad (45)
\end{aligned}
$$

where (i) is based on Lemma 4 that $\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} + \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star}$ is the symmetric tensor. (ii) is due to $\left\| \boldsymbol{\mathcal{E}}_{1}^{t} \right\| + \left\| \boldsymbol{\mathcal{E}}_{2}^{t} \right\| \leq 4 \left\| \boldsymbol{\Delta}^{t} \right\| \left\| \boldsymbol{\mathcal{A}}^{\star} \right\|$. (iii) is based on (13), and the last inequality is because of the lower bound of tube-wise sample complexity in (12).

A.2.3 Bounding for Π_3

To bound the distance between auxiliary sequence and true GD iterates, the proof follows the proof of claim 10 in [8]. Since we want to use Lemma 6 to bound the Π_3 as. For invoking Lemma 6, we can assign

$$
\boldsymbol{\mathcal{X}}_{0} = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix}, \quad \boldsymbol{\mathcal{X}}_{\mathbf{1}} = \begin{bmatrix} \tilde{\boldsymbol{\mathcal{A}}}^{t+1} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\mathcal{X}}_{2} = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t+1} \\ \boldsymbol{\mathcal{B}}^{t+1} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t}. \qquad (46)
$$

The alignment tensor between $\boldsymbol{\mathcal{X}}_{\mathbf{0}}$ and $\boldsymbol{\mathcal{X}}_{2}$ is actually $\boldsymbol{\mathcal{R}}_{1} = \left(\boldsymbol{\mathcal{R}}^{t}\right)^{T} * \boldsymbol{\mathcal{R}}^{t+1}$. Now we prove the alignment tensor between $\boldsymbol{\mathcal{X}}_{0}$ and $\boldsymbol{\mathcal{X}}_{1}$ is actually $\boldsymbol{\mathcal{R}}_{1} = \boldsymbol{\mathcal{I}}_{r}$. Based on tensor orthogonal Procrustes in Lemma 4, this is equivalent to proving that $\left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{A}}}^{t+1} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{B}}}^{t+1}$ is the symmetric positive semi-definite tensor. Based on definition in (38), there is

$$
\begin{aligned}
\boldsymbol{\mathcal{X}}_{0}^{T} * \boldsymbol{\mathcal{X}}_{1} = {}& \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{t} * \boldsymbol{\mathcal{R}}^{t} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\mathcal{B}}^{t} * \boldsymbol{\mathcal{R}}^{t} - \eta \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{Z}}^{c} \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{A}}^{t} * \left(\boldsymbol{\mathcal{B}}^{t}\right)^{T} - \boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T}\right) * \boldsymbol{\mathcal{B}}^{\star} \\
& - \eta \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \left(\boldsymbol{\mathcal{Z}}^{c} \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{A}}^{t} * \left(\boldsymbol{\mathcal{B}}^{t}\right)^{T} - \boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T}\right)\right)^{T} * \boldsymbol{\mathcal{A}}^{\star}. \qquad (47)
\end{aligned}
$$

Based on definition of $\boldsymbol{\mathcal{R}}^{t}$ and Lemma 4, the $\left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{t} * \boldsymbol{\mathcal{R}}^{t} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\mathcal{B}}^{t} * \boldsymbol{\mathcal{R}}^{t}$ is symmetric and semi-definite positive. Thus, $\left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{A}}}^{t+1} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{B}}}^{t+1}$ is symmetric. In addition,

$$
\begin{aligned}
&\left\| \left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{A}}}^{t+1} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \tilde{\boldsymbol{\mathcal{B}}}^{t+1} - \left(\left(\boldsymbol{\mathcal{A}}^{\star}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} + \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star}\right) \right\| \overset{(i)}{\leq} 2\sqrt{\sigma_{\max}} \left(\Pi_{1} + \Pi_{2}\right) \\
&\qquad \overset{(ii)}{\leq} 2\sqrt{\sigma_{\max}} \left(0.012\eta m \sigma_{\min} \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^{3} \kappa^{2} n_{3} \log^{2}(n \vee n_{3}) \sigma_{\max}}{nm}} + \left(1 - 0.9\eta m \sigma_{\min}\right) \cdot \left\| \boldsymbol{\Delta}^{t} \right\| \right) \\
&\qquad \overset{(iii)}{\leq} 2\sqrt{\sigma_{\max}} \left(0.012\eta m \sigma_{\min} \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^{3} \kappa^{2} n_{3} \log^{2}(n \vee n_{3}) \sigma_{\max}}{nm}} \right.
\end{aligned}
$$

$$
\begin{aligned}
&+ (1-0.9\eta m\sigma_{\min}) \cdot 98C\rho^t \sqrt{\frac{(\mu r)^4\kappa^5 n_3 \log^2(n\vee n_3)\sigma_{\max}}{nm}}\Bigg) \\
&\leq 2\sigma_{\min} \cdot (1-0.888\eta m\sigma_{\min}) \cdot 98C\rho^t \sqrt{\frac{(\mu r)^4\kappa^7 n_3 \log^2(n\vee n_3)}{nm}} \\
&\leq 0.5\sigma_{\min}, \qquad (48)
\end{aligned}
$$

where (i) is based on definition of Π_1, Π_2 in (40) and (ii) is substituting results in (41) and (45). (iii) is due to condition (13), and the last inequality is due to the step size condition and tube-wise complexity in (12). Moreover, we know

$$
\sigma_{\min}\left(\left(\boldsymbol{\mathcal{A}}^\star\right)^T * \boldsymbol{\mathcal{A}}^\star + \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \boldsymbol{\mathcal{B}}^\star\right) = \sigma_{\min}\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}^T * \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right) = 2\sigma_{\min} \qquad (49)
$$

Thus, based on Wely's inequality for tensor, which applies matrix Wely's inequality for the block diagonal matrix in the Fourier domain, there is

$$
\sigma_{\min}\left(\left(\boldsymbol{\mathcal{A}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{A}}}^{t+1} + \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\right) \geq 1.5\sigma_{\min}, \qquad (50)
$$

which means that $\left(\boldsymbol{\mathcal{A}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{A}}}^{t+1} + \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{B}}}^{t+1}$ is positive definite tensor.

To use Lemma 6, we should verify the conditions in Lemma 6. We have proved

$$
\left\|\begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \leq \frac{\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_0)}{2}. \qquad (51)
$$

The last thing is to verify that

$$
\left\|\begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1} \\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^t\right\| \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| \leq \frac{\sigma_{\min}^2(\boldsymbol{\mathcal{A}}^\star)}{4} \qquad (52)
$$

as follows. Based on updating of (7) and (8), there is

$$
\begin{aligned}
&\boldsymbol{\mathcal{A}}^{t+1} * \boldsymbol{\mathcal{R}}^t = \boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \eta \boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \frac{m\eta}{2}\left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t\right) * \left(\boldsymbol{\mathcal{R}}^t\right)^T * \left(\left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t\right) * \boldsymbol{\mathcal{R}}^t \\
&\boldsymbol{\mathcal{B}}^{t+1} * \boldsymbol{\mathcal{R}}^t = \boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \eta \left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \frac{m\eta}{2}\left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t\right) * \left(\boldsymbol{\mathcal{R}}^t\right)^T * \left(\left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t - \left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t\right) * \boldsymbol{\mathcal{R}}^t.
\end{aligned}
\qquad (53)
$$

Thus,

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| &= \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1} \\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix}\right\| \\
&= \eta \left\|\begin{bmatrix} \boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t + \frac{m}{2}\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t * \left(\boldsymbol{\mathcal{R}}^t\right)^T * \left(\left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t\right) * \boldsymbol{\mathcal{R}}^t \\ \left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t + \frac{m}{2}\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t * \left(\boldsymbol{\mathcal{R}}^t\right)^T * \left(\left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t - \left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t\right) * \boldsymbol{\mathcal{R}}^t \end{bmatrix}\right\| \\
&\leq \eta\left(\left\|\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\| + \left\|\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\|\right) + m\eta\left\|\left(\boldsymbol{\mathcal{A}}^t\right)^T * \boldsymbol{\mathcal{A}}^t - \left(\boldsymbol{\mathcal{B}}^t\right)^T * \boldsymbol{\mathcal{B}}^t\right\|\left\|\boldsymbol{\Delta}^t\right\| \\
&\leq \eta\left(\left\|\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\| + m\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\|\right. \\
&\quad + \left.\left\|\left(\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\| + m\left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\|\right) \\
&\quad + m\eta\left(2\left\|\boldsymbol{\mathcal{A}}^\star\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\| + 2\left\|\boldsymbol{\mathcal{B}}^\star\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\| + \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\|^2 + \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\|^2\right)\left\|\boldsymbol{\Delta}^t\right\| \\
&\leq \eta\left(\left\|\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\| + \left\|\left(\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\| + 2m\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right\|\left\|\boldsymbol{\Delta}^t\right\|\right). \\
&\quad + m\eta\left(4\left\|\boldsymbol{\mathcal{A}}^\star\right\|\left\|\boldsymbol{\Delta}^t\right\| + 2\left\|\boldsymbol{\Delta}^t\right\|^2\right)\left\|\boldsymbol{\Delta}^t\right\| \\
&\overset{(i)}{\leq} \eta\left(\left\|\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right) * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^t\right\| + \left\|\left(\left(\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}} - m\boldsymbol{\mathcal{I}}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^t\right)\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^t\right\|\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ 4m\eta \left(2 \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\Delta}^{t}\right\| + \left\|\boldsymbol{\Delta}^{t}\right\|^{2}\right) \left\|\boldsymbol{\Delta}^{t}\right\| \\
&\overset{(ii)}{\leq} 4Cm\eta r\sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \max_{ij} \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t}\right)_{ij:}\right\| \left\|\boldsymbol{\Delta}^{t}\right\| + 12m\eta \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\Delta}^{t}\right\|^{2} \\
&\leq 4Cm\eta r\sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \left\|\boldsymbol{\Delta}^{t}\right\| \Bigg(\max_{i} \left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t})_{i::}\right\| \max_{j} \left\|\boldsymbol{\mathcal{B}}_{j::}^{\star}\right\| + \max_{j} \left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t})_{j::}\right\| \max_{i} \left\|\boldsymbol{\mathcal{A}}_{i::}^{\star}\right\| \\
&\quad + \max_{i} \left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t})_{i::}\right\| \max_{j} \left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t})_{j::}\right\| \Bigg) + 12m\eta \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\Delta}^{t}\right\|^{2} \\
&\leq 12Cm\eta r\sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \max_{i} \left\|\boldsymbol{\mathcal{A}}_{i::}^{\star}\right\| \left\|\boldsymbol{\Delta}^{t}\right\| \max_{l} \left\|\boldsymbol{\Delta}_{l::}^{t}\right\| + 12m\eta \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\Delta}^{t}\right\|^{2} \\
&\leq \Bigg(12Cm\eta r\sqrt{\frac{nn_3 \log^2(n \vee n_3)}{m}} \cdot 594C\rho^{t} \sqrt{\frac{(\mu r)^5 \kappa^6 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \\
&\quad + 12\sqrt{\sigma_{\max}} m\eta \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \Bigg) \left\|\boldsymbol{\Delta}^{t}\right\| \\
&\leq 0.08 \frac{\sigma_{\min}}{\kappa} m\eta \left\|\boldsymbol{\Delta}^{t}\right\|,
\end{aligned} \tag{54}
$$

where (i) is because

$$
\begin{aligned}
\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t}\right\| &= \left\|\boldsymbol{\mathcal{A}}^{t} * \left(\boldsymbol{\mathcal{B}}^{t}\right)^{T} - \boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^{T}\right\| \\
&\leq \left\|\boldsymbol{\mathcal{A}}^{t} * \boldsymbol{\mathcal{R}}^{t} - \boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\mathcal{B}}^{\star}\right\| + \left\|\boldsymbol{\mathcal{B}}^{t} * \boldsymbol{\mathcal{R}}^{t} - \boldsymbol{\mathcal{B}}^{\star}\right\| \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| + \left\|\boldsymbol{\mathcal{A}}^{t} * \boldsymbol{\mathcal{R}}^{t} - \boldsymbol{\mathcal{A}}^{\star}\right\| \left\|\boldsymbol{\mathcal{B}}^{t} * \boldsymbol{\mathcal{R}}^{t} - \boldsymbol{\mathcal{B}}^{\star}\right\| \\
&\leq 2 \left\|\boldsymbol{\Delta}^{t}\right\| \left\|\boldsymbol{\mathcal{A}}^{\star}\right\| + \left\|\boldsymbol{\Delta}^{t}\right\|^{2}.
\end{aligned} \tag{55}
$$

(ii) is based on Lemma 2 and last two inequalities are due to the condition (13), (14) and tube-wise sample complexity (12).

Now the condition

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\| &\leq 0.08 \frac{\sigma_{\min}}{\kappa} m\eta \left\|\boldsymbol{\Delta}^{t}\right\| \cdot \sqrt{\sigma_{\max}} \\
&\leq \frac{\sigma_{\min}^{2}(\boldsymbol{\mathcal{A}}^{\star})}{4}
\end{aligned} \tag{56}
$$

is satisfied. The last inequality is due to the step size condition in Theorem 2, condition (13) and tube-wise sample complexity (12). Finally, based on Lemma 6, there is

$$
\begin{aligned}
\Pi_3 &= \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1} \\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1} - \begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1}\end{bmatrix}\right\| \\
&\leq 9\kappa \left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| \\
&\leq 0.72\sigma_{\min} m\eta \left\|\boldsymbol{\Delta}^{t}\right\|.
\end{aligned} \tag{57}
$$

Combining the bound in (41), (45) and (57), there is

$$
\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1} \\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\| &\leq \Pi_1 + \Pi_2 + \Pi_3 \\
&\leq 0.012\eta m\sigma_{\min} \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^3 \kappa^2 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \\
&\quad + (1 - 0.9\eta m\sigma_{\min}) \cdot \left\|\boldsymbol{\Delta}^{t}\right\| + 0.72\eta m\sigma_{\min} \left\|\boldsymbol{\Delta}^{t}\right\| \\
&\overset{(i)}{\leq} (1 - 0.168\eta m\sigma_{\min}) \cdot 98C\rho^{t} \sqrt{\frac{(\mu r)^2 \kappa^2 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}
\end{aligned}
$$

$$\overset{(ii)}{\leq} 98C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}. \tag{58}$$

The inequality (i) uses (13) and (ii) is due to $\rho = (1 - 0.1\eta m \sigma_{\min})$.

A.3 Proof of (22)

The leave-one-out sequence with respect to the l-th horizontal slice is generated by the gradient descent method for solving (19) as

$$\begin{aligned}
\boldsymbol{\mathcal{A}}^{t+1,l} &= \boldsymbol{\mathcal{A}}^{t,l} - \eta \boldsymbol{\mathcal{Z}}^c_{-l,\cdot} \boldsymbol{\mathcal{Z}}_{-l,\cdot} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) * \boldsymbol{\mathcal{B}}^{t,l} - \eta \boldsymbol{\mathcal{P}}_{l,\cdot} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) * \boldsymbol{\mathcal{B}}^{t,l} \\
&\quad - \frac{m\eta}{2} \boldsymbol{\mathcal{A}}^{t,l} * \left(\left(\boldsymbol{\mathcal{A}}^{t,l}\right)^T * \boldsymbol{\mathcal{A}}^{t,l} - \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T * \boldsymbol{\mathcal{B}}^{t,l} \right), \\
\boldsymbol{\mathcal{B}}^{t+1,l} &= \boldsymbol{\mathcal{B}}^{t,l} - \eta \left(\boldsymbol{\mathcal{Z}}^c_{-l,\cdot} \boldsymbol{\mathcal{Z}}_{-l,\cdot} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) \right)^T * \boldsymbol{\mathcal{A}}^{t,l} - \eta \left(\boldsymbol{\mathcal{P}}_{l,\cdot} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) \right)^T * \boldsymbol{\mathcal{A}}^{t,l} \\
&\quad - \frac{m\eta}{2} \boldsymbol{\mathcal{B}}^{t,l} * \left(\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T * \boldsymbol{\mathcal{B}}^{t,l} - \left(\boldsymbol{\mathcal{A}}^{t,l}\right)^T * \boldsymbol{\mathcal{A}}^{t,l} \right).
\end{aligned} \tag{59}$$

For $n + 1 \leq l \leq 2n$, the modified loss function is

$$\begin{aligned}
\min_{\substack{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n \times r \times n_3} \\ \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n \times r \times n_3}}} \mathcal{L}^{(l)} \left(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \right) := &\left\| \boldsymbol{\mathcal{Z}}_{\cdot,-(l-n)} \left(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^T - \boldsymbol{\mathcal{X}}^\star \right) \right\|_F^2 + \left\| \boldsymbol{\mathcal{P}}_{\cdot,l-n} \left(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^T - \boldsymbol{\mathcal{X}}^\star \right) \right\|_F^2 \\
&+ \frac{m}{8} \left\| \boldsymbol{\mathcal{A}}^T * \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{B}}^T * \boldsymbol{\mathcal{B}} \right\|_F^2
\end{aligned} \tag{60}$$

The leave-one-out sequence generated with respect to the lateral slice is the gradient sequence as

$$\begin{aligned}
\boldsymbol{\mathcal{A}}^{t+1,l} &= \boldsymbol{\mathcal{A}}^{t,l} - \eta \boldsymbol{\mathcal{Z}}^c_{\cdot,-(l-n)} \boldsymbol{\mathcal{Z}}_{\cdot,-(l-n)} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) * \boldsymbol{\mathcal{B}}^{t,l} \\
&\quad - \eta \boldsymbol{\mathcal{P}}_{\cdot,-(l-n)} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) * \boldsymbol{\mathcal{B}}^{t,l}\star \\
&\quad - \frac{m\eta}{2} \boldsymbol{\mathcal{A}}^{t,l} * \left(\left(\boldsymbol{\mathcal{A}}^{t,l}\right)^T * \boldsymbol{\mathcal{A}}^{t,l} - \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T * \boldsymbol{\mathcal{B}}^{t,l} \right), \\
\boldsymbol{\mathcal{B}}^{t+1,l} &= \boldsymbol{\mathcal{B}}^{t,l} - \eta \left(\boldsymbol{\mathcal{Z}}^c_{\cdot,-(l-n)} \boldsymbol{\mathcal{Z}}_{\cdot,-(l-n)} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) \right)^T * \boldsymbol{\mathcal{A}}^{t,l} \\
&\quad - \eta \left(\boldsymbol{\mathcal{P}}_{\cdot,-(l-n)} \left(\boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{X}}^\star \right) \right)^T * \boldsymbol{\mathcal{A}}^{t,l} \\
&\quad - \frac{m\eta}{2} \boldsymbol{\mathcal{B}}^{t,l} * \left(\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T * \boldsymbol{\mathcal{B}}^{t,l} - \left(\boldsymbol{\mathcal{A}}^{t,l}\right)^T * \boldsymbol{\mathcal{A}}^{t,l} \right).
\end{aligned} \tag{61}$$

Without loss of generality, we consider the case where $l \in [n]$, then the formula has

$$\begin{aligned}
&\left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t+1,l} \\ \boldsymbol{\mathcal{B}}^{t+1,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right)_{l::} = \underbrace{\boldsymbol{\mathcal{A}}^{t,l}_{l::} * \boldsymbol{\mathcal{R}}^{t,l} - \boldsymbol{\mathcal{A}}^\star_{l::} - \eta m \left(\boldsymbol{\mathcal{A}}^{t,l}_{l::} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{A}}^\star_{l::} * \left(\boldsymbol{\mathcal{B}}^\star\right)^T \right) * \boldsymbol{\mathcal{B}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l}}_{\boldsymbol{\mathcal{E}}_1} \\
&+ \underbrace{\left(\boldsymbol{\mathcal{A}}^{t,l}_{l::} * \boldsymbol{\mathcal{R}}^{t,l} - \boldsymbol{\mathcal{A}}^\star_{l::} - \eta m \left(\boldsymbol{\mathcal{A}}^{t,l}_{l::} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{A}}^\star_{l::} * \left(\boldsymbol{\mathcal{B}}^\star\right)^T \right) * \boldsymbol{\mathcal{B}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} \right) * \left(\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1} * \boldsymbol{\mathcal{R}}^{t+1,l} - \boldsymbol{\mathcal{I}}_r \right)}_{\boldsymbol{\mathcal{E}}_2} \\
&\underbrace{- \frac{\eta m}{2} \boldsymbol{\mathcal{A}}^{t,l}_{l::} * \left(\left(\boldsymbol{\mathcal{A}}^{t,l}\right)^T * \boldsymbol{\mathcal{A}}^{t,l} - \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T * \boldsymbol{\mathcal{B}}^{t,l} \right) * \boldsymbol{\mathcal{R}}^{t+1,l}}_{\boldsymbol{\mathcal{E}}_3}.
\end{aligned} \tag{62}$$

For bounding the spectral norm of $\boldsymbol{\mathcal{E}}_1$, with defining the notation $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l} := \boldsymbol{\mathcal{A}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} - \boldsymbol{\mathcal{A}}^{\star}$, $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l} := \boldsymbol{\mathcal{B}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} - \boldsymbol{\mathcal{B}}^{\star}$, then there is

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{E}}_1\right\| = & \left\| \left(\boldsymbol{\mathcal{I}}_r - m\eta \left(\boldsymbol{\mathcal{B}}^{\star}\right)^T * \boldsymbol{\mathcal{B}}^{\star}\right) * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::} - m\eta \left(\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::} * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right)^T + \boldsymbol{\mathcal{A}}_{l::}^{\star} * \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right)^T \right) * \boldsymbol{\mathcal{B}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} \right. \\
& \left. - \eta m \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::} * \left(\boldsymbol{\mathcal{B}}^{\star}\right)^T * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l} \right\| \\
\leq & \left\|\boldsymbol{\mathcal{I}}_r - m\eta \left(\boldsymbol{\mathcal{B}}^{\star}\right)^T * \boldsymbol{\mathcal{B}}^{\star}\right\| \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| + \eta m \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| \left(\left\|\boldsymbol{\mathcal{B}}^{t,l}\right\| + \left\|\boldsymbol{\mathcal{B}}^{\star}\right\|\right) \\
& + \eta m \left\|\boldsymbol{\mathcal{A}}_{l::}^{\star}\right\| \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| \left\|\boldsymbol{\mathcal{B}}^{t,l}\right\| \\
\leq & \left\|\boldsymbol{\mathcal{I}}_r - m\eta \left(\boldsymbol{\mathcal{B}}^{\star}\right)^T * \boldsymbol{\mathcal{B}}^{\star}\right\| \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| + \eta m \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| \left(\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| + 2\left\|\boldsymbol{\mathcal{B}}^{\star}\right\|\right) \\
& + \eta m \left\|\boldsymbol{\mathcal{A}}_{l::}^{\star}\right\| \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| \left(\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| + \left\|\boldsymbol{\mathcal{B}}^{\star}\right\|\right).
\end{aligned} \tag{63}
$$

Now comes to bound the $\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\|$ as follows

$$
\begin{aligned}
\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\| \leq \left\|\boldsymbol{\Delta}^{t,l}\right\| & = \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right\| \\
& \leq \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} \right\| + \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right\| \\
& \overset{(i)}{\leq} 9\kappa \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} \right\| + \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right\| \\
& \overset{(ii)}{\leq} 9\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^2 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}} + 98C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \\
& \leq 12\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}.
\end{aligned} \tag{64}
$$

(i) reuses the Lemma 6 to bound the first term of RHS of the above inequality as $\boldsymbol{\mathcal{X}}_0 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix}$, $\boldsymbol{\mathcal{X}}_1 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t}$, and $\boldsymbol{\mathcal{X}}_2 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l}$. It is apparent that $\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_0\right\| \left\|\boldsymbol{\mathcal{X}}_0\right\| \leq \frac{\sigma_{\max}^2(\boldsymbol{\mathcal{A}}^{\star})}{2}$ and $\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| \left\|\boldsymbol{\mathcal{X}}_0\right\| \leq \frac{\sigma_{\max}^2(\boldsymbol{\mathcal{A}}^{\star})}{4}$ based on (13), (23) and tube-wise sample complexity (12). (ii) is due to (13) and (23). Substituting above inequality into (63) would have

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{E}}_1\right\| \leq & \left(1 - \eta m \sigma_{\min}\right) \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| + \eta m \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| \cdot 12\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \\
& \cdot \left(2\sqrt{\sigma_{\max}} + 12\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}\right) \\
& + \eta m \sqrt{\frac{\mu r \sigma_{\max}}{n}} \cdot 12\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \\
& \cdot \left(\sqrt{\sigma_{\max}} + 12\kappa \cdot 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}\right) \\
\overset{(i)}{\leq} & \left(1 - \eta m \sigma_{\min}\right) \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| + 1872C\eta m \sigma_{\max}\kappa\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)}{nm}} \left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| \\
& + \eta m \sigma_{\max} \sqrt{\frac{\mu r}{n}} \cdot 1248C\kappa\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(ii)}{\le} (1-0.85\eta m\sigma_{\min})\left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right)_{l::}\right\| + 0.5\eta m\sigma_{\min}\cdot 2496C\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\overset{(iii)}{\le} (1-0.6\eta m\sigma_{\min})\cdot 2496C\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\le \sqrt{\frac{\mu r\sigma_{\max}}{n}}, \qquad (65)
\end{aligned}
$$

where the (i) and (ii) are due to (12). (iii) is due to (22) and step size condition in Theorem 2.

For the $\boldsymbol{\mathcal{E}}_2$, we have the similar bound as (48) in [8] as

$$
\begin{aligned}
\|\boldsymbol{\mathcal{E}}_2\| &\le \|\boldsymbol{\mathcal{E}}_1 + (\boldsymbol{\mathcal{A}}^\star)_{l::}\| \left\|\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1} * \boldsymbol{\mathcal{R}}^{t+1,l} - \boldsymbol{\mathcal{I}}_r\right\|\\
&\overset{(i)}{\le} 2\sqrt{\frac{\mu r\sigma_{\max}}{n}}\left\|\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1} * \boldsymbol{\mathcal{R}}^{t+1,l} - \boldsymbol{\mathcal{I}}_r\right\|\\
&\le 2\sqrt{\frac{\mu r\sigma_{\max}}{n}}\cdot 2\times 1250^2 m\eta C^2\rho^{2t}\left(\sqrt{\frac{(\mu r)^4\kappa^8 n_3\log^2(n\vee n_3)\sigma_{\max}}{nm}}\right)^2 \qquad (66)
\end{aligned}
$$

where the (i) is due to (65) and the last inequality is due to Lemma 8.

For the $\boldsymbol{\mathcal{E}}_3$, there is

$$
\begin{aligned}
\|\boldsymbol{\mathcal{E}}_3\| &\le \frac{\eta m}{2}\left\|\left(\boldsymbol{\mathcal{A}}^{t,l}\right)_{l::}\right\|\left(2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right\|\|\boldsymbol{\mathcal{A}}^\star\| + 2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\|\|\boldsymbol{\mathcal{B}}^\star\| + \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\right\|^2 + \left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right\|^2\right)\\
&\le \frac{\eta m}{2}\left(\left\|\left(\boldsymbol{\Delta}^{t,l}\right)_{l::}\right\| + \|\boldsymbol{\mathcal{A}}^\star_{l::}\|\right)\cdot\left(4\left\|\boldsymbol{\Delta}^{t,l}\right\|\|\boldsymbol{\mathcal{A}}^\star\| + 2\left\|\boldsymbol{\Delta}^{t,l}\right\|^2\right)\\
&\overset{(i)}{\le} \frac{\eta m}{2}\left(2496C\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}} + \sqrt{\frac{\mu r\sigma_{\max}}{n}}\right)\cdot 6\sqrt{\sigma_{\max}}\left\|\boldsymbol{\Delta}^{t,l}\right\|\\
&\overset{(ii)}{\le} 6\eta m\sigma_{\max}\sqrt{\frac{\mu r}{n}}\left\|\boldsymbol{\Delta}^{t,l}\right\|\\
&\le 1872Cm\eta\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}, \qquad (67)
\end{aligned}
$$

where (i) uses the result in (22) and tensor incoherence condition in (10). (ii) and last inequality use the tube-wise sample complexity and (64), respectively.

Finally, combing the (65), (66) and (67) would induce

$$
\begin{aligned}
\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l}\end{bmatrix} * \boldsymbol{\mathcal{R}}^{t+1,l} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right)_{l::}\right\| &\le (1-0.6\eta m\sigma_{\min})\cdot 2496C\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\quad + (2500C)^2 m\eta\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^4\kappa^9 n_3\log^2(n\vee n_3)}{nm}}\\
&\quad \cdot\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\quad + 1872Cm\eta\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\le 2496C\rho^{t+1}\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}. \qquad (68)
\end{aligned}
$$

The last inequality is due to the tube-wise sample complexity in (12).

A.4 Proof of (23)

Based on definition of $\boldsymbol{\mathcal{T}}^{t,l}$ in (21), there is

$$\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t+1}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l}\end{bmatrix}*\boldsymbol{\mathcal{T}}^{t+1,l}\right\| \leq \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l}\end{bmatrix}*\boldsymbol{\mathcal{T}}^{t,l}\right\| \tag{69}$$

Without loss of generality, we can consider the leave-one-out sequence under $l \in [n]$. Based on (7), (8) and (59), there is decomposition

$$\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l}\end{bmatrix}*\boldsymbol{\mathcal{T}}^{t,l}=\boldsymbol{\mathcal{E}}_1+\eta\begin{bmatrix}\boldsymbol{\mathcal{E}}_2\\ \boldsymbol{\mathcal{E}}_3\end{bmatrix}, \tag{70}$$

where the $\boldsymbol{\mathcal{E}}_1, \boldsymbol{\mathcal{E}}_2$ and $\boldsymbol{\mathcal{E}}_3$ are redefined as

$$\boldsymbol{\mathcal{E}}_1 := \left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}\\ \boldsymbol{\mathcal{B}}^{t}\end{bmatrix}-\eta\nabla\mathcal{L}\left(\boldsymbol{\mathcal{A}}^{t},\boldsymbol{\mathcal{B}}^{t}\right)\right)*\boldsymbol{\mathcal{R}}^{t}-\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix}-\eta\nabla\mathcal{L}\left(\boldsymbol{\mathcal{A}}^{t,l},\boldsymbol{\mathcal{B}}^{t,l}\right)\right)*\boldsymbol{\mathcal{T}}^{t,l} \tag{71}$$

$$\boldsymbol{\mathcal{E}}_2 := \left[m\boldsymbol{\mathcal{P}}_{l,\cdot}\left(\boldsymbol{\mathcal{A}}^{t,l}*\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T-\boldsymbol{\mathcal{A}}^{\star}*\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\right)-\boldsymbol{\mathcal{Z}}^c_{l,\cdot}\boldsymbol{\mathcal{Z}}_{l,\cdot}\left(\boldsymbol{\mathcal{A}}^{t,l}*\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T-\boldsymbol{\mathcal{A}}^{\star}*\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\right)\right]*\boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l} \tag{72}$$

$$\boldsymbol{\mathcal{E}}_3 := \left[m\boldsymbol{\mathcal{P}}_{l,\cdot}\left(\boldsymbol{\mathcal{A}}^{t,l}*\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T-\boldsymbol{\mathcal{A}}^{\star}*\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\right)-\boldsymbol{\mathcal{Z}}^c_{l,\cdot}\boldsymbol{\mathcal{Z}}_{l,\cdot}\left(\boldsymbol{\mathcal{A}}^{t,l}*\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T-\boldsymbol{\mathcal{A}}^{\star}*\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\right)\right]^T*\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}, \tag{73}$$

where $\boldsymbol{\mathcal{Z}}_{i,\cdot}, i \in [n_1]$ denotes the tube-wise sensing operator that only tubes in i-th horizontal slice are sensed based on $\boldsymbol{\mathcal{Z}}$ and all other tubes are not sensed, i.e., for given $i \in [n]$, there is

$$\left[\boldsymbol{\mathcal{Z}}_{i,\cdot}\left(\boldsymbol{\mathcal{X}}\right)\right]_{kj:} := \begin{cases}\boldsymbol{Z}_{kj}\boldsymbol{\mathcal{X}}_{kj:}, & k=i, \forall j\in[n],\\ \mathbf{0}_m, & k\neq i, \forall j \in [n].\end{cases} \tag{74}$$

Now we bound each of the error terms $\|\mathcal{E}_1\|$, $\|\mathcal{E}_2\|$, and $\|\mathcal{E}_3\|$ in turn. For $\|\mathcal{E}_1\|$, we apply Lemma 7, which provides restricted strong convexity and smoothness of the non-convex loss, to establish a local contraction of the distance between the auxiliary sequence and the true GD iterates. The terms $\|\mathcal{E}_2\|$ and $\|\mathcal{E}_3\|$ are then controlled by our concentration bounds, ensuring they remain sufficiently small.

For $\boldsymbol{\mathcal{E}}_1$, there exists $\tau \in [0,1]$ such that

$$\begin{aligned}\|\boldsymbol{\mathcal{E}}_1\| &= \left\|\left(\boldsymbol{\mathcal{I}}-\eta\nabla^2\mathcal{L}\left(\tau\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}+(1-\tau)\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t},\tau\boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}+(1-\tau)\boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}\right)\right)\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}\\ \boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}\end{bmatrix}\right\|\\ &\leq \left(1-\frac{m\eta\sigma_{\min}}{5}\right)\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}\\ \boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}\end{bmatrix}\right\|.\end{aligned} \tag{75}$$

The inequality is based on Lemma 7 as follows. Based on (79), (14) and tube-wise sample complexity (12), we can conclude that

$$\begin{aligned}\max_l\left\|\begin{bmatrix}\tau\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}+(1-\tau)\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{U}}^{\star}\\ \tau\boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}+(1-\tau)\boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{V}}^{\star}\end{bmatrix}_{l::}\right\| &\leq \tau\max_l\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{U}}^{\star}\\ \boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{V}}^{\star}\end{bmatrix}_{l::}\right\|\\ &\quad+(1-\tau)\max_l\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{U}}^{\star}\\ \boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{V}}^{\star}\end{bmatrix}_{l::}\right\|\\ &\leq \frac{\sqrt{\sigma_{\max}}}{500\kappa\sqrt{2n}}\end{aligned}$$

$$\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{U}}^{\star}\\ \boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{V}}^{\star}\end{bmatrix}\right\| \leq \sqrt{\sum_{i=1}^{2n}\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{U}}^{\star}\\ \boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}-\boldsymbol{\mathcal{V}}^{\star}\end{bmatrix}\right)_{i::}\right\|^2}$$

$$\le \frac{\sqrt{\sigma_{\max}}}{500\kappa}, \tag{76}$$

where the third inequality is due to norm inequality (92). The above two conditions are satisfied in Lemma 7.

For $\boldsymbol{\mathcal{E}}_2, \boldsymbol{\mathcal{E}}_3$, we have

$$\begin{aligned}
\left\|\boldsymbol{\mathcal{E}}_2\right\| &= \left\| \sum_{j=1}^{n} \left(\boldsymbol{Z}_{lj}^T \boldsymbol{Z}_{lj} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{lj:} * \boldsymbol{\mathcal{B}}_{j::}^{t,l} \right\|, \\
\left\|\boldsymbol{\mathcal{E}}_3\right\| &= \left\| \begin{bmatrix} \left(\boldsymbol{Z}_{l1}^T \boldsymbol{Z}_{l1} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{l1:} \\ \left(\boldsymbol{Z}_{l2}^T \boldsymbol{Z}_{l2} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{l2:} \\ \vdots \\ \left(\boldsymbol{Z}_{ln}^T \boldsymbol{Z}_{ln} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{ln:} \end{bmatrix} * \boldsymbol{\mathcal{A}}_{l::}^{t,l} \right\| \\
&\le \left\| \begin{bmatrix} \left(\boldsymbol{Z}_{l1}^T \boldsymbol{Z}_{l1} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{l1:} \\ \left(\boldsymbol{Z}_{l2}^T \boldsymbol{Z}_{l2} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{l2:} \\ \vdots \\ \left(\boldsymbol{Z}_{ln}^T \boldsymbol{Z}_{ln} - m\boldsymbol{I}\right)\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{ln:} \end{bmatrix} \right\| * \left\|\boldsymbol{\mathcal{A}}_{l::}^{t,l}\right\| \qquad (77)
\end{aligned}$$

where $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} := \boldsymbol{\mathcal{A}}^{t,l} * \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^T - \boldsymbol{\mathcal{A}}^{\star} * (\boldsymbol{\mathcal{B}}^{\star})^T$ and we would bound the $\left\|\boldsymbol{\mathcal{E}}_2\right\|, \left\|\boldsymbol{\mathcal{E}}_3\right\|$ separately. Before that, we focus on uniform tube-wise bound for $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}$. $\forall j \in [n]$,

$$\begin{aligned}
\left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{lj:}\right\|_F &= \left\| \boldsymbol{\mathcal{A}}_{l::}^{t,l} * \left(\boldsymbol{\mathcal{B}}_{j::}^{t,l}\right)^T - \boldsymbol{\mathcal{A}}_{l::}^{\star} * \left(\boldsymbol{\mathcal{B}}_{j::}^{\star}\right)^T \right\|_F \\
&= \left\| \left(\boldsymbol{\mathcal{A}}_{l::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{A}}_{l::}^{\star}\right) * \left(\boldsymbol{\mathcal{B}}_{j::}^{\star}\right)^T + \boldsymbol{\mathcal{A}}_{l::}^{\star} * \left(\boldsymbol{\mathcal{B}}_{j::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{B}}_{j::}^{\star}\right)^T \right. \\
&\quad \left. + \left(\boldsymbol{\mathcal{A}}_{l::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{A}}_{l::}^{\star}\right) * \left(\boldsymbol{\mathcal{B}}_{j::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{B}}_{j::}^{\star}\right)^T \right\|_F \\
&\overset{(i)}{\le} \left\| \boldsymbol{\mathcal{A}}_{l::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{A}}_{l::}^{\star} \right\| \left\|\boldsymbol{\mathcal{B}}_{j::}^{\star}\right\| + \left\|\boldsymbol{\mathcal{A}}_{l::}^{\star}\right\| \left\| \boldsymbol{\mathcal{B}}_{j::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{B}}_{j::}^{\star} \right\| \\
&\quad + \left\| \boldsymbol{\mathcal{A}}_{l::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{A}}_{l::}^{\star} \right\| \left\| \boldsymbol{\mathcal{B}}_{j::}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{B}}_{j::}^{\star} \right\| \\
&\overset{(ii)}{\le} 1292C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n)\sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \\
&\quad + \left(646C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n)\sigma_{\max}}{n^2 m}}\right)^2 \\
&\le 1300C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n)\sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}}. \qquad (78)
\end{aligned}$$

(i) is based on (95), and the last inequality is due to the tube-wise sample complexity in (12). (ii) is based on the tensor incoherence condition (10) and the following inequality.

$$\begin{aligned}
\max_l \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| &\le \max_l \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} \right)_{l::} \right\| \\
&\quad + \max_l \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\|
\end{aligned}$$

$$\begin{aligned}
&\overset{(iv)}{\le} \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} \right\| + \max_{l} \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t} \\ \boldsymbol{\mathcal{B}}^{t} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| \\
&\overset{(v)}{\le} 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^2 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \\
&\quad + 594C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \\
&\le 646C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}}, \qquad (79)
\end{aligned}$$

where (iv) we use inequality in (92) and (v) are due to (23) and (13). Now we bound the $\|\boldsymbol{\mathcal{E}}_2\|$ based on Lemma 2 as

$$\begin{aligned}
\|\boldsymbol{\mathcal{E}}_2\| &\le 2Cmr\sqrt{\frac{n_3}{m}} \log(n \vee n_3) \max_{ij} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{ij:} \right\|_F \left\| \boldsymbol{\mathcal{B}}^{t,l} \right\| \\
&\overset{(vi)}{\le} 2Cmr\sqrt{\frac{n_3}{m}} \log(n \vee n_3) \cdot 1300C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \\
&\quad \cdot \left(\left\| \boldsymbol{\mathcal{B}}^{t,l} * \boldsymbol{\mathcal{T}}^{t,l} - \boldsymbol{\mathcal{B}}^{\star} \right\| + \|\boldsymbol{\mathcal{B}}^{\star}\| \right) \\
&\overset{(vii)}{\le} 2600C^2 m \sqrt{\frac{(\mu r)^3 \log^2(n \vee n_3) n_3 \sigma_{\max}}{nm}} \cdot \rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \\
&\quad \cdot \left(646C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{nm}} + \sqrt{\sigma_{\max}} \right) \\
&\le 5200C^2 m \sigma_{\min} \rho^t \sqrt{\frac{(\mu r)^3 \kappa^2 \log^2(n \vee n_3) n_3}{nm}} \cdot \rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \\
&\le 2Cm\sigma_{\min} \rho^t \sqrt{\frac{(\mu r)^3 \kappa^2 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}}, \qquad (80)
\end{aligned}$$

where (vi) and (vii) are due to (78), (79), respectively. The last two inequalities are due to tube-wise sample complexity (12).

Now comes to bound the $\|\boldsymbol{\mathcal{E}}_3\|$. To obtain a tight bound, we also bound the spectral norm of each tube, $\left(\boldsymbol{Z}_{lj}^T \boldsymbol{Z}_{lj} - m\boldsymbol{I} \right) \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:}$. $\forall j \in [n]$, there is

$$\begin{aligned}
\frac{t^2}{\sum_{s=1}^m K_s^2} &\ge \frac{t^2}{\sum_{s=1}^m \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F^2 \|\boldsymbol{\mathcal{T}}\|_F^2} \ge \frac{t^2}{m \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F^2 \|\boldsymbol{\mathcal{T}}\|_F^2} = \frac{t^2}{mn_3 \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F^2} \\
\frac{t}{\max_s K_s} &\ge \frac{t}{\max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F \|\boldsymbol{\mathcal{T}}\|_F} = \frac{t}{\sqrt{n_3} \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F}. \qquad (81)
\end{aligned}$$

Based on the sub-exponential Bernstein inequality and ϵ-net argument,

$$\left\| \left(\boldsymbol{Z}_{lj}^T \boldsymbol{Z}_{lj} - m\boldsymbol{I} \right) \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\| \le \sqrt{mn_3} \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F \sqrt{\log(n \vee n_3)} + \sqrt{n_3} \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F \log(n \vee n_3) \qquad (82)$$

holds with at least $1 - \exp(-c\log(n \vee n_3))$ probability. Finally using the union bound and inequality (92), we can obtain the bound for $\boldsymbol{\mathcal{E}}_3$ as

$$\|\boldsymbol{\mathcal{E}}_3\| \le C\sqrt{n} \cdot \left(\sqrt{mn_3} \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F \sqrt{\log(n \vee n_3)} + \sqrt{n_3} \max_{l,j} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l} \right)_{lj:} \right\|_F \log(n \vee n_3) \right) \cdot \left\| \boldsymbol{\mathcal{A}}_{l::}^{t,l} \right\|$$

$$
\begin{aligned}
&\leq 2Cm\sqrt{\frac{nn_3}{m}}\log(n\vee n_3)\max_{l,j}\left\|\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{lj:}\right\|_F\cdot\left\|\boldsymbol{\mathcal{A}}_{l::}^{t,l}\right\|\\
&\overset{(i)}{\leq} 2Cm\sqrt{\frac{nn_3\log^2(n\vee n_3)}{m}}\cdot 1300C\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n_3\vee n)\sigma_{\max}}{n^2m}}\cdot\sqrt{\frac{\mu r\sigma_{\max}}{n}}\\
&\quad\cdot\left(\left\|\boldsymbol{\mathcal{A}}_{l::}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{A}}_{l::}^{\star}\right\|+\|\boldsymbol{\mathcal{A}}_{l::}^{\star}\|\right)\\
&\overset{(ii)}{\leq} 2600C^2m\rho^t\sqrt{\frac{\mu rn_3\log^2(n\vee n_3)\sigma_{\max}}{m}}\cdot\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n_3\vee n)\sigma_{\max}}{n^2m}}\\
&\quad\cdot\left(646C\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n_3\vee n)\sigma_{\max}}{n^2m}}+\sqrt{\frac{\mu r\sigma_{\max}}{n}}\right)\\
&\leq 5200C^2m\sigma_{\min}\sqrt{\frac{(\mu r)^2\kappa^2 n_3\log^2(n\vee n_3)}{nm}}\cdot\rho^t\sqrt{\frac{(\mu r)^5\kappa^9 n_3\log^2(n_3\vee n)\sigma_{\max}}{n^2m}}\\
&\leq 2Cm\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^2\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}},
\end{aligned}
\tag{83}
$$

where (i) is due to (78) and triangle inequality. (ii) is because (79) and tensor incoherence condition (10). The last two inequalities are due to tube-wise sample complexity in (12).

Finally, combining the bounds in (75), (80) and (83) with (70) would have

$$
\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1}\\ \boldsymbol{\mathcal{B}}^{t+1}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t+1}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l}\end{bmatrix}*\boldsymbol{\mathcal{T}}^{t+1,l}\right\| &\leq \left(1-\frac{m\eta\sigma_{\min}}{5}\right)\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{A}}^{t}*\boldsymbol{\mathcal{R}}^{t}\\ \boldsymbol{\mathcal{B}}^{t,l}*\boldsymbol{\mathcal{T}}^{t,l}-\boldsymbol{\mathcal{B}}^{t}*\boldsymbol{\mathcal{R}}^{t}\end{bmatrix}\right\|\\
&\quad+2Cm\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^3\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\quad+2Cm\sigma_{\min}\rho^t\sqrt{\frac{(\mu r)^2\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\leq\left(1-\frac{m\eta\sigma_{\min}}{5}\right)\cdot 52C\rho^t\sqrt{\frac{(\mu r)^4\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\quad+0.08\times 52C\rho^t\sqrt{\frac{(\mu r)^3\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}}\\
&\leq 52C\rho^{t+1}\sqrt{\frac{(\mu r)^4\kappa^2 n_3\log^2(n\vee n_3)\sigma_{\max}}{n^2m}},
\end{aligned}
\tag{84}
$$

where the last inequality is due to the $\rho=(1-0.01m\eta\sigma_{\min})$.

A.5 Proof of (14)

This proof is combining (23) and (22) by triangle inequality. Thus. there is decomposition as

$$
\begin{aligned}
\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}\\ \boldsymbol{\mathcal{B}}^{t}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right)_{l::}\right\| &\leq\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}\\ \boldsymbol{\mathcal{B}}^{t}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t,l}\right)_{l::}\right\|+\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t,l}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right)_{l::}\right\|\\
&\leq\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t}\\ \boldsymbol{\mathcal{B}}^{t}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t,l}\right\|+\left\|\left(\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t,l}\end{bmatrix}*\boldsymbol{\mathcal{R}}^{t,l}-\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right)_{l::}\right\|,
\end{aligned}
\tag{85}
$$

where the last inequality is due to tensor norm inequality (92). For bound the $\left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^t \\ \boldsymbol{\mathcal{B}}^t \end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t,l} \right\|$, we apply the Lemma 6 again as

$$\boldsymbol{\mathcal{X}}_0 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix}, \quad \boldsymbol{\mathcal{X}}_1 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^t \\ \boldsymbol{\mathcal{B}}^t \end{bmatrix} * \boldsymbol{\mathcal{R}}^t, \quad \boldsymbol{\mathcal{X}}_2 = \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l}. \tag{86}$$

Based on tube-wise sample complexity, there is

$$\begin{aligned} \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_0\| \|\boldsymbol{\mathcal{X}}_0\| &\overset{(i)}{\leq} 98C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{nm}} \cdot \sqrt{\sigma_{\max}} \\ &= 98C\sigma_{\min}\rho^t \sqrt{\frac{(\mu r)^4 \kappa^7 n_3 \log^2(n \vee n_3)}{nm}} \\ &\leq \frac{\sigma_{\min}}{2}, \end{aligned} \tag{87}$$

where (i) is due to (13) and the last is due to tube-wise sample complexity. In addition,

$$\begin{aligned} \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| \|\boldsymbol{\mathcal{X}}_0\| &\leq 52C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^2 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}} \cdot \sqrt{\sigma_{\max}} \\ &= 52C\sigma_{\min}\rho^t \sqrt{\frac{(\mu r)^4 \kappa^4 n_3 \log^2(n \vee n_3)}{n^2 m}} \\ &\leq \frac{\sigma_{\min}}{4}, \end{aligned} \tag{88}$$

where (i) the first inequality is due to (23) and the last is due to tube-wise sample complexity. Thus,

$$\begin{aligned} \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^t \\ \boldsymbol{\mathcal{B}}^t \end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right)_{l::} \right\| &\leq 9\kappa \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^t \\ \boldsymbol{\mathcal{B}}^t \end{bmatrix} * \boldsymbol{\mathcal{R}}^t - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{t,l} \right\| + \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{t,l} \\ \boldsymbol{\mathcal{B}}^{t,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{t,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right)_{l::} \right\| \\ &\leq 468C\rho^t \sqrt{\frac{(\mu r)^4 \kappa^4 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}} \\ &\quad + 2496C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}} \\ &\leq 2964C\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n \vee n_3)\sigma_{\max}}{n^2 m}}. \end{aligned} \tag{89}$$

A.6 Technical Lemmas

This section presents the technical lemmas used in proofs of Section A. The first lemma is some important tensor norm inequalities under the t-SVD framework, which will be repeatedly used throughout the proof.

Lemma 1. *For* $\forall \boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_2 \times n_4 \times n_3}$*, there are tensor norm inequalities as*

$$\|\boldsymbol{\mathcal{X}}\| \leq \sqrt{n_3} \|\boldsymbol{\mathcal{X}}\|_F; \tag{90}$$

$$\|\boldsymbol{\mathcal{X}}(i,:,:)\|_F \leq \|\boldsymbol{\mathcal{X}}(i,:,:)\|, \quad \forall i \in [n_1]; \tag{91}$$

$$\max_{i \in [n_1]} \|\boldsymbol{\mathcal{X}}_{i::}\| \leq \|\boldsymbol{\mathcal{X}}\| \leq \sqrt{\sum_{i=1}^{n_1} \|\boldsymbol{\mathcal{X}}(i,:,:)\|^2}; \tag{92}$$

$$\|\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}}\|_F \leq \|\boldsymbol{\mathcal{X}}\| \|\boldsymbol{\mathcal{Y}}\|_F \leq \sqrt{n_3} \|\boldsymbol{\mathcal{X}}\|_F \|\boldsymbol{\mathcal{Y}}\|_F; \tag{93}$$

$$\|\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}}\| \leq \|\boldsymbol{\mathcal{X}}\| \|\boldsymbol{\mathcal{Y}}\|; \tag{94}$$

$$\|\boldsymbol{\mathcal{X}}(i,:,:) * \boldsymbol{\mathcal{Y}}(:,j,:)\|_F \leq \|\boldsymbol{\mathcal{X}}(i,:,:)\| \|\boldsymbol{\mathcal{Y}}(:,j,:)\|, \quad \forall i \in [n_1], \forall j \in [n_4]. \tag{95}$$

Remark 3. *The relationship between the spectral norm and the Frobenius norm in the t-SVD tensor algebra differs from the matrix case. For a matrix X, one has*

$$\|X\| \leq \|X\|_F,$$

whereas in the tensor setting an extra factor of $\sqrt{n_3}$ appears in (90)*. Similarly, the matrix inequality*

$$\|XY\|_F \leq \|X\|_F \, \|Y\|_F$$

also acquires a factor of $\sqrt{n_3}$ in (93)*. This $\sqrt{n_3}$ factor is unavoidable in the general third-order tensor context. Interestingly, for a horizontal slice tensor $\boldsymbol{\mathcal{X}}(i,:,:) \in \mathbb{R}^{1 \times n_2 \times n_3}$, the Frobenius norm is bounded by its spectral norm, reversing the usual matrix inequality.*

Proof. For the (90), we have

$$\|\boldsymbol{\mathcal{X}}\|^2 \overset{(a)}{=} \|\text{bcirc}\,(\boldsymbol{\mathcal{X}})\|^2 \overset{(b)}{\leq} \|\text{bcirc}\,(\boldsymbol{\mathcal{X}})\|_F^2 \overset{(c)}{=} \sqrt{n_3}\,\|\boldsymbol{\mathcal{X}}\|_F^2\,, \tag{96}$$

where we use the definition of tensor spectral norm in (a), matrix norm inequality in (b). (c) is due to the definition of the bcirc operator in (3) and equality (102).

For (91),

$$\|\boldsymbol{\mathcal{X}}(i,:,:)\|_F = \sqrt{\frac{\sum_{k=1}^{n_3} \left\|\overline{\boldsymbol{X}}_{i:}^{(k)}\right\|^2}{n_3}} \leq \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}_{i:}^{(k)}\right\| = \|\boldsymbol{\mathcal{X}}(i,:,:)\| \tag{97}$$

For (92),

$$\begin{aligned}
\|\boldsymbol{\mathcal{X}}\| &= \left\|\text{bdiag}\left(\overline{\boldsymbol{\mathcal{X}}}\right)\right\| \\
&= \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}^{(k)}\right\| \\
&\overset{(d)}{\leq} \max_{k\in[n_3]} \sqrt{\sum_{i=1}^{n_1} \left\|\overline{\boldsymbol{X}}_{i,:}^{(k)}\right\|_F^2} \\
&\leq \sqrt{\sum_{i=1}^{n_1} \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}_{i,:}^{(k)}\right\|_F^2} \\
&\overset{(e)}{=} \sqrt{\sum_{i=1}^{n_1} \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}_{i,:}^{(k)}\right\|^2} \\
&\overset{(f)}{=} \sqrt{\sum_{i=1}^{n_1} \|\boldsymbol{\mathcal{X}}(i,:,:)\|^2},
\end{aligned} \tag{98}$$

where (d) uses matrix norm inequality that $\|\boldsymbol{X}\| \leq \|\boldsymbol{X}\|_F$. (e) is based on the fact that for a matrix with only one row, its spectral norm is equal to its Frobenius norm. (f) is due to the definition of the tensor spectral norm.

For (93),

$$\begin{aligned}
\|\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}}\|_F &= \frac{1}{\sqrt{n_3}} \left\|\overline{\boldsymbol{X}\boldsymbol{Y}}\right\|_F \\
&= \frac{\sqrt{\sum_{k=1}^{n_3} \left\|\overline{\boldsymbol{X}}^{(k)}\overline{\boldsymbol{Y}}^{(k)}\right\|_F^2}}{\sqrt{n_3}} \\
&\overset{(g)}{\leq} \frac{\sqrt{\sum_{k=1}^{n_3} \left\|\overline{\boldsymbol{Y}}^{(k)}\right\|_F^2 \left\|\overline{\boldsymbol{X}}^{(k)}\right\|^2}}{\sqrt{n_3}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}^{(k)}\right\| \frac{\sqrt{\sum_{k=1}^{n_3}\left\|\overline{\boldsymbol{Y}}^{(k)}\right\|_F^2}}{\sqrt{n_3}} \\
&= \|\boldsymbol{\mathcal{X}}\| \, \|\boldsymbol{\mathcal{Y}}\|_F \,.
\end{aligned} \tag{99}
$$

(g) is based on matrix norm inequality that $\|\boldsymbol{X}\boldsymbol{Y}\|_F \leq \|\boldsymbol{Y}\|_F \, \|\boldsymbol{X}\|$.

For (94), $\|\boldsymbol{\mathcal{X}} * \boldsymbol{\mathcal{Y}}\| = \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}^{(k)}\overline{\boldsymbol{Y}}^{(k)}\right\| \leq \max_{k\in[n_3]} \left\|\overline{\boldsymbol{X}}^{(k)}\right\| \max_{k\in[n_3]} \left\|\overline{\boldsymbol{Y}}^{(k)}\right\| = \|\boldsymbol{\mathcal{X}}\| \, \|\boldsymbol{\mathcal{Y}}\|$.

For (95), it is directly result from (92) and (93) as $\|\boldsymbol{\mathcal{X}}(i,:,:) * \boldsymbol{\mathcal{Y}}(:,j,:)\|_F \leq \|\boldsymbol{\mathcal{X}}(i,:,:)\|_F \, \|\boldsymbol{\mathcal{Y}}(:,j,:)\| \leq \|\boldsymbol{\mathcal{X}}(i,:,:)\| \, \|\boldsymbol{\mathcal{Y}}(:,j,:)\|$. □

The following lemma is the key lemma in our proof, which measures the concentration of the tube-wise sensing map $\boldsymbol{\mathcal{Z}} : \mathbb{R}^{n\times n\times n_3} \to \mathbb{R}^{n\times n\times m}$ of Definition 1 in the tensor spectral norm metric.

Lemma 2. *For the defined tube-wise sensing operator $\boldsymbol{\mathcal{Z}}$, all entries in $\{\boldsymbol{Z}_{ij}\}$ obey i.i.d standard Gaussian distribution, then $\forall$ tensor $\boldsymbol{\Delta} \in \mathbb{R}^{n\times n\times n_3}, \boldsymbol{\mathcal{F}} \in \mathbb{R}^{n\times r\times n_3}$, the*

$$
\|(m\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}})(\boldsymbol{\Delta}) * \boldsymbol{\mathcal{F}}\| \leq 2Cmr\sqrt{\frac{nn_3}{m}}\log(n\vee n_3)\max_{ij}\|\boldsymbol{\Delta}_{ij:}\|_F \, \|\boldsymbol{\mathcal{F}}\| \tag{100}
$$

holds with high probability at least $1 - \frac{1}{(n\vee n_3)^{10}}$.

Proof. Due to the independence of each horizontal slice from $\boldsymbol{\mathcal{Z}}$, the core idea is to bound each horizontal slice separately and utilize the union bound to control concentration. Denote $\boldsymbol{\mathcal{C}} := (m\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}})(\boldsymbol{\Delta}) * \boldsymbol{\mathcal{F}}$, then $\boldsymbol{\mathcal{C}}_{i::} = \sum_{j=1}^{n}\left(\boldsymbol{Z}_{ij}^T\boldsymbol{Z}_{ij}\boldsymbol{\Delta}_{ij:} - m\boldsymbol{I}\right) * \boldsymbol{\mathcal{F}}(j,:,:) \in \mathbb{R}^{1\times r\times n_3}$. Based on the variational formula of the matrix spectral norm, there is

$$
\begin{aligned}
\|\boldsymbol{\mathcal{C}}_{i::}\| &\overset{(i)}{=} \max_{\boldsymbol{w}\in\boldsymbol{\Theta}^1,\boldsymbol{z}\in\boldsymbol{\Theta}^r} \langle\overline{\boldsymbol{C}_{i::}}, \boldsymbol{w}\boldsymbol{z}^T\rangle \\
&\overset{(ii)}{=} \max_{\boldsymbol{w}\in\boldsymbol{\Theta}^1,\boldsymbol{z}\in\boldsymbol{\Theta}^r} \langle(\boldsymbol{F}_{n_3}\otimes\boldsymbol{I}_1)\,\mathrm{bcirc}(\boldsymbol{\mathcal{C}}_{i::})\left(\boldsymbol{F}_{n_3}^{-1}\otimes\boldsymbol{I}_r\right), \boldsymbol{w}\boldsymbol{z}^T\rangle \\
&= \max_{\boldsymbol{w}\in\boldsymbol{\Theta}^1,\boldsymbol{z}\in\boldsymbol{\Theta}^r} \left\langle \sum_{j=1}^{n}\left(\boldsymbol{Z}_{ij}^T\boldsymbol{Z}_{ij}\boldsymbol{\Delta}_{ij:} - m\boldsymbol{\Delta}_{ij:}\right) * \boldsymbol{\mathcal{F}}(j,:,:), \underbrace{\mathrm{bcirc}^*\left(\left(\boldsymbol{F}_{n_3}^{-1}\otimes\boldsymbol{I}_1\right)\boldsymbol{w}\boldsymbol{z}^T\left(\boldsymbol{F}_{n_3}\otimes\boldsymbol{I}_r\right)\right)}_{\boldsymbol{\mathcal{T}}\in\mathbb{R}^{1\times r\times n_3}} \right\rangle \\
&= \max_{\boldsymbol{w}\in\boldsymbol{\Theta}^1,\boldsymbol{z}\in\boldsymbol{\Theta}^r} \sum_{j=1}^{n}\left\langle \boldsymbol{Z}_{ij}\boldsymbol{\Delta}_{ij:}, \boldsymbol{Z}_{ij}\left(\boldsymbol{\mathcal{T}} * (\boldsymbol{\mathcal{F}}(j,:,:))^T\right)\right\rangle - m\left\langle\boldsymbol{\Delta}_{ij:}, \boldsymbol{\mathcal{T}} * (\boldsymbol{\mathcal{F}}(j,:,:))^T\right\rangle \\
&= \max_{\boldsymbol{w}\in\boldsymbol{\Theta}^1,\boldsymbol{z}\in\boldsymbol{\Theta}^r} \sum_{j=1}^{n}\sum_{s=1}^{m}\left(\underbrace{\boldsymbol{z}_{ijs}^T\boldsymbol{\Delta}_{ij:}\boldsymbol{z}_{ijs}^T\left(\boldsymbol{\mathcal{T}} * (\boldsymbol{\mathcal{F}}(j,:,:))^T\right) - \left\langle\boldsymbol{\Delta}_{ij:}, \boldsymbol{\mathcal{T}} * (\boldsymbol{\mathcal{F}}(j,:,:))^T\right\rangle}_{K_{js}}\right).
\end{aligned} \tag{101}
$$

(i) is due to the definition of the tensor spectral norm. (ii) is because the block circulant matrix in (3) can be block diagonalized, i.e.,

$$
(\boldsymbol{F}_{n_3}\otimes\boldsymbol{I}_{n_1})\mathrm{bcirc}(\boldsymbol{\mathcal{X}})(\boldsymbol{F}_{n_3}^{-1}\otimes\boldsymbol{I}_{n_2}) = \overline{\boldsymbol{X}}, \tag{102}
$$

where F_n is the Fourier matrix and $\overline{\boldsymbol{X}}$ is the block diagonal of $\boldsymbol{\mathcal{X}}$ as

$$
\overline{\boldsymbol{X}} = \begin{bmatrix} \overline{\boldsymbol{X}}^{(1)} & & & \\ & \overline{\boldsymbol{X}}^{(2)} & & \\ & & \ddots & \\ & & & \overline{\boldsymbol{X}}^{(n_3)}. \end{bmatrix} \tag{103}
$$

The set $\boldsymbol{\Theta}^k$ is defined as $\boldsymbol{\Theta}^k := \boldsymbol{\mathcal{B}}^k \bigcap \boldsymbol{\mathcal{S}}^k$. The set $\boldsymbol{\mathcal{B}}^k$ denote the block sparse vectors, specially, $\boldsymbol{\mathcal{B}}^k := \left\{\boldsymbol{x} \in \mathbb{R}^{kn_3} | \boldsymbol{x} = \left[\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_i^T, \cdots, \boldsymbol{x}_{n_3}^T\right]\right\}$, where $\boldsymbol{x}_i \in \mathbb{R}^k$ and there exists a j such that $\boldsymbol{x}_j \neq \mathbf{0}$ and $\boldsymbol{x}_i = \mathbf{0}$ for all $i \neq j$. The $\boldsymbol{\mathcal{S}}^k$ is defined as $\boldsymbol{\mathcal{S}}^k := \left\{\boldsymbol{x} \in \mathbb{R}^{kn_3} | \left\|\boldsymbol{x}\right\| = 1\right\}$.

Due to independence of $\{\boldsymbol{z}_{ijs}\}_{j=1,s=1}^{j=n,s=m}, \forall i \in [n]$, the K_{js} are independent sub-exponential scalars with zero means and have bounded sub-exponential norm as $\left\|K_{js}\right\|_{\psi_1} \leq \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \left\|\boldsymbol{\mathcal{T}} * \left(\boldsymbol{\mathcal{F}}(j,:,:)\right)^T\right\|_F$. $\forall t > 0$, there are

$$
\begin{aligned}
\frac{t^2}{\sum_{j=1}^n \sum_{s=1}^m \left\|K_{js}\right\|_{\psi_1}^2} &\geq \frac{t^2}{\sum_{j=1}^n \sum_{s=1}^m \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{T}} * \left(\boldsymbol{\mathcal{F}}(j,:,:)\right)^T\right\|_F^2} \\
&\geq \frac{t^2}{m \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \sum_{j=1}^n \left\|\boldsymbol{\mathcal{T}} * \left(\boldsymbol{\mathcal{F}}(j,:,:)\right)^T\right\|_F^2} \\
&= \frac{t^2}{m \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{T}} * \boldsymbol{\mathcal{F}}^T\right\|_F^2} \\
&\overset{(i)}{\geq} \frac{t^2}{m \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{T}}\right\|_F^2 \left\|\boldsymbol{\mathcal{F}}\right\|^2} \\
&\overset{(ii)}{=} \frac{t^2}{mn_3 \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{F}}\right\|^2},
\end{aligned} \tag{104}
$$

$$
\begin{aligned}
\frac{t}{\max_{js} \left\|K_{js}\right\|_{\psi_1}} &\geq \frac{t}{\max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \max_j \left\|\boldsymbol{\mathcal{T}} * \left(\boldsymbol{\mathcal{F}}(j,:,:)\right)^T\right\|_F} \\
&\overset{(iii)}{\geq} \frac{t}{\max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \left\|\boldsymbol{\mathcal{T}}\right\|_F \max_j \left\|\boldsymbol{\mathcal{F}}_{j::}\right\|} \\
&\overset{(iv)}{=} \frac{t}{\sqrt{n_3} \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \max_j \left\|\boldsymbol{\mathcal{F}}_{j::}\right\|}.
\end{aligned} \tag{105}
$$

(i) and (iii) are due to norm inequality in (93). (ii) and (iv) are because $\left\|\boldsymbol{\mathcal{T}}\right\|_F = \sqrt{n_3}$. Thus, based on the sub-exponential Bernstein's inequality and ϵ-net argument, $\forall i \in [n]$ there is

$$
\mathbb{P}\left(\left\|\boldsymbol{\mathcal{C}}_{i::}\right\| \geq t\right) \leq \exp\left(r \log n_3 - c \min\left\{\frac{t^2}{mn_3 \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{F}}\right\|^2}, \frac{t}{\sqrt{n_3} \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \max_j \left\|\boldsymbol{\mathcal{F}}_{j::}\right\|}\right\}\right). \tag{106}
$$

The $\exp\left(r \log n_3\right)$ is due to the covering number of for union set $\boldsymbol{\Theta}^1 \times \boldsymbol{\Theta}^2$ is $\mathcal{O}\left(\exp\left((r+1)\log n_3\right)\right)$. Based on the independence among $\boldsymbol{\mathcal{C}}_{i::}, \forall i \in [n]$ due to tube-wise sensing operator $\boldsymbol{\mathcal{Z}}$, there is union bound

$$
\begin{aligned}
&\mathbb{P}\left(\sqrt{\sum_{i=1}^n \left\|\boldsymbol{\mathcal{C}}_{i::}\right\|^2} \geq \sqrt{n} t\right) \leq \sum_{i=1}^n \mathbb{P}\left(\left\|\boldsymbol{\mathcal{C}}_{i::}\right\| \geq t\right) \\
&\qquad \leq n \exp\left(r \log n_3\right) \cdot \exp\left(-c \min\left\{\frac{t^2}{mn_3 \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F^2 \left\|\boldsymbol{\mathcal{F}}\right\|^2}, \frac{t}{\sqrt{n_3} \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \max_j \left\|\boldsymbol{\mathcal{F}}_{j::}\right\|}\right\}\right).
\end{aligned} \tag{107}
$$

Thus,

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{C}}\right\| &\leq \sqrt{\sum_{i=1}^n \left\|\boldsymbol{\mathcal{C}}_{i::}\right\|^2} \\
&\leq C\sqrt{n} \cdot \left(\sqrt{mn_3 r \log(n \vee n_3)} \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \left\|\boldsymbol{\mathcal{F}}\right\| + \sqrt{n_3} r \log(n \vee n_3) \max_{ij} \left\|\boldsymbol{\Delta}_{ij:}\right\|_F \max_j \left\|\boldsymbol{\mathcal{F}}_{j::}\right\|\right)
\end{aligned}
$$

$$\leq 2Cmr\sqrt{\frac{nn_3}{m}}\log(n\vee n_3)\max_{ij}\left\|\boldsymbol{\Delta}_{ij:}\right\|_F\left\|\boldsymbol{\mathcal{F}}\right\| \tag{108}$$

holds with probability at least $1-\frac{1}{(n\vee n_3)^{10}}$. The first and last inequalities are due to the tensor norm inequality (92). □

Besides the above concentration inequality in the tensor spectral norm metric, we also need the following concentration inequality in the inner product metric, which would be used in proving the restricted strong convexity and smooth property of the non-convex loss landscape in (6).

Lemma 3. *For the tube-wise sensing map $\boldsymbol{\mathcal{Z}}:\mathbb{R}^{n\times n\times n_3}\rightarrow\mathbb{R}^{n\times n\times m}$ in Definition 1, $\forall\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{D}}\in\mathbb{R}^{n\times r\times n_3}$ and the defined concentration distance*

$$D\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T,\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right):=\left\langle\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T\right),\boldsymbol{\mathcal{Z}}\left(\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right)\right\rangle-m\left\langle\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T,\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right\rangle, \tag{109}$$

there is

$$\begin{aligned}\left|D\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T,\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right)\right|&\leq\sqrt{m}\log(n\vee n_3)\sqrt{\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F^2\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2}\\&\leq\sqrt{mn_3}\log(n\vee n_3)\left(\left(\left\|\boldsymbol{\mathcal{A}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{C}}_{i::}\right\|\right)\wedge\left(\left\|\boldsymbol{\mathcal{C}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{C}}_{i::}\right\|\right)\right)\left\|\boldsymbol{\mathcal{B}}\right\|_F\left\|\boldsymbol{\mathcal{D}}\right\|_F\end{aligned} \tag{110}$$

with probability at least $1-\frac{1}{(n\vee n_3)^{10}}$.

Proof.

$$\begin{aligned}D\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T,\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right)&=\sum_{i=1}^n\sum_{j=1}^n\left\langle\boldsymbol{Z}_{ij}\left(\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right),\boldsymbol{Z}_{ij}\left(\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right)\right\rangle\\&\quad-m\left\langle\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T,\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\rangle\\&=\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\left\{\underbrace{\boldsymbol{z}_{ijs}^T\left(\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right)\boldsymbol{z}_{ijs}^T\left(\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right)-\left\langle\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T,\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\rangle}_{g_{ijs}}\right\}.\end{aligned} \tag{111}$$

It is apparent that g_{ijs} are independent sub-exponential variables with zero means and bounded sub-exponential norm $\left\|g_{ijs}\right\|_{\psi_1}\leq\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F$. Thus, there is

$$\begin{aligned}\frac{t^2}{\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\left\|g_{ijs}\right\|_{\psi_1}^2}&\geq\frac{t^2}{m\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F^2\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2},\\\frac{t}{\max_{ijs}\left\|g_{ijs}\right\|_{\psi_1}}&\geq\frac{t}{\max_{ij}\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F}.\end{aligned} \tag{112}$$

Based on the sub-exponential Bernstein inequality, there is

$$\begin{aligned}\left|D\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^T,\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^T\right)\right|&\leq\frac{\sqrt{m\log(n\vee n_3)}}{2}\sqrt{\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F^2\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2}\\&\quad+\frac{\log(n\vee n_3)}{2}\max_{ij}\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F\\&\leq\sqrt{m}\log(n\vee n_3)\sqrt{\sum_{i=1}^n\sum_{j=1}^n\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F^2\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2}\end{aligned} \tag{113}$$

holds with probability at least $1-\frac{1}{(n\vee n_3)^{10}}$. Further, we have

$$\begin{aligned}
&\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F^2\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2}\leq\max_{ij}\left\|\boldsymbol{\mathcal{A}}_{i::}*\left(\boldsymbol{\mathcal{C}}_{j::}\right)^T\right\|_F\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{\mathcal{B}}_{i::}*\left(\boldsymbol{\mathcal{D}}_{j::}\right)^T\right\|_F^2}\\
&\quad\leq\left(\left(\|\boldsymbol{\mathcal{A}}\|_F\max_i\|\boldsymbol{\mathcal{C}}_{i::}\|\right)\wedge\left(\|\boldsymbol{\mathcal{C}}\|_F\max_i\|\boldsymbol{\mathcal{A}}_{i::}\|\right)\right)\sqrt{n_3\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{\mathcal{B}}_{i::}\|_F^2\|\boldsymbol{\mathcal{D}}_{j::}\|_F^2}\\
&\quad=\sqrt{n_3}\left(\left(\|\boldsymbol{\mathcal{A}}\|_F\max_i\|\boldsymbol{\mathcal{C}}_{i::}\|\right)\wedge\left(\|\boldsymbol{\mathcal{C}}\|_F\max_i\|\boldsymbol{\mathcal{A}}_{i::}\|\right)\right)\|\boldsymbol{\mathcal{B}}\|_F\|\boldsymbol{\mathcal{D}}\|_F, \qquad (114)
\end{aligned}$$

where the inequality is due to norm inequalities (93). □

The following lemma gives the property of alignment tensor $\boldsymbol{\mathcal{R}}^t$ in (11), which shares a similar result in the matrix alignment setting [32].

Lemma 4. *For the minimization problem* (11)*, the*

$$\left(\boldsymbol{\mathcal{A}}^t*\boldsymbol{\mathcal{R}}^t\right)^T*\boldsymbol{\mathcal{A}}^\star+\left(\boldsymbol{\mathcal{B}}^t*\boldsymbol{\mathcal{R}}^t\right)^T*\boldsymbol{\mathcal{B}}^\star \qquad (115)$$

is the symmetric positive semidefinite tensor and $\left(\boldsymbol{\mathcal{A}}^t*\boldsymbol{\mathcal{R}}^t-\boldsymbol{\mathcal{A}}^\star\right)^T*\boldsymbol{\mathcal{A}}^\star+\left(\boldsymbol{\mathcal{B}}^t*\boldsymbol{\mathcal{R}}^t-\boldsymbol{\mathcal{B}}^\star\right)^T*\boldsymbol{\mathcal{B}}^\star\in\mathbb{R}^{r\times r\times n_3}$ *is also symmetric tensor.*

Proof. Solving (11) is equivalent to solving the following problem in the frequency domain as

$$\overline{\boldsymbol{R}^t}:=\arg\min_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r}\left\|\begin{bmatrix}\begin{bmatrix}\overline{\boldsymbol{A}^t}^{(1)}\\ \overline{\boldsymbol{B}^t}^{(1)}\end{bmatrix}\overline{\boldsymbol{R}^t}^{(1)}-\begin{bmatrix}\overline{\boldsymbol{A}^\star}^{(1)}\\ \overline{\boldsymbol{B}^\star}^{(1)}\end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix}\overline{\boldsymbol{A}^t}^{(n_3)}\\ \overline{\boldsymbol{B}^t}^{(n_3)}\end{bmatrix}\overline{\boldsymbol{R}^t}^{(n_3)}-\begin{bmatrix}\overline{\boldsymbol{A}^\star}^{(n_3)}\\ \overline{\boldsymbol{B}^\star}^{(n_3)}\end{bmatrix}\end{bmatrix}\right\|_F \qquad (116)$$

Due to the block diagonal structure, to find the orthogonal alignment tensor. We just need to find each orthogonal alignment matrix for each frontal slice in the Fourier domain as

$$\overline{\boldsymbol{R}^t}^{(k)}:=\arg\min_{\boldsymbol{R}\in\boldsymbol{O}_r}\left\|\begin{bmatrix}\overline{\boldsymbol{A}^t}^{(k)}\\ \overline{\boldsymbol{B}^t}^{(k)}\end{bmatrix}\boldsymbol{R}-\begin{bmatrix}\overline{\boldsymbol{A}^\star}^{(k)}\\ \overline{\boldsymbol{B}^\star}^{(k)}\end{bmatrix}\right\|_F,\quad\forall k\in[n_3]. \qquad (117)$$

$\forall k\in[n_3]$, the (117) is the matrix orthogonal Procrustes problem [14] which concludes that the each optimal alignment matrix as

$$\begin{aligned}
&\overline{\boldsymbol{R}^t}^{(k)}=\boldsymbol{L}_t^{(k)}\left(\boldsymbol{P}_t^{(k)}\right)^T,\\
&\left(\overline{\boldsymbol{A}^t}^{(k)}\right)^T\overline{\boldsymbol{A}^\star}^{(k)}+\left(\overline{\boldsymbol{B}^t}^{(k)}\right)^T\overline{\boldsymbol{B}^\star}^{(k)}=\overline{\boldsymbol{L}_t}^{(k)}\overline{\boldsymbol{S}_t}^{(k)}\left(\overline{\boldsymbol{P}_t}^{(k)}\right)^T, \qquad (118)
\end{aligned}$$

where the second equation denotes the SVD decomposition of LHS. Also there is

$$\begin{aligned}
\left(\overline{\boldsymbol{A}^\star}^{(k)}\right)^T\overline{\boldsymbol{A}^t}^{(k)}\overline{\boldsymbol{R}^t}^{(k)}+\left(\overline{\boldsymbol{B}^\star}^{(k)}\right)^T\overline{\boldsymbol{B}^t}^{(k)}\overline{\boldsymbol{R}^t}^{(k)}&=\left(\overline{\boldsymbol{A}^t}^{(k)}\overline{\boldsymbol{R}^t}^{(k)}\right)^T\overline{\boldsymbol{A}^\star}^{(k)}+\left(\overline{\boldsymbol{B}^t}^{(k)}\overline{\boldsymbol{R}^t}^{(k)}\right)^T\overline{\boldsymbol{B}^\star}^{(k)}\\
&=\overline{\boldsymbol{P}_t}^{(k)}\overline{\boldsymbol{S}_t}^{(k)}\left(\overline{\boldsymbol{P}_t}^{(k)}\right)^T. \qquad (119)
\end{aligned}$$

Translating the above equation into the original domain, we obtain

$$\left(\boldsymbol{\mathcal{A}}^\star\right)^T*\left(\boldsymbol{\mathcal{A}}^t*\boldsymbol{\mathcal{R}}^t\right)+\left(\boldsymbol{\mathcal{B}}^\star\right)^T*\left(\boldsymbol{\mathcal{B}}^t*\boldsymbol{\mathcal{R}}^t\right)=\left(\boldsymbol{\mathcal{A}}^t*\boldsymbol{\mathcal{R}}^t\right)^T*\boldsymbol{\mathcal{A}}^\star+\left(\boldsymbol{\mathcal{B}}^t*\boldsymbol{\mathcal{R}}^t\right)^T*\boldsymbol{\mathcal{B}}^\star$$

$$= \boldsymbol{\mathcal{P}}_t * \boldsymbol{\mathcal{S}}_t * \boldsymbol{\mathcal{P}}_t^T, \tag{120}$$

which is the symmetric positive semidefinite tensor.

$$\begin{aligned}\left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{A}}^\star\right)^T * \boldsymbol{\mathcal{A}}^\star + \left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t - \boldsymbol{\mathcal{B}}^\star\right)^T * \boldsymbol{\mathcal{B}}^\star &= \left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t\right)^T * \boldsymbol{\mathcal{A}}^\star + \left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t\right)^T * \boldsymbol{\mathcal{B}}^\star \\ &\quad - \left((\boldsymbol{\mathcal{A}}^\star)^T * \boldsymbol{\mathcal{A}}^\star + (\boldsymbol{\mathcal{B}}^\star)^T * \boldsymbol{\mathcal{B}}^\star\right), \end{aligned} \tag{121}$$

where the RHS of the above equation is the symmetric tensor due to (120). The general result for tensor is similar as matrix that $\boldsymbol{\mathcal{R}}^t$ is the minimizer of (11) iff that $\left(\boldsymbol{\mathcal{A}}^t * \boldsymbol{\mathcal{R}}^t\right)^T * \boldsymbol{\mathcal{A}}^\star + \left(\boldsymbol{\mathcal{B}}^t * \boldsymbol{\mathcal{R}}^t\right)^T * \boldsymbol{\mathcal{B}}^\star$ is symmetric and positive semidefinite. □

The next lemma gives the perturbation bound in the tensor spectral norm under the t-SVD algebra.

Lemma 5. *Let $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r \times r \times n_3}$ be a nonsingular tensor, i.e., $\sigma_r(\overline{\boldsymbol{C}}^{(i)}) > 0, \forall i \in [n_3]$. Then for any tensor $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{r \times r \times n_3}$ with $\|\boldsymbol{\mathcal{E}}\| \leq \sigma_{\min}(\boldsymbol{\mathcal{C}})$, there is*

$$\|sgn(\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{E}}) - sgn(\boldsymbol{\mathcal{C}})\| \leq \frac{\|\boldsymbol{\mathcal{E}}\|}{\sigma_{\min}(\boldsymbol{\mathcal{C}}) - \|\boldsymbol{\mathcal{E}}\|}, \tag{122}$$

where $sgn(\boldsymbol{\mathcal{X}}) := \boldsymbol{\mathcal{X}} * (\boldsymbol{\mathcal{X}}^T * \boldsymbol{\mathcal{X}})^{-\frac{1}{2}}$.

Proof.

$$\begin{aligned}\|\mathrm{sgn}(\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{E}}) - \mathrm{sgn}(\boldsymbol{\mathcal{C}})\| &= \left\| \overline{\boldsymbol{C}} \left(\overline{\boldsymbol{C}}^c \overline{\boldsymbol{C}}\right)^{-\frac{1}{2}} - (\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}}) \left((\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}})^c (\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}})\right)^{-\frac{1}{2}} \right\| \\ &\overset{(i)}{\leq} -\log\left(1 - \frac{\|\overline{\boldsymbol{E}}\|}{\sigma_{\min}(\overline{\boldsymbol{C}})}\right) \\ &\overset{(ii)}{\leq} \frac{\frac{\|\overline{\boldsymbol{E}}\|}{\sigma_{\min}(\overline{\boldsymbol{C}})}}{1 - \frac{\|\overline{\boldsymbol{E}}\|}{\sigma_{\min}(\overline{\boldsymbol{C}})}} \\ &= \frac{\|\boldsymbol{\mathcal{E}}\|}{\sigma_{\min}(\boldsymbol{\mathcal{C}}) - \|\boldsymbol{\mathcal{E}}\|}, \end{aligned} \tag{123}$$

where (i) is based on Theorem 3.2 in [35], which is a perturbation bound for the complex matrices. (ii) is based on logarithm inequality $\log(1+x) \geq \frac{x}{1+x}, \forall x > -1$. □

To bound the distance between two tensors after transformation based on the corresponding orthogonal alignment tensor, we need the following lemma, which is a generalized version of the matrix setting in [32] and based on Lemma 5.

Lemma 6. *Assume the tensors $\boldsymbol{\mathcal{X}}_0, \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \in \mathbb{R}^{n \times r \times n_3}$ satisfying*

$$\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_0\| \|\boldsymbol{\mathcal{X}}_0\| \leq \frac{\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_0)}{2}, \quad \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| \|\boldsymbol{\mathcal{X}}_0\| \leq \frac{\sigma_{\min}^2(\boldsymbol{\mathcal{X}}_0)}{4}. \tag{124}$$

Denote the alignment tensors as

$$\begin{aligned}\boldsymbol{\mathcal{R}}_1 &:= \underset{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r}{\arg\min} \|\boldsymbol{\mathcal{X}}_1 * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{X}}_0\|_F \\ \boldsymbol{\mathcal{R}}_2 &:= \underset{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r}{\arg\min} \|\boldsymbol{\mathcal{X}}_2 * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{X}}_0\|_F . \end{aligned} \tag{125}$$

Then there is

$$\|\boldsymbol{\mathcal{X}}_1 * \boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{X}}_2 * \boldsymbol{\mathcal{R}}_2\| \leq 9 \left(\frac{\sigma_{\max}(\boldsymbol{\mathcal{X}}_0)}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_0)}\right)^2 \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| . \tag{126}$$

This can be proved based on the proof for the matrix setting in [32].

$$\|\boldsymbol{\mathcal{X}}_1 * \boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{X}}_2 * \boldsymbol{\mathcal{R}}_2\| = \|(\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2) * \boldsymbol{\mathcal{R}}_2 + \boldsymbol{\mathcal{X}}_1 * (\boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{R}}_2)\|$$

$$
\begin{aligned}
&\le \left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| + \left\|\boldsymbol{\mathcal{X}}_1\right\| \left\|\boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{R}}_2\right\| \\
&\le \left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| + 2\left\|\boldsymbol{\mathcal{X}}_0\right\| \left\|\boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{R}}_2\right\|, \qquad (127)
\end{aligned}
$$

where the last inequality is due to the second condition in (124). Now we come to bound the distance between two alignment tensors. Based on (125), $\forall k \in [n_3]$, there is

$$
\overline{\boldsymbol{R}}_1^{(k)} = \mathop{\arg\min}_{\boldsymbol{R}\in\boldsymbol{O}_r} \left\|\overline{\boldsymbol{X}_1}^{(k)}\boldsymbol{R} - \overline{\boldsymbol{X}_0}^{(k)}\right\|_F \qquad (128)
$$

$$
\overline{\boldsymbol{R}}_2^{(k)} = \mathop{\arg\min}_{\boldsymbol{R}\in\boldsymbol{O}_r} \left\|\overline{\boldsymbol{X}_2}^{(k)}\boldsymbol{R} - \overline{\boldsymbol{X}_0}^{(k)}\right\|_F. \qquad (129)
$$

Denote the $\boldsymbol{\mathcal{C}} := \boldsymbol{\mathcal{X}}_1^T * \boldsymbol{\mathcal{X}}_0$ and $\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{E}} := \boldsymbol{\mathcal{X}}_2^T * \boldsymbol{\mathcal{X}}_0$, there is

$$
\left\|\overline{\boldsymbol{C}} - \overline{\boldsymbol{X}_0^T\boldsymbol{X}_0}\right\| = \left\|\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{X}}_0^T * \boldsymbol{\mathcal{X}}_0\right\| \le \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{2}. \qquad (130)
$$

Based on Wely's inequality, we can know $\sigma_{\min}(\overline{\boldsymbol{C}}) \ge \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{2}$ which means that all $\overline{\boldsymbol{C}}^{(k))}$ are invertible. Thus, $\boldsymbol{\mathcal{C}}$ is an invertible tensor. In addition,

$$
\left\|\overline{\boldsymbol{E}}\right\| = \left\|\left(\overline{\boldsymbol{X}_2}^T - \overline{\boldsymbol{X}_1}\right)^T\overline{\boldsymbol{X}_0}\right\| \le \left\|\overline{\boldsymbol{X}_2} - \overline{\boldsymbol{X}_1}\right\|\left\|\overline{\boldsymbol{X}_0}\right\| = \left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\|\left\|\boldsymbol{\mathcal{X}}_0\right\| \le \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{4}. \qquad (131)
$$

Thus, $\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}}$ is invertible and $\overline{\boldsymbol{R}_1}^{(k)}, \overline{\boldsymbol{R}_2}^{(k)}$ have closed formulation as

$$
\overline{\boldsymbol{R}_1}^{(k)} = \overline{\boldsymbol{C}}^{(k)}\left(\left(\overline{\boldsymbol{C}}^{(k)}\right)^T\overline{\boldsymbol{C}}^{(k)}\right)^{-\frac{1}{2}}, \quad \overline{\boldsymbol{R}_2}^{(k)} = \left(\overline{\boldsymbol{C}}^{(k)} + \overline{\boldsymbol{E}}^{(k)}\right)\left(\left(\overline{\boldsymbol{C}}^{(k)} + \overline{\boldsymbol{E}}^{(k)}\right)^T\left(\overline{\boldsymbol{C}}^{(k)} + \overline{\boldsymbol{E}}^{(k)}\right)\right)^{-\frac{1}{2}} \qquad (132)
$$

Therefore, there is

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{R}}_2\right\| &= \left\|\overline{\boldsymbol{R}_1} - \overline{\boldsymbol{R}_2}\right\| \\
&= \left\|\overline{\boldsymbol{C}}\left(\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}\right)^{-\frac{1}{2}} - \left(\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}}\right)\left(\left(\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}}\right)^T\left(\overline{\boldsymbol{C}} + \overline{\boldsymbol{E}}\right)\right)^{-\frac{1}{2}}\right\| \\
&\overset{(i)}{\le} \frac{\left\|\boldsymbol{\mathcal{E}}\right\|}{\sigma_{\min}(\boldsymbol{\mathcal{C}}) - \left\|\boldsymbol{\mathcal{E}}\right\|} \\
&\le \frac{\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\|\left\|\boldsymbol{\mathcal{X}}_0\right\|}{\frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{2} - \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{4}}, \qquad (133)
\end{aligned}
$$

where (i) is based on Lemma 5 and $\left\|\overline{\boldsymbol{E}}\right\| \le \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{4} < \frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{2} \le \sigma_{\min}(\overline{\boldsymbol{C}})$, which also induce the last inequality with (131). Substituting above inequality into (127) would have

$$
\begin{aligned}
\left\|\boldsymbol{\mathcal{X}}_1 * \boldsymbol{\mathcal{R}}_1 - \boldsymbol{\mathcal{X}}_2 * \boldsymbol{\mathcal{R}}_2\right\| &\le \left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| + \frac{2\left\|\boldsymbol{\mathcal{X}}_0\right\|^2}{\frac{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}{4}}\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| \\
&\le \left(1 + 8\frac{\sigma^2_{\max}(\boldsymbol{\mathcal{X}}_0)}{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}\right)\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\| \\
&\le 9\frac{\sigma^2_{\max}(\boldsymbol{\mathcal{X}}_0)}{\sigma^2_{\min}(\boldsymbol{\mathcal{X}}_0)}\left\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\right\|. \qquad (134)
\end{aligned}
$$

The following lemma provides the benign landscape of the non-convex problem (6), which has restricted strong convexity and smoothness properties in incoherent regions near the global optimum $\boldsymbol{\mathcal{X}}^\star$. The sample complexity in (12) is sufficient to guarantee the existence of this benign landscape, which can obtain the linear rate.

Lemma 7. *With the same setting as Theorem 2, if the tube-wise sample complexity satisfies* (12)*, then*

$$
\mathit{Vec}\left(\begin{bmatrix}\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\end{bmatrix}\right)^T \mathit{Mat}\left(\nabla^2 f(\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}})\right)\mathit{Vec}\left(\begin{bmatrix}\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\end{bmatrix}\right) \ge \frac{\sigma_{\min}}{5}\left\|\begin{bmatrix}\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\end{bmatrix}\right\|_F^2 \quad \textit{and} \quad \left\|\nabla^2 f(\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}})\right\| \le 5\sigma_{\max} \qquad (135)
$$

hold with probability at least $1-\frac{1}{(n\vee n_3)^{10}}$ *for* $\forall \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n\times r\times n_3}$ *that satisfy*

$$\max_{l\in[2n]} \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right)_{l::} \right\| \leq \frac{\sqrt{\sigma_{\max}}}{500\kappa\sqrt{2n}} \tag{136}$$

and $\forall \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{n\times r\times n_3}$ *stay in set*

$$\left\{ \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 \\ \boldsymbol{\mathcal{B}}_1 \end{bmatrix} * \hat{\boldsymbol{\mathcal{R}}} - \begin{bmatrix} \boldsymbol{\mathcal{A}}_2 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix} : \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}_2 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right\| \leq \frac{\sqrt{\sigma_{\max}}}{500\kappa}, \hat{\boldsymbol{\mathcal{R}}} := \arg\min_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r} \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}_1 \\ \boldsymbol{\mathcal{B}}_1 \end{bmatrix} * \boldsymbol{\mathcal{R}} - \begin{bmatrix} \boldsymbol{\mathcal{A}}_2 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix} \right\|_F \right\}. \tag{137}$$

Proof. Based on (6), there is definition $\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}} := \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{U}}^{\star}, \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}} := \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{V}}^{\star}$ and population level has expression as

$$\begin{aligned} \mathbb{E}\left[\mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right)^T \mathrm{Mat}\left(\nabla^2 f(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})\right) \mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right) \right] &= m\Bigg(\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{V}}^{\star}\|_F^2 + \left\|\boldsymbol{\mathcal{U}}^{\star} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T\right\|_F^2 \\ &+ \frac{1}{2}\left\| \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}^T * \boldsymbol{\mathcal{U}}^{\star} - \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T * \boldsymbol{\mathcal{V}}^{\star} \right\|_F^2 + \frac{1}{2}\left\| \boldsymbol{\mathcal{A}}_2^T * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} + \boldsymbol{\mathcal{B}}_2^T * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \right\|_F^2 + \mathcal{W}_1 + \mathcal{W}_2 \Bigg), \end{aligned} \tag{138}$$

where the $\mathcal{W}_1 + \mathcal{W}_2$ has following bound

$$\begin{aligned} |\mathcal{W}_1 + \mathcal{W}_2| \leq 9 \Big[& (\|\boldsymbol{\mathcal{U}}^{\star} - \boldsymbol{\mathcal{A}}_2\| + \|\boldsymbol{\mathcal{V}}^{\star} - \boldsymbol{\mathcal{B}}_2\|) \cdot (\|\boldsymbol{\mathcal{A}}_2\| + \|\boldsymbol{\mathcal{B}}_2\|) + (\|\boldsymbol{\mathcal{U}}^{\star} - \boldsymbol{\mathcal{A}}_2\| + \|\boldsymbol{\mathcal{V}}^{\star} - \boldsymbol{\mathcal{B}}_2\|)^2 \\ & + (\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\| + \|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}\|) \cdot (\|\boldsymbol{\mathcal{U}}^{\star}\| + \|\boldsymbol{\mathcal{V}}^{\star}\|) + (\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\| + \|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}\|)^2 \Big] \cdot \left(\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\|_F^2 + \|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\|_F^2 \right). \end{aligned} \tag{139}$$

Due to spectral norm inequality (92), there is

$$\begin{aligned} \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right\| &\leq \sqrt{ \sum_{i=1}^{2n} \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right)_{i::} \right\|^2 } \\ &\leq \max_{l\in[2n]} \sqrt{2n} \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{B}} - \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right)_{l::} \right\| \\ &\leq \frac{\sqrt{\sigma_{\max}}}{500\kappa}. \end{aligned} \tag{140}$$

Thus, based on $\left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}_2 \\ \boldsymbol{\mathcal{B}}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} \right\| \leq \frac{\sqrt{\sigma_{\max}}}{500\kappa}$ and above inequality, there is

$$|\mathcal{W}_1 + \mathcal{W}_2| \leq \frac{\sigma_{\min}}{5} \left\| \begin{bmatrix} \boldsymbol{\mathcal{D}}_{\mathcal{A}} \\ \boldsymbol{\mathcal{D}}_{\mathcal{B}} \end{bmatrix} \right\|_F^2. \tag{141}$$

Next, we need to bound the difference between the population level and the empirical level by the following decomposition as

$$\begin{aligned} & \mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right)^T \mathrm{Mat}\left(\nabla^2 f(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})\right) \mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right) - \mathbb{E}\left[\mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right)^T \mathrm{Mat}\left(\nabla^2 f(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}})\right) \mathrm{Vec}\left(\begin{bmatrix} \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} \\ \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}} \end{bmatrix} \right) \right] \\ & \quad = 2\left\langle \boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^T - \boldsymbol{\mathcal{U}}^{\star} * (\boldsymbol{\mathcal{V}}^{\star})^T \right), \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right\rangle - 2\left\langle \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}^T - \boldsymbol{\mathcal{U}}^{\star} * (\boldsymbol{\mathcal{V}}^{\star})^T, \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right\rangle \\ & \qquad + \left\| \boldsymbol{\mathcal{Z}} \left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{B}}^T + \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right) \right\|_F^2 - \left\| \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{B}}^T + \boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right\|_F^2 \\ & = \Pi, \end{aligned} \tag{142}$$

where Π has decomposition as

$$\Pi = 2D\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}} * (\boldsymbol{\mathcal{V}}^{\star})^T, \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right) + 2D\left(\boldsymbol{\mathcal{U}}^{\star} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^T, \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right) + 2D\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^T, \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}} * \boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^T \right)$$

$$
\begin{aligned}
&+2D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\left(\boldsymbol{\mathcal{V}}^{\star}\right)^{T},\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right)+2D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\mathcal{U}}^{\star}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right)+2D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right);\\
&+D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\left(\boldsymbol{\mathcal{V}}^{\star}\right)^{T},\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\left(\boldsymbol{\mathcal{V}}^{\star}\right)^{T}\right)+D\left(\boldsymbol{\mathcal{U}}^{\star}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\mathcal{U}}^{\star}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right)+2D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\left(\boldsymbol{\mathcal{V}}^{\star}\right)^{T},\boldsymbol{\mathcal{U}}^{\star}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right);\\
&+D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{T}\right)+D\left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right)+2D\left(\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\left(\boldsymbol{\mathcal{V}}^{\star}\right)^{T},\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{T}\right)\\
&+2D\left(\boldsymbol{\mathcal{U}}^{\star}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T},\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}*\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}^{T}\right).
\end{aligned}
\tag{143}
$$

The concentration distance $D\left(\boldsymbol{\mathcal{A}}*\boldsymbol{\mathcal{C}}^{T},\boldsymbol{\mathcal{B}}*\boldsymbol{\mathcal{D}}^{T}\right)$ is defined in Lemma 3. Based on the Lemma 3, we can bound the conccentration level Π as follows

$$
\begin{aligned}
|\Pi|\leq{}& m\sqrt{\frac{n_3}{m}}\log(n\vee n_3)\cdot\Big(2\left\|\boldsymbol{\mathcal{V}}^{\star}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\mathcal{A}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F+2\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\mathcal{A}}\right\|_F\\
&\cdot\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F+2\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\mathcal{A}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F+2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{V}}^{\star}_{i::}\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F
\end{aligned}
\tag{144}
$$

$$
\begin{aligned}
&+2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F+2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}})_{i::}\right\|\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\mathcal{B}}\right\|_F\\
&+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{V}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{V}}^{\star}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F+\max_i\left\|\boldsymbol{\mathcal{U}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\\
&+2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{V}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}\right\|_F\\
&+\max_i\left\|(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}})_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F+2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\max_i\left\|\boldsymbol{\mathcal{V}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}\right\|_F\\
&+2\max_i\left\|\boldsymbol{\mathcal{U}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\left\|\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\Big)\\
\overset{(i)}{\leq}{}& m\sqrt{\frac{n_3}{m}}\log(n\vee n_3)\cdot\Big(3\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\max_l\left\|\boldsymbol{\Delta}_{l::}\right\|+3\max_i\left\|\boldsymbol{\mathcal{U}}^{\star}_{i::}\right\|\left\|\boldsymbol{\Delta}\right\|_F+3\left\|\boldsymbol{\Delta}\right\|_F\max_l\left\|\boldsymbol{\Delta}_{l::}\right\|\\
&+2\max_i\left\|\boldsymbol{\mathcal{U}}^{\star}_{i::}\right\|\left\|\boldsymbol{\mathcal{U}}^{\star}\right\|_F\Big)\cdot\left(\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F^2+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F^2\right)\\
\overset{(ii)}{\leq}{}& m\sqrt{\frac{n_3}{m}}\log(n\vee n_3)\cdot\Bigg(3\sqrt{n}\sqrt{\frac{\mu r\sigma_{\max}}{n}}\cdot\frac{\sqrt{\sigma_{\max}}}{500\kappa\sqrt{2n}}+3\sqrt{\frac{\mu r\sigma_{\max}}{n}}\cdot\frac{\sqrt{\sigma_{\max}}}{500\kappa}+3\frac{\sqrt{\sigma_{\max}}}{500\kappa}\\
&\cdot\frac{\sqrt{\sigma_{\max}}}{500\kappa\sqrt{2n}}+2\sqrt{\frac{\mu r\sigma_{\max}}{n}}\cdot\sqrt{n}\sqrt{\frac{\mu r\sigma_{\max}}{n}}\Bigg)\cdot\left(\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F^2+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F^2\right)\\
\leq{}&\frac{m\sigma_{\min}}{5}\left(\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F^2+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F^2\right),
\end{aligned}
\tag{145}
$$

where (i) is due to inequality $2\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F\leq\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{A}}}\right\|_F^2+\left\|\boldsymbol{\mathcal{D}}_{\boldsymbol{\mathcal{B}}}\right\|_F^2$. (ii) is due to the tensor incoherence condition (10), (136), (140) and tensor norm inequality $\left\|\boldsymbol{\mathcal{X}}\right\|_F=\sqrt{\sum_{i=1}^n\left\|\boldsymbol{\mathcal{X}}_{i::}\right\|_F^2}\leq\sqrt{\sum_{i=1}^n\left\|\boldsymbol{\mathcal{X}}_{i::}\right\|^2}\leq\sqrt{n}\max_i\left\|\boldsymbol{\mathcal{X}}_{i::}\right\|$. The last inequality is due to the tube-wise sample complexity in (12). □

Lemma 8 claims that the distance between two consecutive alignment tensors can be well bounded, which would be useful in proving (22).

Lemma 8. *With the same setting as the main Theorem 2, then*

$$
\left\|\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1}*\boldsymbol{\mathcal{R}}^{t+1,l}-\boldsymbol{\mathcal{I}}_r\right\|\leq 2\times 1250^2 m\eta C^2\rho^{2t}\left(\sqrt{\frac{(\mu r)^4\kappa^8 n_3\log^2(n\vee n_3)\sigma_{\max}}{nm}}\right)^2
\tag{146}
$$

holds with probability at least $1-\frac{1}{5(n\vee n_3)^{10}}$.

Proof. This proof follows the proof of Claim 11 in [8] and the proof of Lemma 12 in [32], which makes us use the proof idea for (13) and then utilize the tensor perturbation bound in Lemma 5. Defining the auxiliary iterations as (38)

$$
\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}=\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}*\boldsymbol{\mathcal{R}}^{t,l}-\eta\boldsymbol{\mathcal{Z}}^{c}_{-l,\cdot}\boldsymbol{\mathcal{Z}}_{-l,\cdot}\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)*\boldsymbol{\mathcal{B}}^{\star}-\eta\boldsymbol{\mathcal{P}}_{l,\cdot}\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)*\boldsymbol{\mathcal{B}}^{\star}
$$

$$
\begin{aligned}
&\quad -\frac{m\eta}{2}\boldsymbol{\mathcal{A}}^{\star} * \left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{T} * \left(\left(\boldsymbol{\mathcal{A}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{A}}^{t,l} - \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{B}}^{t,l}\right) * \boldsymbol{\mathcal{R}}^{t,l} \\
\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} &= \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \eta\left(\boldsymbol{\mathcal{Z}}^{c}_{-l,\cdot}\boldsymbol{\mathcal{Z}}_{-l,\cdot}\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} - \eta\left(\boldsymbol{\mathcal{P}}_{l,\cdot}\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} \\
&\quad -\frac{m\eta}{2}\boldsymbol{\mathcal{B}}^{\star} * \left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{T} * \left(\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{B}}^{t,l} - \left(\boldsymbol{\mathcal{A}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{A}}^{t,l}\right) * \boldsymbol{\mathcal{R}}^{t,l}. \qquad (147)
\end{aligned}
$$

Based on definition of $\boldsymbol{\mathcal{R}}^{t,l}$ in (21), we can conclude

$$
\begin{aligned}
\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1} * \boldsymbol{\mathcal{R}}^{t+1,l} &= \arg\min_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r}\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l}\end{bmatrix} * \boldsymbol{\mathcal{R}} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|_F \\
&\overset{(i)}{=} (\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{E}}) * \left((\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{E}})^{T} * (\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{E}})\right)^{\frac{1}{2}} \\
&= \operatorname{sgn}(\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{E}}) \qquad (148)
\end{aligned}
$$

where the (i) is due to Eq.(134) in [32] in frequency domain and (117) and

$$
\boldsymbol{\mathcal{C}} = (\boldsymbol{\mathcal{A}}^{\star})^{T} * \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} + (\boldsymbol{\mathcal{B}}^{\star})^{T} * \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} \qquad (149)
$$

$$
\boldsymbol{\mathcal{E}} = \left(\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right)^{T} * \boldsymbol{\mathcal{A}}^{\star} + \left(\boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right)^{T} * \boldsymbol{\mathcal{B}}^{\star}. \qquad (150)
$$

To use the Lemma 5, we need to prove

$$
\boldsymbol{\mathcal{I}}_r = \arg\min_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r}\left\|\begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix} * \boldsymbol{\mathcal{R}} - \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|_F \qquad (151)
$$

which is equivalent to $\operatorname{sgn}(\boldsymbol{\mathcal{C}}) = \boldsymbol{\mathcal{I}}_r$. Then we can use the Lemma 5 as

$$
\begin{aligned}
\left\|\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1} * \boldsymbol{\mathcal{R}}^{t+1,l} - \boldsymbol{\mathcal{I}}_r\right\| &= \left\|\operatorname{sgn}(\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{E}}) - \operatorname{sgn}(\boldsymbol{\mathcal{C}})\right\| \\
&\le \frac{\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}^{T} * \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|}{\sigma_{\min}\left(\begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}^{T} * \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right) - \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}^{T} * \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|} \\
&\le \frac{\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}\right\|\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|}{\sigma_{\min}\left(\begin{bmatrix}\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}^{T} * \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right) - \left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}^{T} * \begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}\\ \boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|} \\
&\qquad (152)
\end{aligned}
$$

To prove the (151), we should prove that $\boldsymbol{\mathcal{C}}$ is the symmetric positive definite tensor and bound $\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}\right\|$.

We first bound the $\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}\right\|$ as

$$
\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\\ \boldsymbol{\mathcal{B}}^{t+1,l} * \boldsymbol{\mathcal{R}}^{t,l} - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\end{bmatrix}\right\| &= \eta\left\|\begin{bmatrix}\left(-\boldsymbol{\mathcal{Z}}^{c}\boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{Z}}^{c}_{l,\cdot}\boldsymbol{\mathcal{Z}}_{l,\cdot} - m\boldsymbol{\mathcal{P}}_{l,\cdot}\right)\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right) * \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{A}}}\\ \left(\left(-\boldsymbol{\mathcal{Z}}^{c}\boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{Z}}^{c}_{l,\cdot}\boldsymbol{\mathcal{Z}}_{l,\cdot} - m\boldsymbol{\mathcal{P}}_{l,\cdot}\right)\left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)\right)^{T} * \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{B}}}\end{bmatrix}\right\| \\
&\quad + \frac{\eta}{2}\left\|\begin{bmatrix}\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{A}}} * \left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{T} * \left(\left(\boldsymbol{\mathcal{A}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{A}}^{t,l} - \left(\boldsymbol{\mathcal{B}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{B}}^{t,l}\right) * \boldsymbol{\mathcal{R}}^{t,l}\\ \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{B}}} * \left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{T} * \left(\left(\boldsymbol{\mathcal{B}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{B}}^{t,l} - \left(\boldsymbol{\mathcal{A}}^{t,l}\right)^{T} * \boldsymbol{\mathcal{A}}^{t,l}\right) * \boldsymbol{\mathcal{R}}^{t,l}\end{bmatrix}\right\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \eta \underbrace{\left\| \left(-\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{Z}}^c_{l,\cdot} \boldsymbol{\mathcal{Z}}_{l,\cdot} - m \boldsymbol{\mathcal{P}}_{l,\cdot}\right) \left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right) \right\| \left\| \boldsymbol{\Delta}^{t,l} \right\|}_{a_1} + \frac{\eta m}{2} \\
&\quad \cdot \underbrace{\left(\left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{A}}} \right\| + \left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{B}}} \right\| \right) \left(2 \left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{A}}} \right\| \left\| \boldsymbol{\mathcal{A}}^\star \right\| + 2 \left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{B}}} \right\| \left\| \boldsymbol{\mathcal{B}}^\star \right\| + \left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{A}}} \right\|^2 + \left\| \boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{B}}} \right\|^2 \right)}_{a_2} \\
&\leq 2 \times 1250^2 m \eta C^2 \rho^{2t} \sqrt{\sigma_{\max}} \left(\sqrt{\frac{(\mu r)^4 \kappa^7 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \right)^2
\end{aligned}
\tag{153}
$$

where the last two inequalities are due to the bounds for a_1 and a_2 as follows.

For the a_1, there is

$$
\begin{aligned}
a_1 &\leq \left(\left\| \left(\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} - m \boldsymbol{\mathcal{I}}_r\right) \left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right) \right\| + m \left\| \boldsymbol{\mathcal{A}}^{t,l} * (\boldsymbol{\mathcal{B}}^{t,l})^T - \boldsymbol{\mathcal{A}}^\star * (\boldsymbol{\mathcal{B}}^\star)^T \right\| \right. \\
&\quad \left. + \left\| \left(\boldsymbol{\mathcal{Z}}^c_{l,\cdot} \boldsymbol{\mathcal{Z}}_{l,\cdot} - m \boldsymbol{\mathcal{P}}_{l,\cdot}\right) \left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right) \right\| \right) \left\| \boldsymbol{\Delta}^{t,l} \right\| \\
&\overset{(i)}{\leq} \left(2Cmr \sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \max_{ij} \left\| \left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)_{ij:} \right\|_F + 2Cmr \sqrt{\frac{n_3}{m}} \log(n \vee n_3) \max_{j} \left\| \left(\boldsymbol{\Delta}^{t,l}_{\boldsymbol{\mathcal{X}}}\right)_{lj:} \right\|_F \right. \\
&\quad \left. + m \left\| \boldsymbol{\Delta}^{t,l} \right\| \left(\left\| \boldsymbol{\mathcal{B}}^\star \right\| + \left\| \boldsymbol{\mathcal{A}}^\star \right\| + \left\| \boldsymbol{\Delta}^{t,l} \right\| \right) \right) \left\| \boldsymbol{\Delta}^{t,l} \right\| \\
&\overset{(ii)}{\leq} \left(4Cmr \sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \cdot 1300 C \rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n_3 \vee n) \sigma_{\max}}{n^2 m}} \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \right. \\
&\quad \left. + 3m \cdot 12\kappa \cdot 52 C \rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \cdot \sqrt{\sigma_{\max}} \right) \\
&\quad \cdot 12\kappa \cdot 52 C \rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \\
&\leq (1248C)^2 m \rho^{2t} \sqrt{\sigma_{\max}} \left(\sqrt{\frac{(\mu r)^4 \kappa^7 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \right)^2,
\end{aligned}
\tag{154}
$$

where the (i) is due to concentration Lemma 2 and (ii) are because (64) and (79). The last inequality is due to the tube-wise sample complexity in (12).

For a_2, there is

$$
\begin{aligned}
a_2 &\leq 2 \left\| \boldsymbol{\Delta}^{t,l} \right\| \left(4 \left\| \boldsymbol{\Delta}^{t,l} \right\| \left\| \boldsymbol{\mathcal{A}}^\star \right\| + \left\| \boldsymbol{\Delta}^{t,l} \right\|^2 \right) \\
&\overset{(i)}{\leq} 1248 C \rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \left(2496 C \rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \cdot \sqrt{\sigma_{\max}} \right. \\
&\quad \left. + \left(624 C \rho^t \sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \right)^2 \right) \\
&\leq 1248 \times 2500 C^2 \rho^{2t} \sqrt{\sigma_{\max}} \left(\sqrt{\frac{(\mu r)^4 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{nm}} \right)^2.
\end{aligned}
\tag{155}
$$

Next, we prove that $\boldsymbol{\mathcal{C}}$ is a symmetric positive definite tensor. The proof idea is similar to that of Section A.2.3. Based on the definition of (147), there is

$$
\begin{aligned}
\boldsymbol{\mathcal{C}} = &\left(\boldsymbol{\mathcal{A}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{A}}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} + \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{B}}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} - \eta \left(\boldsymbol{\mathcal{A}}^\star\right)^T * \boldsymbol{\mathcal{Z}}_{l,\cdot}^c \boldsymbol{\mathcal{Z}}_{l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{B}}^\star \\
&- \eta \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \boldsymbol{\mathcal{Z}}_{l,\cdot}^c \boldsymbol{\mathcal{Z}}_{l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{A}}^\star - \eta \left(\boldsymbol{\mathcal{A}}^\star\right)^T * \boldsymbol{\mathcal{P}}_{l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{B}}^\star - \eta \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \boldsymbol{\mathcal{P}}_{l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{A}}^\star
\end{aligned}
\tag{156}
$$

Based on definition of $\boldsymbol{\mathcal{R}}^{t,l}$ in (21) and Lemma 4, the $\left(\boldsymbol{\mathcal{A}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{A}}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l} + \left(\boldsymbol{\mathcal{B}}^\star\right)^T * \tilde{\boldsymbol{\mathcal{B}}}^{t,l} * \boldsymbol{\mathcal{R}}^{t,l}$ is symmetric positive definite. Thus, $\boldsymbol{\mathcal{C}}$ is also symmetric. In addition,

$$
\begin{aligned}
\left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix}^T \begin{bmatrix} \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix}^T * \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| &\le \left\| \begin{bmatrix} \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} - \boldsymbol{\mathcal{A}}^\star \\ \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} - \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| \\
&\le \left(\left\| \begin{bmatrix} \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right] - \boldsymbol{\mathcal{A}}^\star \\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right] - \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| + \left\| \begin{bmatrix} \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} \\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} \end{bmatrix} \right\| \right) \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| \\
&\le 0.5\sigma_{\min},
\end{aligned}
\tag{157}
$$

where the last inequality uses the results in (158) and (160) as follows.

Bounding the $\left\| \begin{bmatrix} \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} \\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} \end{bmatrix} \right\|$ is similar to proof in Section A.2.1 as

$$
\begin{aligned}
\left\| \begin{bmatrix} \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{A}}}^{t+1,l} \\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right] - \tilde{\boldsymbol{\mathcal{B}}}^{t+1,l} \end{bmatrix} \right\| &\le \eta \left\| \begin{bmatrix} \boldsymbol{\mathcal{Z}}_{-l,\cdot}^c \boldsymbol{\mathcal{Z}}_{-l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{B}}^\star - m \boldsymbol{\mathcal{P}}_{-l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) * \boldsymbol{\mathcal{B}}^\star \\ \left(\boldsymbol{\mathcal{Z}}_{-l,\cdot}^c \boldsymbol{\mathcal{Z}}_{-l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)\right)^T * \boldsymbol{\mathcal{A}}^\star - m \left(\boldsymbol{\mathcal{P}}_{-l,\cdot} \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)\right)^T * \boldsymbol{\mathcal{B}}^\star \end{bmatrix} \right\| \\
&\overset{(i)}{\le} 2\eta \left\|\boldsymbol{\mathcal{A}}^\star\right\| \left\| \left(\boldsymbol{\mathcal{Z}}^c \boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{I}}_n\right) \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right) \right\| \\
&\overset{(ii)}{\le} 2C\sqrt{\sigma_{\max}} m\eta r \sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \max_{ij} \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{X}}}^{t,l}\right)_{ij:} \right\|_F \\
&\overset{(iii)}{\le} 2C\sqrt{\sigma_{\max}} m\eta r \sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \cdot \Big(\max_i \left\| (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l})_{i::} \right\| \max_i \left\|\boldsymbol{\mathcal{B}}_{i::}^\star\right\| \\
&\qquad + \max_i \left\| (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l})_{i::} \right\| \max_i \left\|\boldsymbol{\mathcal{A}}_{i::}^\star\right\| + \max_i \left\| (\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l})_{i::} \right\| \max_i \left\| \left(\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\right)_{i::} \right\| \Big) \\
&\overset{(iv)}{\le} 2C\sqrt{\sigma_{\max}} m\eta r \sqrt{\frac{nn_3}{m}} \log(n \vee n_3) \cdot \sqrt{\frac{\mu r \sigma_{\max}}{n}} \\
&\qquad \cdot 1938\rho^t \sqrt{\frac{(\mu r)^5 \kappa^9 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}} \\
&\le 0.25 \frac{\sqrt{\sigma_{\max}}}{\kappa},
\end{aligned}
\tag{158}
$$

where the leave-one-out projection operator $\boldsymbol{\mathcal{P}}_{-i,\cdot}$ is defined as

$$
\left[\boldsymbol{\mathcal{P}}_{-i,\cdot}\left(\boldsymbol{\mathcal{X}}\right)\right]_{kj:} := \begin{cases} \boldsymbol{\mathcal{X}}_{kj:}, & k \neq i, \forall j \in [n], \\ \mathbf{0}_m, & k = i, \forall j \in [n]. \end{cases}
\tag{159}
$$

(i) is because the spectral norm of the sub-tensor is not larger than the original tensor. (ii) and (iii) are due to the Lemma 2 and tensor norm inequality (95), respectively. The (iv) is based on the tensor incoherence condition (10), (64), and tube-wise complexity in (12).

For the term $\left\|\begin{bmatrix}\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right]-\boldsymbol{\mathcal{A}}^{\star}\\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right]-\boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\|$, we can follow the (45) in Section A.2.2 as

$$\begin{aligned}
&\left\|\begin{bmatrix}\mathbb{E}\left[\tilde{\boldsymbol{\mathcal{A}}}^{t+1,l}\right]-\boldsymbol{\mathcal{A}}^{\star}\\ \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{B}}}^{t+1,l}\right]-\boldsymbol{\mathcal{B}}^{\star}\end{bmatrix}\right\| \\
&\quad=\left\|\frac{1}{2}\begin{bmatrix}\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\\ \boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\end{bmatrix}*\left(\boldsymbol{\mathcal{I}}_r-2\eta m\left(\boldsymbol{\mathcal{A}}^{\star}\right)^T*\boldsymbol{\mathcal{A}}^{\star}\right)+\frac{1}{2}\begin{bmatrix}\left(\boldsymbol{\mathcal{I}}_n-2\eta m\boldsymbol{\mathcal{A}}^{\star}\left(\boldsymbol{\mathcal{A}}^{\star}\right)^T\right)*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{A}}}^{t,l}\\ \left(\boldsymbol{\mathcal{I}}_n-2\eta m\boldsymbol{\mathcal{B}}^{\star}\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\right)*\boldsymbol{\Delta}_{\boldsymbol{\mathcal{B}}}^{t,l}\end{bmatrix}+\eta m\begin{bmatrix}\boldsymbol{\mathcal{E}}_1^{t,l}\\ \boldsymbol{\mathcal{E}}_2^{t,l}\end{bmatrix}\right\| \\
&\quad\leq\frac{1}{2}\left\|\boldsymbol{\mathcal{I}}_r-2\eta m\left(\boldsymbol{\mathcal{A}}^{\star}\right)^T*\boldsymbol{\mathcal{A}}^{\star}\right\|\left\|\boldsymbol{\Delta}^{t,l}\right\|+\frac{1}{2}\left\|\boldsymbol{\mathcal{I}}_{2n}-2\eta m\begin{bmatrix}\boldsymbol{\mathcal{A}}^{\star}*\left(\boldsymbol{\mathcal{A}}^{\star}\right)^T & \boldsymbol{\mathcal{O}}\\ \boldsymbol{\mathcal{O}} & \boldsymbol{\mathcal{B}}^{\star}*\left(\boldsymbol{\mathcal{B}}^{\star}\right)^T\end{bmatrix}\right\|\left\|\boldsymbol{\Delta}^{t,l}\right\| \\
&\leq\left(1-\eta m\sigma_{\min}\right)\cdot\left\|\boldsymbol{\Delta}^{t,l}\right\|+4\eta m\left\|\boldsymbol{\Delta}^{t,l}\right\|^2\left\|\boldsymbol{\mathcal{A}}^{\star}\right\| \\
&\leq\left(1-0.9\eta m\sigma_{\min}\right)\left\|\boldsymbol{\Delta}^{t,l}\right\| \\
&\leq 0.25\frac{\sqrt{\sigma_{\max}}}{\kappa},
\end{aligned}\tag{160}$$

where the last two inequalities are due to tube-wise sample complexity.

Thus, based on (157) and Wely's inequality, we can also conclude that $\sigma_{\min}\left(\boldsymbol{\mathcal{C}}\right)\geq 1.5\sigma_{\min}$, which means that is positive definite tensor. Moreover, based on (153), we can bound the spectral norm of $\boldsymbol{\mathcal{E}}$ as

$$\begin{aligned}
\|\boldsymbol{\mathcal{E}}\|&\leq 2\sqrt{\sigma_{\max}}\cdot 2\times 1250^2 m\eta C^2\rho^{2t}\sqrt{\sigma_{\max}}\left(\sqrt{\frac{(\mu r)^4\kappa^7 n_3\log^2(n\vee n_3)\sigma_{\max}}{nm}}\right)^2 \\
&\leq 0.5\sigma_{\min},
\end{aligned}\tag{161}$$

where the last inequality is due to the tube-wise sample complexity and step size in Theorem 2. Thus, $\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{E}}$ satisfy the condition in Lemma 5, applying Lemma 5 as (152) would have

$$\begin{aligned}
\left\|\left(\boldsymbol{\mathcal{R}}^{t,l}\right)^{-1}*\boldsymbol{\mathcal{R}}^{t+1,l}-\boldsymbol{\mathcal{I}}_r\right\|&\leq\frac{2\times 1250^2 m\eta C^2\rho^{2t}\sqrt{\sigma_{\max}}\left(\sqrt{\frac{(\mu r)^4\kappa^7 n_3\log^2(n\vee n_3)\sigma_{\max}}{nm}}\right)^2\cdot\sqrt{\sigma_{\max}}}{1.5\sigma_{\min}-0.5\sigma_{\min}} \\
&=2\times 1250^2 m\eta C^2\rho^{2t}\left(\sqrt{\frac{(\mu r)^4\kappa^8 n_3\log^2(n\vee n_3)\sigma_{\max}}{nm}}\right)^2.
\end{aligned}\tag{162}$$

□

B Proof for Spectral Initialization

To use the same symmetric dilation technique for the matrix [1, 8, 40] in our setting, we construct the auxiliary tensor

$$\hat{\boldsymbol{\mathcal{X}}^{\star}}:=\begin{bmatrix}\mathbf{0}_{n\times n_\times n_3} & \boldsymbol{\mathcal{X}}^{\star}\\ \left(\boldsymbol{\mathcal{X}}^{\star}\right)^T & \mathbf{0}_{n\times n_\times n_3}\end{bmatrix}\in\mathbb{R}^{2n\times 2n\times n_3},\tag{163}$$

which is the symmetric tensor. This is because

$$\hat{\boldsymbol{X}^{\star}}^{(1)}=\begin{bmatrix}\mathbf{0}_{n\times n} & \boldsymbol{X}^{\star(1)}\\ \left(\boldsymbol{X}^{\star(1)}\right)^T & \mathbf{0}_{n\times n}\end{bmatrix}\tag{164}$$

is symmetric. $\forall k = 2, \cdots, n_3$, there is

$$\hat{\boldsymbol{X}^{\star}}^{(k)} = \begin{bmatrix} \mathbf{0}_{n\times n} & \boldsymbol{X}^{\star(k)} \\ \left(\boldsymbol{X}^{\star(n_3+2-k)}\right)^T & \mathbf{0}_{n\times n} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{X}^{\star}}^{(n_3+2-k)} = \begin{bmatrix} \mathbf{0}_{n\times n} & \boldsymbol{X}^{\star(n_3+2-k)} \\ \left(\boldsymbol{X}^{\star(k)}\right)^T & \mathbf{0}_{n\times n} \end{bmatrix}. \tag{165}$$

Thus, there is

$$\left(\hat{\boldsymbol{X}^{\star}}^{(k)}\right)^T = \hat{\boldsymbol{X}^{\star}}^{(n_3+2-k)}, \tag{166}$$

which obeys the definition of the symmetric tensor in Definition 3. Thus, each frontal slice of $\hat{\boldsymbol{\mathcal{X}}^{\star}}$ is the symmetric matrix and has tensor eigenvalue decomposition as

$$\hat{\boldsymbol{\mathcal{X}}^{\star}} = \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\mathcal{U}}^{\star} & \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{V}}^{\star} & -\boldsymbol{\mathcal{V}}^{\star} \end{bmatrix} * \begin{bmatrix} \boldsymbol{\mathcal{S}}^{\star} & \mathbf{0}_{n\times n\times n_3} \\ \mathbf{0}_{n\times n\times n_3} & -\boldsymbol{\mathcal{S}}^{\star} \end{bmatrix} * \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\mathcal{U}}^{\star} & \boldsymbol{\mathcal{U}}^{\star} \\ \boldsymbol{\mathcal{V}}^{\star} & -\boldsymbol{\mathcal{V}}^{\star} \end{bmatrix}^T. \tag{167}$$

Similarly, we can also define

$$\hat{\boldsymbol{\mathcal{X}}^{0}} := \begin{bmatrix} \mathbf{0}_{n\times n\times n_3} & \boldsymbol{\mathcal{X}}^{0} \\ \left(\boldsymbol{\mathcal{X}}^{0}\right)^T & \mathbf{0}_{n\times n\times n_3} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\mathcal{X}}^{0,l}} := \begin{bmatrix} \mathbf{0}_{n\times n\times n_3} & \boldsymbol{\mathcal{X}}^{0,l} \\ \left(\boldsymbol{\mathcal{X}}^{0,l}\right)^T & \mathbf{0}_{n\times n\times n_3} \end{bmatrix}. \tag{168}$$

Then we have

$$\left\|\hat{\boldsymbol{\mathcal{X}}^{0,l}} - \hat{\boldsymbol{\mathcal{X}}^{\star}}\right\| = \begin{bmatrix} \mathbf{0}_{n\times n\times n_3} & \frac{1}{m}\boldsymbol{\mathcal{Z}}^c_{-l,\cdot}\boldsymbol{\mathcal{Z}}_{-l,\cdot}\left(\boldsymbol{\mathcal{X}}^{\star}\right) + \boldsymbol{\mathcal{P}}_{l,\cdot}(\boldsymbol{\mathcal{X}}^{\star}) - \boldsymbol{\mathcal{X}}^{\star} \\ \left(\frac{1}{m}\boldsymbol{\mathcal{Z}}^c_{-l,\cdot}\boldsymbol{\mathcal{Z}}_{-l,\cdot}\left(\boldsymbol{\mathcal{X}}^{\star}\right) + \boldsymbol{\mathcal{P}}_{l,\cdot}(\boldsymbol{\mathcal{X}}^{\star}) - \boldsymbol{\mathcal{X}}^{\star}\right)^T & \mathbf{0}_{n\times n\times n_3} \end{bmatrix}, \tag{169}$$

which is actually the sub-part of the $\hat{\boldsymbol{\mathcal{X}}^{0}} - \hat{\boldsymbol{\mathcal{X}}^{\star}}$. In addition, based on the definition of $\boldsymbol{\mathcal{X}}^{0}$, there is

$$\begin{aligned} \left\|\hat{\boldsymbol{\mathcal{X}}^{0}} - \hat{\boldsymbol{\mathcal{X}}^{\star}}\right\| &= \left\|\begin{bmatrix} \mathbf{0}_{n\times n\times n_3} & \frac{1}{m}\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{X}}^{\star}) - \boldsymbol{\mathcal{X}}^{\star} \\ \left(\frac{1}{m}\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{X}}^{\star}) - \boldsymbol{\mathcal{X}}^{\star}\right)^T & \mathbf{0}_{n\times n\times n_3} \end{bmatrix}\right\| \\ &= \left\|\frac{1}{m}\boldsymbol{\mathcal{Z}}^c\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{X}}^{\star}) - \boldsymbol{\mathcal{X}}^{\star}\right\| \\ &\le 2Cr\sqrt{\frac{nn_3}{m}}\log(n\vee n_3)\max_{ij}\left\|\boldsymbol{\mathcal{X}}^{\star}_{ij:}\right\|_F \\ &\overset{(i)}{\le} 2Cr\sqrt{\frac{nn_3}{m}}\log(n\vee n_3)\left(\sqrt{\frac{\mu r\sigma_{\max}}{n}}\right)^2 \\ &\le 2C\sqrt{\frac{(\mu r)^4 n_3\log^2(n\vee n_3)\kappa^2}{mn}}\sigma_{\min} \\ &\le \frac{1}{40\kappa^2}\sigma_{\min} \\ &\le \frac{\sigma_{\min}}{4}, \end{aligned} \tag{170}$$

where (i) is due to the tensor norm inequality (95) and the tensor incoherence condition in (10). The last inequality is due to the tube-wise sample complexity in (25). Thus, we also have

$$\left\|\hat{\boldsymbol{\mathcal{X}}^{0,l}} - \hat{\boldsymbol{\mathcal{X}}^{\star}}\right\| \le \left\|\hat{\boldsymbol{\mathcal{X}}^{0}} - \hat{\boldsymbol{\mathcal{X}}^{\star}}\right\| \le \frac{\sigma_{\min}}{4}. \tag{171}$$

Because $\hat{\boldsymbol{\mathcal{X}}^{0,l}}, \hat{\boldsymbol{\mathcal{X}}^{0}}$ are symmetric tensor, if the top-r skinny of t-SVDs are as

$$\begin{aligned} \boldsymbol{\mathcal{U}}^{0} * \boldsymbol{\mathcal{S}}^{0} * \left(\boldsymbol{\mathcal{V}}^{0}\right)^T &= \boldsymbol{\mathcal{X}}^{0} \\ \boldsymbol{\mathcal{U}}^{0,l} * \boldsymbol{\mathcal{S}}^{0,l} * \left(\boldsymbol{\mathcal{V}}^{0,l}\right)^T &= \boldsymbol{\mathcal{X}}^{0,l}. \end{aligned} \tag{172}$$

Then, the top-r T-eigenvalue decompositions [72] for $\hat{\boldsymbol{\mathcal{X}}}^{0}, \hat{\boldsymbol{\mathcal{X}}}^{0,l}$ are

$$\frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^0\\ \boldsymbol{\mathcal{V}}^0\end{bmatrix} * \boldsymbol{\mathcal{S}}^0 * \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^0\\ \boldsymbol{\mathcal{V}}^0\end{bmatrix}^T \quad \text{and} \quad \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^{0,l}\\ \boldsymbol{\mathcal{V}}^{0,l}\end{bmatrix} * \boldsymbol{\mathcal{S}}^{0,l} * \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^{0,l}\\ \boldsymbol{\mathcal{V}}^{0,l}\end{bmatrix}^T, \tag{173}$$

respectively. Eq.(170) means that in the block diagonal formula, there is

$$\left\|\overline{\hat{\boldsymbol{X}}^{0,l}} - \overline{\hat{\boldsymbol{X}}^{\star}}\right\| \leq \frac{\sigma_{\min}}{4} \quad \text{and} \quad \left\|\overline{\hat{\boldsymbol{X}}^{0}} - \overline{\hat{\boldsymbol{X}}^{\star}}\right\| \leq \frac{\sigma_{\min}}{4}. \tag{174}$$

Based on Wely's inequality [61] and the above inequalities, there are

$$\frac{3}{4}\sigma_{\min} \leq \sigma_{rn_3}(\overline{\boldsymbol{S}^0}) \leq \sigma_{\max}(\overline{\boldsymbol{S}^0}) \leq \frac{5}{4}\sigma_{\max} \tag{175}$$

$$\frac{3}{4}\sigma_{\min} \leq \sigma_{rn_3}(\overline{\boldsymbol{S}^{0,l}}) \leq \sigma_{\max}(\overline{\boldsymbol{S}^{0,l}}) \leq \frac{5}{4}\sigma_{\max}. \tag{176}$$

Recall the definition of $\boldsymbol{\mathcal{A}}^0 = \boldsymbol{\mathcal{U}}^0 * \left(\boldsymbol{\mathcal{S}}^0\right)^{\frac{1}{2}}, \boldsymbol{\mathcal{B}}^0 = \boldsymbol{\mathcal{V}}^0 * \left(\boldsymbol{\mathcal{S}}^0\right)^{\frac{1}{2}}, \boldsymbol{\mathcal{A}}^{0,l} = \boldsymbol{\mathcal{U}}^{0,l} * \left(\boldsymbol{\mathcal{S}}^{0,l}\right)^{\frac{1}{2}}, \boldsymbol{\mathcal{B}}^0 = \boldsymbol{\mathcal{V}}^{0,l} * \left(\boldsymbol{\mathcal{S}}^{0,l}\right)^{\frac{1}{2}}$. Define the tensor corresponding eigen-spaces

$$\tilde{\boldsymbol{\mathcal{G}}} := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^\star\\ \boldsymbol{\mathcal{V}}^\star\end{bmatrix}, \quad \boldsymbol{\mathcal{G}} := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix} \tag{177}$$

and

$$\tilde{\boldsymbol{\mathcal{H}}}^0 := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^0\\ \boldsymbol{\mathcal{V}}^0\end{bmatrix}, \quad \boldsymbol{\mathcal{H}}^0 := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{A}}^0\\ \boldsymbol{\mathcal{B}}^0\end{bmatrix}, \quad \tilde{\boldsymbol{\mathcal{H}}}^{0,l} := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{U}}^{0,l}\\ \boldsymbol{\mathcal{V}}^{0,l}\end{bmatrix}, \quad \boldsymbol{\mathcal{H}}^{0,l} := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{\mathcal{A}}^{0,l}\\ \boldsymbol{\mathcal{B}}^{0,l}\end{bmatrix}. \tag{178}$$

Further, the alignment tensors between these eigen-spaces are defined as

$$\boldsymbol{\mathcal{Q}}^0 := \arg\min_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \left\|\tilde{\boldsymbol{\mathcal{H}}}^0 * \boldsymbol{\mathcal{R}} - \tilde{\boldsymbol{\mathcal{G}}}\right\|_F \tag{179}$$

$$\boldsymbol{\mathcal{Q}}^{0,l} := \arg\min_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \left\|\tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{R}} - \tilde{\boldsymbol{\mathcal{G}}}\right\|_F. \tag{180}$$

B.1 Proof of (26)

$$\begin{aligned}
\left\|\begin{bmatrix}\boldsymbol{\mathcal{A}}^0\\ \boldsymbol{\mathcal{B}}^0\end{bmatrix} * \boldsymbol{\mathcal{R}}^0 - \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star\\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}\right\| &= \sqrt{2}\left\|\boldsymbol{\mathcal{H}}^0 * \boldsymbol{\mathcal{R}}^0 - \boldsymbol{\mathcal{G}}\right\| \\
&= \sqrt{2}\Big\|\tilde{\boldsymbol{\mathcal{H}}}^0 * (\boldsymbol{\mathcal{S}}^0)^{\frac{1}{2}}\left(\boldsymbol{\mathcal{R}}^0 - \boldsymbol{\mathcal{Q}}^0\right) + \tilde{\boldsymbol{\mathcal{H}}}^0\left((\boldsymbol{\mathcal{S}}^0)^{\frac{1}{2}} * \boldsymbol{\mathcal{Q}}^0 - \boldsymbol{\mathcal{Q}}^0 * (\boldsymbol{\mathcal{S}}^\star)^{\frac{1}{2}}\right) \\
&\quad + \left(\tilde{\boldsymbol{\mathcal{H}}}^0 * \boldsymbol{\mathcal{Q}}^0 - \tilde{\boldsymbol{\mathcal{G}}}\right) * (\boldsymbol{\mathcal{S}}^\star)^{\frac{1}{2}}\Big\| \\
&\leq \sqrt{2}\left\|\boldsymbol{\mathcal{S}}^0\right\|^{\frac{1}{2}}\left\|\boldsymbol{\mathcal{R}}^0 - \boldsymbol{\mathcal{Q}}^0\right\| + \sqrt{2}\left\|(\boldsymbol{\mathcal{S}}^0)^{\frac{1}{2}} * \boldsymbol{\mathcal{Q}}^0 - \boldsymbol{\mathcal{Q}}^0 * (\boldsymbol{\mathcal{S}}^\star)^{\frac{1}{2}}\right\| \\
&\quad + \sqrt{2}\left\|\boldsymbol{\mathcal{S}}^\star\right\|^{\frac{1}{2}}\left\|\tilde{\boldsymbol{\mathcal{G}}}^0 * \boldsymbol{\mathcal{Q}}^0 - \tilde{\boldsymbol{\mathcal{G}}}\right\| \\
&\overset{(i)}{\leq} 15\sqrt{2}\frac{\sigma_{\max}^{\frac{1}{2}}\kappa^{\frac{3}{2}}}{\sigma_{\min}}\left\|\hat{\boldsymbol{\mathcal{X}}}^0 - \hat{\boldsymbol{\mathcal{X}}}^\star\right\| + 15\sqrt{2}\frac{\kappa}{\sigma_{\min}^{\frac{1}{2}}}\left\|\hat{\boldsymbol{\mathcal{X}}}^0 - \hat{\boldsymbol{\mathcal{X}}}^\star\right\| \\
&\quad + \frac{3\sqrt{2}\sigma_{\max}^{\frac{1}{2}}}{\sigma_{\min}}\left\|\hat{\boldsymbol{\mathcal{X}}}^0 - \hat{\boldsymbol{\mathcal{X}}}^\star\right\| \\
&\overset{(ii)}{\leq} 2C\sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3)\kappa\sigma_{\max}}{mn}}\sqrt{\sigma_{\min}} \\
&\quad \cdot \left(22\frac{\sigma_{\max}^{\frac{1}{2}}\kappa^{\frac{3}{2}}}{\sigma_{\min}} + 22\frac{\kappa}{\sigma_{\min}^{\frac{1}{2}}} + 5\frac{\sigma_{\max}^{\frac{1}{2}}}{\sigma_{\min}}\right) \\
&\leq 98C\sqrt{\frac{(\mu r)^4\kappa^5 n_3 \log^2(n \vee n_3)\sigma_{\max}}{mn}},
\end{aligned} \tag{181}$$

where we use Lemma 10 and 12 in (i) and (170) in (ii).

B.2 Proof of (28)

$$
\begin{aligned}
&\left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| = \sqrt{2} \left\| \left(\boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{R}}^{0,l} - \boldsymbol{\mathcal{G}} \right)_{l::} \right\| \\
&\quad \leq \sqrt{2} \left\| \left(\boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{G}} \right)_{l::} \right\| + \sqrt{2} \left\| \boldsymbol{\mathcal{H}}^{0,l}_{l::} * (\boldsymbol{\mathcal{R}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l}) \right\| \\
&\quad = \sqrt{2} \left\| \left(\hat{\boldsymbol{\mathcal{X}}}^{0,l} * \tilde{\boldsymbol{\mathcal{H}}}^{0,l} * (\boldsymbol{\mathcal{S}}^{0,l})^{-\frac{1}{2}} * \boldsymbol{\mathcal{Q}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{\star} * \tilde{\boldsymbol{\mathcal{G}}} * (\boldsymbol{\mathcal{S}}^{\star})^{-\frac{1}{2}} \right)_{l::} \right\| \\
&\qquad + \sqrt{2} \left\| \boldsymbol{\mathcal{H}}^{0,l}_{l::} \right\| \left\| \boldsymbol{\mathcal{R}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l} \right\| \\
&\quad = \sqrt{2} \left\| \hat{\boldsymbol{\mathcal{X}}}^{\star}_{l::} * \left(\tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \left(\left(\boldsymbol{\mathcal{S}}^{0,l} \right)^{-\frac{1}{2}} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l} * (\boldsymbol{\mathcal{S}}^{\star})^{-\frac{1}{2}} \right) \right. \right. \\
&\qquad \left. \left. + \left(\tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \tilde{\boldsymbol{\mathcal{G}}} \right) * (\boldsymbol{\mathcal{S}}^{\star})^{-\frac{1}{2}} \right) \right\| + \sqrt{2} \left\| \boldsymbol{\mathcal{H}}^{0,l}_{l::} \right\| \left\| \boldsymbol{\mathcal{R}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l} \right\| \\
&\quad \leq \sqrt{2} \left\| \hat{\boldsymbol{\mathcal{X}}}^{\star}_{l::} \right\| \left(\left\| \left(\boldsymbol{\mathcal{S}}^{0,l} \right)^{-\frac{1}{2}} * \left(\left(\boldsymbol{\mathcal{S}}^{0,l} \right)^{\frac{1}{2}} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l} * (\boldsymbol{\mathcal{S}}^{\star})^{\frac{1}{2}} \right) * (\boldsymbol{\mathcal{S}}^{\star})^{-\frac{1}{2}} \right\| \right. \\
&\qquad \left. + \frac{\left\| \tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \tilde{\boldsymbol{\mathcal{G}}} \right\|}{\sqrt{\sigma_{\min}}} \right) + \sqrt{2} \left(\max_{l} \left\| \boldsymbol{\mathcal{G}}_{l::} \right\| + \left\| \left(\boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{G}} \right)_{l::} \right\| \right) \left\| \boldsymbol{\mathcal{R}}^{0,l} - \boldsymbol{\mathcal{Q}}^{0,l} \right\| \\
&\quad \overset{(i)}{\leq} \sqrt{2} \max_{i} \left\| \boldsymbol{\mathcal{A}}^{\star}_{i::} \right\| \cdot \sqrt{\sigma_{\max}} \left(\cdot \sqrt{\frac{4}{3\sigma_{\min}}} \cdot 15 \frac{\sigma_{\max}}{\sigma_{\min}^{\frac{3}{2}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{\star} \right\| \cdot \frac{1}{\sqrt{\sigma_{\min}}} \right. \\
&\qquad \left. + \frac{3}{\sigma_{\min}^{\frac{3}{2}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{\star} \right\| \right) + \sqrt{2} \left(\sqrt{\frac{\mu r \sigma_{\max}}{2n}} + \left\| \left(\boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{G}} \right)_{l::} \right\| \right) \\
&\qquad \cdot 15 \frac{\sigma_{\max}^{\frac{3}{2}}}{\sigma_{\min}^{\frac{5}{2}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{\star} \right\| \\
&\quad \overset{(ii)}{\leq} \frac{33\kappa\sigma_{\max}}{\sigma_{\min}^{\frac{3}{2}}} \cdot 2C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^2}{mn^2}} \sigma_{\min} + \sqrt{2} \left(\sqrt{\frac{\mu r \sigma_{\max}}{2n}} \right. \\
&\qquad \left. + \left\| \left(\boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{Q}}^{0,l} - \boldsymbol{\mathcal{G}} \right)_{l::} \right\| \right) \cdot 15 \frac{\sigma_{\max}^{\frac{3}{2}}}{\sigma_{\min}^{\frac{5}{2}}} \cdot 2C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^2}{mn}} \sigma_{\min} \\
&\quad \leq 66C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{mn^2}} + \sqrt{2} \left(66C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{mn^2}} \right. \\
&\qquad \left. + \sqrt{\frac{\mu r \sigma_{\max}}{2n}} \right) \cdot 15 \frac{\sigma_{\max}^{\frac{3}{2}}}{\sigma_{\min}^{\frac{5}{2}}} \cdot 2C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^2}{mn}} \sigma_{\min} \\
&\quad \overset{(iii)}{\leq} 66C \sqrt{\frac{(\mu r)^4 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{mn^2}} + 60C \sqrt{\frac{(\mu r)^5 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{mn^2}} \\
&\quad \leq 126C \sqrt{\frac{(\mu r)^5 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{mn^2}}, \qquad (182)
\end{aligned}
$$

where (i) uses the Lemma 12, (ii) is due to (171), and (iii) is due to the sample complexity bound in (25).

B.3 Proof of (29)

Without loss of generality, we consider the case $l \in [n]$, the denote the alignment tensor

$$\boldsymbol{\mathcal{J}}^{0,l} := \mathop{\arg\min}_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \left\| \tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{R}} - \tilde{\boldsymbol{\mathcal{H}}}^{0} \right\|_F . \tag{183}$$

Then based on definition of $\boldsymbol{\mathcal{T}}^{0.l}$, there is

$$\begin{aligned}
\left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{0,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0} \\ \boldsymbol{\mathcal{B}}^{0} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0} \right\| &= \sqrt{2} \left\| \boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{T}}^{0,l} - \boldsymbol{\mathcal{H}}^{0} * \boldsymbol{\mathcal{R}}^{0} \right\| \\
&\leq \sqrt{2} \left\| \boldsymbol{\mathcal{H}}^{0,l} * \boldsymbol{\mathcal{J}}^{0,l} - \boldsymbol{\mathcal{H}}^{0} * \boldsymbol{\mathcal{R}}^{0} \right\| \\
&\leq \sqrt{2} \left\| \tilde{\boldsymbol{\mathcal{H}}}^{0,l} \left(\left(\boldsymbol{\mathcal{S}}^{0,l} \right)^{\frac{1}{2}} * \boldsymbol{\mathcal{J}}^{0,l} - \boldsymbol{\mathcal{J}}^{0,l} * \left(\boldsymbol{\mathcal{S}}^{0} \right)^{\frac{1}{2}} \right) \right\| \\
&\quad + \sqrt{2} \left\| \left(\tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{J}}^{0,l} - \tilde{\boldsymbol{\mathcal{H}}}^{0} \right) * \left(\boldsymbol{\mathcal{S}}^{0} \right)^{\frac{1}{2}} \right\| \\
&\overset{(i)}{\leq} 15\sqrt{2} \frac{\sigma_{\max}}{\sigma_{\min}^{\frac{3}{2}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{0} \right\| + \sqrt{2} \cdot \sqrt{\frac{5\sigma_{\max}}{4}} \cdot \left\| \tilde{\boldsymbol{\mathcal{H}}}^{0,l} * \boldsymbol{\mathcal{J}}^{0,l} - \tilde{\boldsymbol{\mathcal{H}}}^{0} \right\| \\
&\overset{(ii)}{\leq} 15\sqrt{2} \frac{\sigma_{\max}}{\sigma_{\min}^{\frac{3}{2}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{0} \right\| + \sqrt{\frac{5\sigma_{\max}}{2}} \cdot \frac{2\sqrt{2}}{\sigma_{\min}} \left\| \left(\hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{0} \right) * \tilde{\boldsymbol{\mathcal{H}}}^{0,l} \right\| \\
&\leq 26 \frac{\kappa}{\sqrt{\sigma_{\min}}} \left\| \hat{\boldsymbol{\mathcal{X}}}^{0,l} - \hat{\boldsymbol{\mathcal{X}}}^{0} \right\| \\
&\overset{(iii)}{=} 26 \frac{\kappa}{\sqrt{\sigma_{\min}}} \cdot \left\| \boldsymbol{\mathcal{P}}_{l,\cdot}(\boldsymbol{\mathcal{X}}^{\star}) - \frac{1}{m} \boldsymbol{\mathcal{Z}}_{l,\cdot}^{c} \boldsymbol{\mathcal{Z}}_{l,\cdot}(\boldsymbol{\mathcal{X}}^{\star}) \right\| \\
&\overset{(iv)}{\leq} 26 \frac{\kappa}{\sqrt{\sigma_{\min}}} \cdot 2C \sqrt{\frac{n_3}{m}} r \log(n \vee n_3) \max_{ij} \left\| \boldsymbol{\mathcal{X}}_{ij:}^{\star} \right\|_F \\
&\overset{(v)}{\leq} 52C \sqrt{\frac{n_3 r^2 \log^2(n \vee n_3) \kappa}{m \sigma_{\min}}} \cdot \left(\sqrt{\frac{\mu r \sigma_{\max}}{n}} \right)^2 \\
&\leq 52C \sqrt{\frac{(\mu r)^4 \kappa^2 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}},
\end{aligned} \tag{184}$$

where (i) is due to Lemma 12 and (175). (ii) is based on matrix version of Davis-Kahan SinΘ theorem [9] and (iii) is based on definition of $\hat{\boldsymbol{\mathcal{X}}}^{0,l}, \hat{\boldsymbol{\mathcal{X}}}^{0}$ in (168). (iv) and (v) are due to our Lemma 2 and tensor incoherence condition in (10), respectively.

B.4 Proof of (27)

Proof. Its proof is the same as proving (14) that $\forall l \in [2n]$, there is

$$\begin{aligned}
\left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{0} \\ \boldsymbol{\mathcal{B}}^{0} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| &\leq \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0} \\ \boldsymbol{\mathcal{B}}^{0} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0,l} \right\| + \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| \\
&\overset{(i)}{\leq} 9\kappa \left\| \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0} \\ \boldsymbol{\mathcal{B}}^{0} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{T}}^{0,l} \right\| + \left\| \left(\begin{bmatrix} \boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l} \end{bmatrix} * \boldsymbol{\mathcal{R}}^{0,l} - \begin{bmatrix} \boldsymbol{\mathcal{A}}^{\star} \\ \boldsymbol{\mathcal{B}}^{\star} \end{bmatrix} \right)_{l::} \right\| \\
&\overset{(ii)}{\leq} 9\kappa \cdot 52C \sqrt{\frac{(\mu r)^4 \kappa^2 n_3 \log^2(n \vee n_3) \sigma_{\max}}{n^2 m}} \\
&\quad + 126C \sqrt{\frac{(\mu r)^5 n_3 \log^2(n \vee n_3) \kappa^5 \sigma_{\max}}{m n^2}} \\
&\leq 594C \sqrt{\frac{(\mu r)^5 \kappa^5 n_3 \log^2(n \vee n_3) \sigma_{\max}}{m n^2}}.
\end{aligned} \tag{185}$$

We apply the Lemma 6 in (i) as

$$\boldsymbol{\mathcal{X}}_0 = \begin{bmatrix}\boldsymbol{\mathcal{A}}^\star \\ \boldsymbol{\mathcal{B}}^\star\end{bmatrix}, \quad \boldsymbol{\mathcal{X}}_1 = \begin{bmatrix}\boldsymbol{\mathcal{A}}^0 \\ \boldsymbol{\mathcal{B}}^0\end{bmatrix} * \boldsymbol{\mathcal{R}}^0, \quad \boldsymbol{\mathcal{X}}_2 = \begin{bmatrix}\boldsymbol{\mathcal{A}}^{0,l} \\ \boldsymbol{\mathcal{B}}^{0,l}\end{bmatrix} * \boldsymbol{\mathcal{T}}^{0,l}. \tag{186}$$

The conditions $\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_0\| \, \|\boldsymbol{\mathcal{X}}_0\| \leq \frac{\sigma_{\min}}{2}$ and $\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| \, \|\boldsymbol{\mathcal{X}}_0\| \leq \frac{\sigma_{\min}}{4}$ can be satisfied based on (29),(26) and tube-wise sample complexity in (25). (ii) is due to the result in (28) and (29). □

B.5 Technical lemmas

This section provides the necessary lemmas used in proving that spectral initialization satisfied the results in Theorem 4. These can obtained based on results of matrix setting in [32, 8].

Lemma 9. *(T-eigenvalue decomposition [72]) Any symmetric tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times n \times n_3}$ (i.e., $\boldsymbol{\mathcal{X}}^T = \boldsymbol{\mathcal{X}}$) could be decomposed into $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{U}}^T$, where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n \times n \times n_3}$ is orthogonal tensor and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{n \times n \times n_3}$ is a f-diagonal tensor that diagonal entries of $\overline{\boldsymbol{S}}$ are T-eigenvalues of $\boldsymbol{\mathcal{X}}$.*

Lemma 10. *Let the $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times n \times n_3}$ be a symmetric tensor with top-r T-eigenvalue decomposition as $\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{U}}^T$ and $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}^\star\| \leq \frac{\sigma_{\min}}{2}$ and denote*

$$\hat{\boldsymbol{\mathcal{Q}}} := \arg\min_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \|\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{U}}^\star\|_F, \tag{187}$$

where $\boldsymbol{\mathcal{X}}^\star$ has T-eigenvalue decomposition as $\boldsymbol{\mathcal{U}}^\star * \boldsymbol{\mathcal{S}}^\star * (\boldsymbol{\mathcal{U}}^\star)^T$. Then there is

$$\left\|\boldsymbol{\mathcal{U}} * \hat{\boldsymbol{\mathcal{Q}}} - \boldsymbol{\mathcal{U}}^\star\right\| \leq \frac{3}{\sigma_{\min}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}^\star\|. \tag{188}$$

Proof. The proof applies the result under the matrix case for each frontal slice of the tensor in the frequency domain. Based on (117), $\forall k \in [n_3]$

$$\overline{\hat{\boldsymbol{Q}}}^{(k)} := \arg\min_{\boldsymbol{R} \in \boldsymbol{C}_r} \left\|\overline{\boldsymbol{U}}^{(k)} * \boldsymbol{R} - \overline{\boldsymbol{U}^\star}^{(k)}\right\|_F, \tag{189}$$

where $\overline{\boldsymbol{U}}^{(k)}, \overline{\boldsymbol{U}^\star}^{(k)} \in \mathbb{C}^{n \times r}$ are orthogonal matrices. Since $\left\|\overline{\boldsymbol{X}}^{(k)} - \overline{\boldsymbol{X}^\star}^{(k)}\right\| \leq \frac{\sigma_{\min}}{2} \leq \frac{\sigma_{\min}\left(\overline{\boldsymbol{X}^\star}^{(k)}\right)}{2}$, based on Lemma 32 in [32], there is $\forall k \in [n_3]$

$$\left\|\overline{\boldsymbol{U}}^{(k)} * \overline{\hat{\boldsymbol{Q}}}^{(k)} - \overline{\boldsymbol{U}^\star}^{(k)}\right\| \leq \frac{3}{\sigma_{\min}(\overline{\boldsymbol{X}^\star}^{(k)})} \left\|\overline{\boldsymbol{X}}^{(k)} - \overline{\boldsymbol{X}^\star}^{(k)}\right\|, \tag{190}$$

which induces

$$\max_{k \in [n_3]} \left\|\overline{\boldsymbol{U}}^{(k)} * \overline{\hat{\boldsymbol{Q}}}^{(k)} - \overline{\boldsymbol{U}^\star}^{(k)}\right\| \leq \frac{3}{\min_{k \in [n_3]} \sigma_{\min}(\overline{\boldsymbol{X}^\star}^{(k)})} \max_{k \in [n_3]} \left\|\overline{\boldsymbol{X}}^{(k)} - \overline{\boldsymbol{X}^\star}^{(k)}\right\|, \tag{191}$$

which is the conclusion of the lemma. □

Lemma 11. *Let $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \in \mathbb{R}^{n \times n \times n_3}$ be symmetric tensors with top-r T-eigenvalue decomposition $\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{S}}_1 * \boldsymbol{\mathcal{U}}_1^T, \boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{S}}_2 * \boldsymbol{\mathcal{U}}_2^T$ correspondingly. Assume for each frontal slice $\lambda_1(\overline{\boldsymbol{X}}_1^{(k)}) = -\lambda_n(\overline{\boldsymbol{X}}_1^{(k)}), \lambda_r(\overline{\boldsymbol{X}}_1^{(k)}) > 0, \lambda_{r+1}(\overline{\boldsymbol{X}}_1^{(k)}) = 0$ and*

$$\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| \leq \frac{1}{40} \frac{\lambda_{\min}^{\frac{5}{2}}(\boldsymbol{\mathcal{X}}_1)}{\lambda_{\max}^{\frac{3}{2}}(\boldsymbol{\mathcal{X}}_1)}. \tag{192}$$

Denote $\boldsymbol{\mathcal{A}}_1 = \boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{S}}_1^{\frac{1}{2}}, \boldsymbol{\mathcal{A}}_2 = \boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{S}}_2^{\frac{1}{2}}$ and define

$$\hat{\boldsymbol{\mathcal{Q}}} := \arg\min_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \|\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{U}}_1\|_F, \tag{193}$$

$$\boldsymbol{\mathcal{Q}} := \arg\min_{\boldsymbol{\mathcal{R}} \in \boldsymbol{\mathcal{O}}_r} \|\boldsymbol{\mathcal{A}}_2 * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{A}}_1\|_F. \tag{194}$$

Then there is

$$\left\|\hat{\boldsymbol{\mathcal{Q}}} - \boldsymbol{\mathcal{Q}}\right\| \leq 15 \frac{\lambda_{\max}^{\frac{3}{2}}(\boldsymbol{\mathcal{X}}_1)}{\lambda_{\min}^{\frac{5}{2}}(\boldsymbol{\mathcal{X}}_1)} \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\|. \tag{195}$$

Proof. The proof is similar to the proof in Lemma 10. Based on condition in (192), $\forall k \in [n_3]$, there is

$$\left\|\overline{\boldsymbol{X}}_1^{(k)} - \overline{\boldsymbol{X}}_2^{(k)}\right\| \leq \frac{1}{40}\frac{\lambda_{\min}^{\frac{5}{2}}(\boldsymbol{\mathcal{X}}_1)}{\lambda_{\max}^{\frac{3}{2}}(\boldsymbol{\mathcal{X}}_1)} \leq \frac{1}{40}\frac{\lambda_{\min}^{\frac{5}{2}}(\overline{\boldsymbol{X}}_1^{(k)})}{\lambda_{\max}^{\frac{3}{2}}(\overline{\boldsymbol{X}}_1^{(k)})}. \tag{196}$$

Combined with result in Lemma 34 in [32], we have

$$\left\|\overline{\hat{\boldsymbol{Q}}}^{(k)} - \overline{\boldsymbol{Q}}^{(k)}\right\| \leq 15\frac{\lambda_{\max}^{\frac{3}{2}}\left(\overline{\boldsymbol{X}}_1^{(k)}\right)}{\lambda_{\min}^{\frac{5}{2}}(\overline{\boldsymbol{X}}_1^{(k)})}\left\|\overline{\boldsymbol{X}}_1^{(k)} - \overline{\boldsymbol{X}}_2^{(k)}\right\|, \tag{197}$$

which induce

$$\max_{k\in[n_3]}\left\|\overline{\hat{\boldsymbol{Q}}}^{(k)} - \overline{\boldsymbol{Q}}^{(k)}\right\| \leq 15\frac{\lambda_{\max}^{\frac{3}{2}}(\boldsymbol{\mathcal{X}}_1)}{\lambda_{\min}^{\frac{5}{2}}(\boldsymbol{\mathcal{X}}_1)}\max_{k\in[n_3]}\left\|\overline{\boldsymbol{X}}_1^{(k)} - \overline{\boldsymbol{X}}_2^{(k)}\right\|, \tag{198}$$

which is exactly the result of the lemma. □

Lemma 12. *Let $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \boldsymbol{\mathcal{X}}_3 \in \mathbb{R}^{n\times n\times n_3}$ be symmetric tensors with top-r T-eigenvalue decomposition $\boldsymbol{\mathcal{U}}_1 * \boldsymbol{\mathcal{S}}_1 * \boldsymbol{\mathcal{U}}_1^T$,$\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{S}}_2 * \boldsymbol{\mathcal{U}}_2^T$ and $\boldsymbol{\mathcal{U}}_3 * \boldsymbol{\mathcal{S}}_3 * \boldsymbol{\mathcal{U}}_3^T$ correspondingly. Assume for each frontal slice $\lambda_1(\overline{\boldsymbol{X}}_1^{(k)}) = -\lambda_n(\overline{\boldsymbol{X}}_1^{(k)}), \lambda_r(\overline{\boldsymbol{X}}_1^{(k)}) > 0, \lambda_{r+1}(\overline{\boldsymbol{X}}_1^{(k)}) = 0$ and $\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\| \leq \frac{\lambda_{\min}(\boldsymbol{\mathcal{X}}_1)}{4}, \|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_3\| \leq \frac{\lambda_{\min}(\boldsymbol{\mathcal{X}}_1)}{4}$. Denote*

$$\hat{\boldsymbol{\mathcal{Q}}} := \arg\min_{\boldsymbol{\mathcal{R}}\in\boldsymbol{\mathcal{O}}_r}\|\boldsymbol{\mathcal{U}}_2 * \boldsymbol{\mathcal{R}} - \boldsymbol{\mathcal{U}}_3\|_F\,. \tag{199}$$

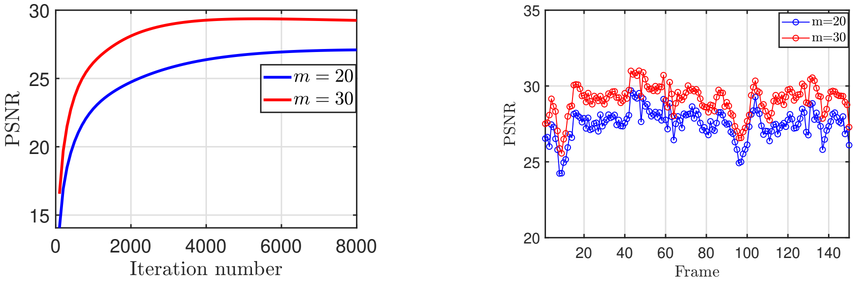
Then

$$\left\|\boldsymbol{\mathcal{S}}_2^{\frac{1}{2}} * \hat{\boldsymbol{\mathcal{Q}}} - \hat{\boldsymbol{\mathcal{Q}}} * \boldsymbol{\mathcal{S}}_3^{\frac{1}{2}}\right\| \leq 15\frac{\lambda_{\max}(\boldsymbol{\mathcal{X}}_1)}{\lambda_{\min}^{\frac{3}{2}}(\boldsymbol{\mathcal{X}}_1)}\|\boldsymbol{\mathcal{X}}_2 - \boldsymbol{\mathcal{X}}_3\|\,. \tag{200}$$

Proof. The proof follows the idea in Lemma 10 and Lemma 11 and applies the Lemma 33 in [32] to each frontal slice of $\overline{\hat{\boldsymbol{Q}}}^{(k)}$, $\overline{\boldsymbol{S}_2}^{(k)}$ and $\overline{\boldsymbol{S}_3}^{(k)}$ to obtain the final result based on definition of $\lambda_{\max}(\boldsymbol{\mathcal{X}}_1), \lambda_{\min}(\boldsymbol{\mathcal{X}}_1)$ for third-order tensor. □

C Experiments on real world data

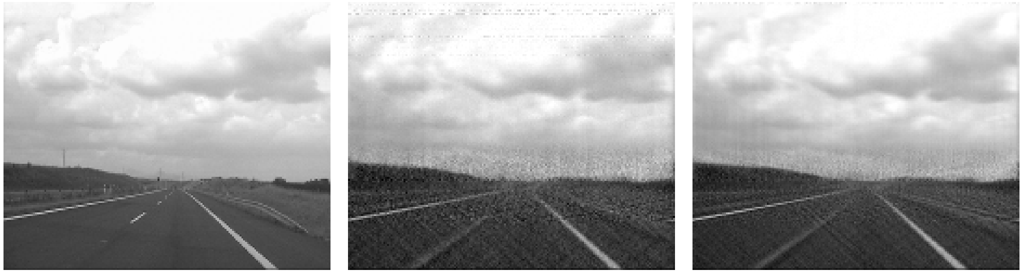
All experiments were conducted using MATLAB R2022a on a Windows system equipped with a 12th Gen Intel(R) Core(TM) i7-12700 CPU at 2.10 GHz and 16.0 GB of RAM.



(a) PSNR vs. iteration

(b) PSNR for each frame

Figure 3: Convergence and PSNR values for video sequence recovery.



(a) Ground truth

(b) $m = 20, r = 3$

(c) $m = 30, r = 3$

Figure 4: Visual comparison at 17th frame .

We test the proposed local TCS model on the "highway" video sequence, which is reshaped into third-order tensor with size $144 \times 150 \times 176$. 144, 150 denote the spatial dimension, and 176 denotes the tube size. When we compare with FGD [26] and Alt-PGD-Min [64] methods, both methods require large sample sizes that exceed our machine's memory capacity, making them infeasible to run in our setting. Because the video has a full tubal rank and only approximately has a low tubal rank, we have no prior setting of the parameter r in our model. Thus, we set $r = 3$ in our algorithm. For this r, we select the two tube-wise sample sizes as $m = 20, 30$, which are much smaller than the tube size 176. For setting the step size η, the video tensor has a large condition number $\kappa = 1.1408 \times 10^6$, which means the ground truth is an ill-conditioned tensor. We have to select small step sizes (0.00005) to guarantee the convergence, which leads to slow convergence rates. We use the PSNR metric to evaluate the recovery performance and plot the PSNR vs. iteration number in Fig. 3(a). We also plot the PSNR values across video frames in Fig. 3(b). The visual recovery results are shown in Fig. 4.

We can observe that although our recovered tensor only has $r = 3$, it has a very small reconstruction error for the ground-truth video sequence. This means that the original video sequence approximately has a low tubal rank. Besides the video sequence, MRI medical images and hyperspectral images often exhibit a low tubal rank structure because spectral (wavelength), temporal (frame), or slice axes are highly redundant. Low tubal rank-based tensor methods are often effective for these tasks, as shown in existing works in [23, 42, 57, 71]. Thus, based on these experimental results on real-world data, our low tubal rank Assumption 1 is realistic in practical applications.

We can observe that the GD converges slowly due to the very large condition number of the video tensor. This motivates us to use the preconditioning technique as [51, 52, 64] to get rid of dependence on κ to accelerate the convergence rate. Despite the slow convergence and the fact that the ground truth video tensor is not strictly low-tubal-rank, GD still achieves satisfactory recovery performance with small sample sizes. Moreover, as the sample size increases, the recovery quality improves accordingly.