Guaranteed Nonconvex Factorization Approach for Tensor Train Recovery

Zhen Qin   QIN.660@OSU.EDU
Department of Computer Science and Engineering
Ohio State University
Columbus, Ohio 43201, USA

Michael B. Wakin   MWAKIN@MINES.EDU
Department of Electrical Engineering
Colorado School of Mines
Golden, Colorado 80401, USA

Zhihui Zhu   ZHU.3440@OSU.EDU
Department of Computer Science and Engineering
Ohio State University
Columbus, Ohio 43201, USA

Editor: Shiqian Ma

Abstract

Tensor train (TT) decomposition represents an order-N tensor using $O(N)$ order-3 tensors (i.e., factors of small dimension), achieved through products among these factors. Due to its compact representation, TT decomposition has been widely used in the fields of signal processing, machine learning, and quantum physics. It offers benefits such as reduced memory requirements, enhanced computational efficiency, and decreased sampling complexity. Nevertheless, existing optimization algorithms with guaranteed performance concentrate exclusively on using the TT format for reducing the optimization space in recovery problems, while still operating on the entire tensor in each iteration. There is a lack of comprehensive theoretical analysis for optimization involving the factors directly, despite the proven efficacy of such factorization methods in practice. In this paper, we provide the first convergence guarantee for the factorization approach in a TT-based recovery problem. Specifically, to avoid the scaling ambiguity and to facilitate theoretical analysis, we optimize over the so-called left-orthogonal TT format which enforces orthonormality among most of the factors. To ensure the orthonormal structure, we utilize the Riemannian gradient descent (RGD) for optimizing those factors over the Stiefel manifold. We first delve into the TT factorization/decomposition problem and establish the local linear convergence of RGD. Notably, the rate of convergence only experiences a linear decline as the tensor order increases. We then study the sensing problem that aims to recover a TT format tensor from linear measurements. Assuming the sensing operator satisfies the restricted isometry property (RIP), we show that with a proper initialization, which could be obtained through spectral initialization, RGD also converges to the ground-truth tensor at a linear rate. Furthermore, we expand our analysis to encompass scenarios involving Gaussian noise in the measurements. We prove that RGD can reliably recover the ground truth at a linear rate, with the recovery error exhibiting only polynomial growth in relation to the tensor order N. We conduct various experiments to validate our theoretical findings.

Keywords: Tensor-train decomposition, factorization approach, Riemannian gradient descent, linear convergence

1. Introduction

Tensor estimation is a crucial task in various scientific and engineering fields, including signal processing and machine learning (Cichocki et al., 2015; Sidiropoulos et al., 2017), communication (Sidiropoulos et al., 2000), chemometrics (Smilde et al., 2005; Acar and Yener, 2008), genetic engineering (Hore et al., 2016), and so on. When dealing with an order-N tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, its exponentially increasing size with respect to N poses significant challenges in both memory and computation. To address this issue, tensor decomposition, which provides a compact representation of a tensor, has gained popularity in practical applications. The widely used tensor decompositions include the canonical polyadic (CP) (Bro, 1997), Tucker (Tucker, 1966), and tensor train (TT) (Oseledets, 2011) decompositions. These three formats have their pros and cons. The CP decomposition offers a storage advantage as it requires the least amount of storage, scaling linearly with N. However, determining the CP rank of a tensor is generally an NP-hard problem, as are tasks such as CP decomposition (Håstad, 1989; De Silva and Lim, 2008; Kolda and Bader, 2009; Cai et al., 2019). On the contrary, the Tucker decomposition can be approximately computed using the higher-order singular value decomposition. However, when representing a tensor using the Tucker decomposition, the size of the core tensor still grows exponentially in terms of N. This leads to significant memory consumption, making the Tucker decomposition more suitable for low-order tensors than for high-order ones.

In comparison, the *TT format* provides a balanced representation: in many cases it requires $O(N)$ parameters, while its quasi-optimal decomposition can be obtained through a sequential singular value decomposition (SVD) algorithm, commonly referred to as the tensor train SVD (TT-SVD) (Oseledets, 2011). Specifically, the $(s_1, \ldots, s_N)$-th element of $\mathcal{X}$ in the *TT format* can be expressed as the following matrix product form (Oseledets, 2011)

$$\mathcal{X}(s_1, \ldots, s_N) = \boldsymbol{X}_1(:, s_1, :)\boldsymbol{X}_2(:, s_2, :) \cdots \boldsymbol{X}_N(:, s_N, :), \tag{1}$$

where the tensor factors $\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}, i = 1, \ldots, N$ with $r_0 = r_N = 1$. See Figure 1 for an illustration. Thus, the TT format can be represented by N order-3 tensor factors $\{\boldsymbol{X}_i\}_{i \geq 1}$, with a total of $O(N\bar{d}\bar{r}^2)$ parameters, where $\bar{d} = \max_i d_i$ and $\bar{r} = \max_i r_i$. The dimensions $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$ of such a decomposition are called the *TT ranks* of $\mathcal{X}$. Any tensor can be decomposed in the TT format (1) with sufficiently large TT ranks (Oseledets, 2011, Theorem 2.1). Indeed, there always exists a TT decomposition with $r_i \leq \min\{\Pi_{j=1}^i d_j, \Pi_{j=i+1}^N d_j\}$ for any $i \geq 1$. We say a TT format tensor is low-rank if r_i is much smaller compared to $\min\{\Pi_{j=1}^i d_j, \Pi_{j=i+1}^N d_j\}$ for most indices[1] i so that the total number of parameters in the tensor factors $\{\boldsymbol{X}_i\}$ is much smaller than the number of entries in $\mathcal{X}$. We refer to any tensor for which such a low-rank TT decomposition exists as a *low-TT-rank* tensor.

Due to its compact representation, the TT decomposition with small TT ranks has found extensive applications in various fields. For instance, it has been widely used for image compression (Latorre, 2005; Bengua et al., 2017), analyzing theoretical properties of deep networks (Khrulkov et al., 2017), network compression or tensor networks (Stoudenmire and Schwab, 2016; Novikov et al., 2015; Yang et al., 2017; Tjandra et al., 2017; Yu et al., 2017; Ma et al., 2019), recommendation systems (Frolov and Oseledets, 2017), probabilistic model estimation (Novikov et al., 2021), learning of Hidden Markov Models (Kuznetsov and Oseledets, 2019), and more. Notably, as equivalents

---

1. When $i = 1$ or $N - 1$, r_1 or r_{N-1} may not be much smaller than d_1 or d_N.
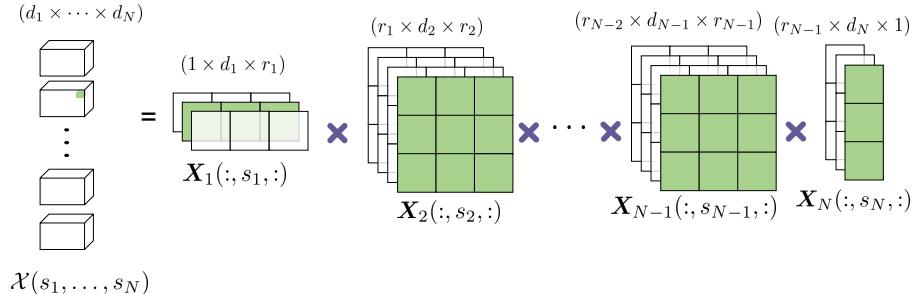
Figure 1: Illustration of the TT format (1).

to the TT decomposition, the matrix product state (MPS) and matrix product operator (MPO) decompositions have been introduced in the quantum physics community for efficiently and concisely representing quantum states. In this context, the parameter N represents the number of qubits in the many-body system (Verstraete and Cirac, 2006; Verstraete et al., 2008; Schollwöck, 2011). The concise representation provided by MPS and MPO is particularly valuable in quantum state tomography, as it allows us to observe a quantum state using computational and experimental resources that grow polynomially rather than exponentially with the number of qubits N (Ohliger et al., 2013).

A fundamental challenge in many of the aforementioned applications is to construct a low-TT-rank tensor from highly incomplete measurements of that tensor. The work (Bengua et al., 2017; Wang et al., 2019a) extends a nuclear-norm based convex relaxation approach from the matrix case to the TT case, but its high computational complexity makes it impractical for higher-order tensors. Alternating minimization (Wang et al., 2016) and gradient descent (Yuan et al., 2019b) have been employed to efficiently estimate the factors in the TT format, but theoretical guarantees regarding recovery error or convergence properties are not provided. Besides these heuristic algorithms, iterative hard thresholding (IHT) (Rauhut et al., 2017, 2015) and Riemannian gradient descent on the TT manifold (Budzinskiy and Zamarashkin, 2021; Wang et al., 2019b; Cai et al., 2022) have been proposed with local convergence guarantees. However, both methods necessitate the estimation of the entire tensor $\mathcal{X}$ in each iteration, which poses a challenge due to its exponential size in terms of N. As a result, both methods demand an exponential amount of storage or memory. In addition, theoretical results of IHT hinge on an unverified perturbation bound for TT-SVD. The Riemannian gradient descent method relies on the curvature information at the target tensor, which is often unknown a priori.

Instead of optimizing over the tensor $\mathcal{X}$ directly, in this paper, we focus on optimizing over the factors $\{\boldsymbol{X}_i\}_{i \geq 1}$ in the TT format. This factorization approach can significantly reduce the memory cost and has found widespread applications. For instance, gradient descent-based optimization on the TT factors has been successfully applied in various areas, including the TT deep computation model (Zhang et al., 2018), TT deep neural networks (Qi et al., 2022), TT completion (Yuan et al., 2019a), channel estimation (Zhang et al., 2021b) and quantum tomography (Lidiak et al., 2022). However, to the best of our knowledge, there is a lack of rigorous convergence analysis for the TT factorization approaches.

3

Challenges: One of the main challenges in studying the convergence analysis of iterative algorithms for the factorization approach lies in the form of products among multiple matrices in (1). For instance, the TT factorization is not unique, and there exist infinitely many equivalent factorizations. In particular, for any factorization $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, $\{\boldsymbol{X}_1\boldsymbol{P}_1, \boldsymbol{P}_1^{-1}\boldsymbol{X}_2\boldsymbol{P}_2, \ldots, \boldsymbol{P}_{N-1}^{-1}\boldsymbol{X}_N\}$ is also a TT factorization of $\mathcal{X}$ for any invertible matrices $\boldsymbol{P}_i \in \mathbb{R}^{r_i \times r_i}, i \in [N-1]$, where $[N-1] = \{1, \ldots, N-1\}$ and $\boldsymbol{P}_{i-1}^{-1}\boldsymbol{X}_i\boldsymbol{P}_i$ refers to $\boldsymbol{P}_{i-1}^{-1}\boldsymbol{X}_i(:, s_i, :)\boldsymbol{P}_i$ for all $s_i \in [d_i]$. This implies that the factors could be unbalanced (e.g., $\boldsymbol{P}_i = t\mathbf{I}$ with either very large or small t), which makes the convergence analysis difficult. In matrix factorization, a regularizer is often used to address this scaling ambiguity (i.e., to reduce the search space of factors) and balance the energy of the two factors (Tu et al., 2016; Park et al., 2017; Zhu et al., 2018). Motivated by these results, one may adopt the same trick by adding regularizers to balance any pair of consecutive factors $(\boldsymbol{X}_i, \boldsymbol{X}_{i+1})$. However, this strategy could be intricate, given that modifying $\boldsymbol{X}_i$ to achieve a balance between the pair $(\boldsymbol{X}_i, \boldsymbol{X}_{i+1})$ will similarly impact another pair, namely, $(\boldsymbol{X}_{i-1}, \boldsymbol{X}_i)$.

Our contributions: In this paper, we study the factorization approach for the TT sensing problem, where the goal is to recover the underlying low-TT-rank tensor $\mathcal{X}^\star$ through its linear measurements $\boldsymbol{y} = \mathcal{A}(\mathcal{X}^\star)$, where the linear mapping $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ denotes the sensing operator. To address the ambiguity issue in the factorization approach, we consider the so-called left-canonical TT format that restricts all of the factors except the last one to be orthonormal, i.e., $\sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}, i \in [N-1]$. Further details on the left-canonical form are described in Section 3. For a collection of factors $\{\boldsymbol{X}_i\}$, to simplify the notation, we will denote by $[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ the corresponding TT format tensor $\mathcal{X}$ with entries expressed in (1). We then attempt to recover the underlying low-TT-rank tensor by solving the following TT factorized optimization problem

$$\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}, \\ i \in [N]}} \frac{1}{2m}\|\mathcal{A}([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]) - \boldsymbol{y}\|_2^2,$$

$$\text{s.t.} \sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}, i \in [N-1]. \tag{2}$$

Noting that each constraint defines a Stiefel manifold, to guarantee the exact preservation of the orthonormal structure in each iteration, we propose a (hybrid) Riemannian gradient descent (RGD) algorithm to solve the above TT factorization problem. Our main contribution focuses on the convergence analysis of RGD for solving this problem.

- We first study the TT factorization problem where $\mathcal{A}$ is an identity operator. With an appropriate distance metric on the factors, we establish the local linear convergence of RGD. Notably, the accuracy requirement on the initialization only depends polynomially on the tensor order N and the rate of convergence only experiences a linear decline as N increases. This demonstrates potential advantages over introducing additional regularizers to enforce orthogonality for each factor, as used in (Han et al., 2022) for the Tucker factorization, which only ensures approximate orthogonality in each iteration of gradient descent and thus is likely to suffer from exponential dependence on the tensor order N.

- We then extend the convergence analysis to the more general TT sensing problem. Under the assumption that the sensing operator satisfies the restricted isometry property (RIP)—a

condition that can be satisfied with $m \gtrsim N\bar{d}\bar{r}^2 \log(N\bar{r})$ generic subgaussian measurements (Rauhut et al., 2017; Qin et al., 2024) (where, again, $\bar{r} = \max_i r_i$ and $\bar{d} = \max_i d_i$)—we show that RGD, given an appropriate initialization, converges to the ground-truth tensor at a linear rate. Additionally, spectral initialization provides a valid starting point for ensuring the convergence of RGD. Furthermore, we expand our analysis to noisy measurements and prove that RGD can reliably recover the ground truth at a linear rate up to an error proportional to the noise level and exhibiting only polynomial growth in the tensor order N.

Paper organization   The rest of this paper is organized as follows. In Section 2, we introduce the basic definitions of the TT format. Section 3 and Section 4 analyze the local convergence of RGD for the TT factorization and sensing problems, respectively. Section 5 presents numerical experiments. Lastly, we conclude the paper in Section 6.

Notations   We use calligraphic letters (e.g., $\mathcal{Y}$) to denote tensors, bold capital letters (e.g., $\boldsymbol{Y}$) to denote matrices, except for $\boldsymbol{X}_i$ which denotes the i-th order-3 tensor factor in the TT format, bold lowercase letters (e.g., $\boldsymbol{y}$) to denote vectors, and italic letters (e.g., y) to denote scalar quantities. $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_F$ respectively represent the spectral norm and Frobenius norm of the matrix $\boldsymbol{X}$, while $\sigma_i(\boldsymbol{X})$ is the i-th singular value of $\boldsymbol{X}$. $\|\boldsymbol{x}\|_2$ denotes the l_2 norm of the vector $\boldsymbol{x}$. Elements of matrices and tensors are denoted in parentheses, as in Matlab notation. For example, $\mathcal{X}(s_1, s_2, s_3)$ denotes the element in position (s_1, s_2, s_3) of the order-3 tensor $\mathcal{X}$. The inner product of $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ can be denoted as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{s_1=1}^{d_1} \cdots \sum_{s_N=1}^{d_N} \mathcal{A}(s_1, \ldots, s_N)\mathcal{B}(s_1, \ldots, s_N)$. The vectorization of $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, denoted as $\text{vec}(\mathcal{X})$, transforms the tensor $\mathcal{X}$ into a vector. The $(s_1, \ldots, s_N)$-th element of $\mathcal{X}$ can be found in the vector $\text{vec}(\mathcal{X})$ at the position $s_1 + d_1(s_2 - 1) + \cdots + d_1 d_2 \cdots d_{N-1}(s_N - 1)$. $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$ is the Frobenius norm of $\mathcal{X}$. $\overline{\otimes}$ denotes the Kronecker product between submatrices in two block matrices. Its detailed definition and properties are shown in Appendix A. For a positive integer K, $[K]$ denotes the set $\{1, \ldots, K\}$. For two positive quantities $a, b \in \mathbb{R}$, the inequality $b \lesssim a$ or $b = O(a)$ means $b \leq ca$ for some universal constant c; likewise, $b \gtrsim a$ or $b = \Omega(a)$ indicates that $b \geq ca$ for some universal constant c.

2. Preliminaries of Tensor Train Decomposition and Stiefel Manifold

2.1 Tensor Train Decomposition

Recall the TT format of $\mathcal{X}$ in (1). Since $\boldsymbol{X}_i(:, s_i, :)$ will be extensively used, we will denote it by $\boldsymbol{X}_i(s_i) \in \mathbb{R}^{r_{i-1} \times r_i}$; this matrix comprises one "slice" of $\boldsymbol{X}_i$ with the second index being fixed at s_i. The $(s_1, \ldots, s_N)$-th element in $\mathcal{X}$ can then be written as $\mathcal{X}(s_1, \ldots, s_N) = \prod_{i=1}^{N} \boldsymbol{X}_i(s_i)$.

   In addition, for any two TT format tensors $\widetilde{\mathcal{X}}, \widehat{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ with factors $\{\widetilde{\boldsymbol{X}}_i(s_i) \in \mathbb{R}^{\widetilde{r}_{i-1} \times \widetilde{r}_i}\}$ and $\{\widehat{\boldsymbol{X}}_i(s_i) \in \mathbb{R}^{\widehat{r}_{i-1} \times \widehat{r}_i}\}$, each element of the summation $\mathcal{X} = \widetilde{\mathcal{X}} + \widehat{\mathcal{X}}$ can be represented by

$$\mathcal{X}(s_1, \ldots, s_N) = \begin{bmatrix} \widetilde{\boldsymbol{X}}_1(s_1) & \widehat{\boldsymbol{X}}_1(s_1) \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{X}}_2(s_2) & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{X}}_2(s_2) \end{bmatrix} \cdots \begin{bmatrix} \widetilde{\boldsymbol{X}}_{N-1}(s_{N-1}) & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{X}}_{N-1}(s_{N-1}) \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{X}}_N(s_N) \\ \widehat{\boldsymbol{X}}_N(s_N) \end{bmatrix},$$
$$(3)$$

 which implies that $\mathcal{X}$ can also be represented in the TT format with ranks $r_i \leq \widetilde{r}_i + \widehat{r}_i$ for $i = 1, \ldots, N-1$.

Canonical form   The decomposition of a tensor $\mathcal{X}$ into the form (1) is generally not unique: not only are the factors $\boldsymbol{X}_i(s_i)$ not unique, but also the dimensions of these factors can vary. To intro-

duce the factorization with the smallest possible dimensions $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$, for convenience, for each i, we put $\{\boldsymbol{X}_i(s_i)\}_{s_i=1}^{d_i}$ together into the following two forms

$$L(\boldsymbol{X}_i) = \begin{bmatrix} \boldsymbol{X}_i(1) \\ \vdots \\ \boldsymbol{X}_i(d_i) \end{bmatrix} \in \mathbb{R}^{(r_{i-1}d_i) \times r_i},$$

$$R(\boldsymbol{X}_i) = \begin{bmatrix} \boldsymbol{X}_i(1) & \cdots & \boldsymbol{X}_i(d_i) \end{bmatrix} \in \mathbb{R}^{r_{i-1} \times (d_i r_i)},$$

where $L(\boldsymbol{X}_i)$ and $R(\boldsymbol{X}_i)$ are often called the left and right unfoldings of $\boldsymbol{X}_i$, respectively, if we view $\boldsymbol{X}_i$ as a tensor. We say the decomposition (1) is *minimal* if the rank of the left unfolding matrix $L(\boldsymbol{X}_i)$ is r_i and the rank of the right unfolding matrix $R(\boldsymbol{X}_i)$ is r_{i-1} for all i. The dimensions $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$ of such a minimal decomposition are called the *TT ranks* of $\mathcal{X}$. To simplify the notation and presentation, we may also refer to $\bar{r} = \max_i r_i$ as the *TT rank*. According to (Holtz et al., 2012), there is exactly one set of ranks $\boldsymbol{r}$ that $\mathcal{X}$ admits a minimal TT decomposition.

Under the minimal decomposition, there always exists a factorization such that $L(\boldsymbol{X}_i)$ are orthonormal matrices for all $i \in [N-1]$:

$$L^\top(\boldsymbol{X}_i)L(\boldsymbol{X}_i) = \boldsymbol{I}_{r_i}, \ \forall i = 1, \ldots, N-1. \tag{4}$$

Such a decomposition is unique up to the insertion of orthonormal matrices between adjacent factors (Holtz et al., 2012, Theorem 1). That is, $\Pi_{i=1}^N \boldsymbol{X}_i(s_i) = \Pi_{i=1}^N \boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i(s_i) \boldsymbol{R}_i$ for any orthonormal matrix $\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}$ (with $\boldsymbol{R}_0 = \boldsymbol{R}_N = 1$). The resulting TT factors $\{\boldsymbol{X}_i\}$ or the TT decomposition is called the *left-orthogonal form*, or *left-canonical form*. Similarly, the decomposition is said to be right-orthogonal if $R(\boldsymbol{X}_i)$ satisfies $R(\boldsymbol{X}_i)R^\top(\boldsymbol{X}_i) = \boldsymbol{I}_{r_{i-1}}, \ \forall i = 2, \ldots, N$. Since the two forms are equivalent (Holtz et al., 2012), in this paper, we always focus on the left-orthogonal form. Unless otherwise specified, we will always assume that the factors are in left-orthogonal form.

Moreover, r_i also relates to the rank of the i-th unfolding matrix[2] $\mathcal{X}^{\langle i \rangle} \in \mathbb{R}^{(d_1 \cdots d_i) \times (d_{i+1} \cdots d_N)}$ of the tensor $\mathcal{X}$, where the $(s_1 \cdots s_i, s_{i+1} \cdots s_N)$-th element[3] of $\mathcal{X}^{\langle i \rangle}$ is given by

$$\mathcal{X}^{\langle i \rangle}(s_1 \cdots s_i, s_{i+1} \cdots s_N) = \mathcal{X}(s_1, \ldots, s_N).$$

This can also serve as an alternative way to define the TT ranks. With the i-th unfolding matrix $\mathcal{X}^{\langle i \rangle}$ and the TT ranks, we can obtain its smallest singular value $\underline{\sigma}(\mathcal{X}) = \min_{i=1}^{N-1} \sigma_{r_i}(\mathcal{X}^{\langle i \rangle})$, its largest singular value $\bar{\sigma}(\mathcal{X}) = \max_{i=1}^{N-1} \sigma_1(\mathcal{X}^{\langle i \rangle})$ and its condition number $\kappa(\mathcal{X}) = \frac{\bar{\sigma}(\mathcal{X})}{\underline{\sigma}(\mathcal{X})}$.

---

2. We can also define the i-th unfolding matrix as $\mathcal{X}^{\langle i \rangle} = \boldsymbol{X}^{\leq i} \boldsymbol{X}^{\geq i+1}$, where each row of the left part $\boldsymbol{X}^{\leq i}$ and each column of the right part $\boldsymbol{X}^{\geq i+1}$ can be represented as

$$\boldsymbol{X}^{\leq i}(s_1 \cdots s_i, :) = \boldsymbol{X}_1(s_1) \cdots \boldsymbol{X}_i(s_i),$$

$$\boldsymbol{X}^{\geq i+1}(:, s_{i+1} \cdots s_N) = \boldsymbol{X}_{i+1}(s_{i+1}) \cdots \boldsymbol{X}_N(s_N).$$

Note that when the factors are in the left-orthogonal form, we have $(\boldsymbol{X}^{\leq i})^\top \boldsymbol{X}^{\leq i} = \boldsymbol{I}_{r_i}$. Similarly, for the right-orthogonal form, $\boldsymbol{X}^{\geq i+1}(\boldsymbol{X}^{\geq i+1})^\top = \boldsymbol{I}_{r_i}$.

3. Specifically, $s_1 \cdots s_i$ and $s_{i+1} \cdots s_N$ respectively represent the $(s_1 + d_1(s_2 - 1) + \cdots + d_1 \cdots d_{i-1}(s_i - 1))$-th row and $(s_{i+1} + d_{i+1}(s_{i+2} - 1) + \cdots + d_{i+1} \cdots d_{N-1}(s_N - 1))$-th column.

Distance between factors We now introduce an appropriate metric to quantify the distinctions between the left-orthogonal form factors $\{\boldsymbol{X}_i\}$ and $\{\boldsymbol{X}_i^\star\}$ of two TT format tensors $\mathcal{X} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_N]$ and $\mathcal{X}^\star = [\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_N^\star]$.

To capture this rotation ambiguity, by defining the rotated factors $L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)$ as

$$L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star) = \begin{bmatrix} \boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i^\star(1) \boldsymbol{R}_i \\ \vdots \\ \boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i^\star(d_i) \boldsymbol{R}_i \end{bmatrix}, \tag{5}$$

we then define the distance between the two sets of factors as

$$\text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) = \min_{\substack{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}, \\ i \in [N-1]}} \sum_{i=1}^{N-1} \overline{\sigma}^2(\mathcal{X}^\star) \|L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 + \|L(\boldsymbol{X}_N) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2, \tag{6}$$

where we note that $L(\boldsymbol{X}_N), L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star) \in \mathbb{R}^{(r_{N-1} d_N) \times 1}$ are vectors. Here, the coefficients $\overline{\sigma}^2(\mathcal{X}^\star)$ and 1 are incorporated to harmonize the energy between $\{L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\}_{i \leq N-1}$ and $L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)$ since $\|L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|^2 = 1, i \in [N-1]$ and $\|L(\boldsymbol{X}_N) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2 = \|R(\boldsymbol{X}_N) - \boldsymbol{R}_{N-1}^\top R(\boldsymbol{X}_N^\star)\|_F^2$, $\|\boldsymbol{R}_{N-1}^\top R(\boldsymbol{X}_N^\star)\|^2 = \|R(\boldsymbol{X}_N^\star)\|^2 = \sigma_1^2(\mathcal{X}^{\star\langle N-1 \rangle}) \leq \overline{\sigma}^2(\mathcal{X}^\star)$. The following result establishes a connection between $\text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\})$ and $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$.

Lemma 1 *For any two TT format tensors $\mathcal{X}$ and $\mathcal{X}^\star$ with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$, let $\{\boldsymbol{X}_i\}$ and $\{\boldsymbol{X}_i^\star\}$ be the corresponding left-orthogonal form factors. Assume $\overline{\sigma}^2(\mathcal{X}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$. Then $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$ and $\text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\})$ defined in (6) satisfy*

$$\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \geq \frac{1}{8(N + 1 + \sum_{i=2}^{N-1} r_i) \kappa^2(\mathcal{X}^\star)} \text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}), \tag{7}$$

$$\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \leq \frac{9N}{4} \text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}). \tag{8}$$

The proof is given in Appendix A. Lemma 1 ensures that $\mathcal{X}$ is close to $\mathcal{X}^\star$ once the corresponding factors are close with respect to the proposed distance measure, and the convergence behavior of $\|\mathcal{X} - \mathcal{X}^\star\|_F^2$ is reflected by the convergence in terms of the factors. In the next sections, we will study the convergence with respect to the factors.

2.2 Stiefel Manifold

Since we will focus on the left-canonical form where the left unfolding matrices of a TT factorization are orthonormal, i.e., reside on the Stiefel manifold, we will introduce several essential definitions concerning the Stiefel manifold and its tangent space to clarify our discussion of optimization on the Stiefel manifold. The Stiefel manifold $\text{St}(m, n) = \{\boldsymbol{Y} \in \mathbb{R}^{m \times n} : \boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{I}_n\}$ is a Riemannian manifold that is composed of all $m \times n$ orthonormal matrices. We can regard $\text{St}(m, n)$ as an embedded submanifold of a Euclidean space and further define $T_{\boldsymbol{Y}} \text{St} := \{\boldsymbol{A} \in \mathbb{R}^{m \times n} : \boldsymbol{A}^\top \boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{A} = \boldsymbol{0}\}$ as the tangent space to the Stiefel manifold $\text{St}(m, n)$ at the point $\boldsymbol{Y} \in \text{St}(m, n)$. For any $\boldsymbol{B} \in \mathbb{R}^{m \times n}$, the projection of $\boldsymbol{B}$ onto $T_{\boldsymbol{Y}} \text{St}$ is given by (Absil et al., 2008)

$$\mathcal{P}_{T_{\boldsymbol{Y}} \text{St}}(\boldsymbol{B}) = \boldsymbol{B} - \frac{1}{2} \boldsymbol{Y} \left(\boldsymbol{B}^\top \boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{B} \right), \tag{9}$$

and the projection of B onto the orthogonal complement of $\mathrm{T}_Y\,\mathrm{St}$ is given by

$$\mathcal{P}^{\perp}_{\mathrm{T}_Y\,\mathrm{St}}(B) = B - \mathcal{P}_{\mathrm{T}_Y\,\mathrm{St}}(B) = \frac{1}{2}Y\left(B^{\top}Y + Y^{\top}B\right). \tag{10}$$

Note that when we have a gradient B, we can use the projection operator (9) to compute the Riemannian gradient $\mathcal{P}_{\mathrm{T}_Y\,\mathrm{St}}(B)$ on the tangent space of the Stiefel manifold. Riemannian gradient descent involves the update $\widehat{Y} = Y - \mu \mathcal{P}_{\mathrm{T}_Y\,\mathrm{St}}(B)$ with a step size $\mu > 0$, which is then projected back to the Stiefel manifold, such as via the polar decomposition-based retraction, i.e.,

$$\mathrm{Retr}_Y(\widehat{Y}) = \widehat{Y}(\widehat{Y}^{\top}\widehat{Y})^{-\frac{1}{2}}. \tag{11}$$

3. Warm-up: Low-rank Tensor-train Factorization

To provide a baseline for the convergence of iterative algorithms for the TT recovery problem with the factorization approach in (2), we first study the following TT factorization problem

$$\min_{\substack{X_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ i \in [N]}} f(X_1,\ldots,X_N) = \frac{1}{2}\|[X_1,\ldots,X_N] - \mathcal{X}^{\star}\|_F^2, \tag{12}$$

$$\mathrm{s.\,t.}\ \ L^{\top}(X_i)L(X_i) = I_{r_i}, i \in [N-1],$$

Except for the scaling difference in the object function, the above problem is a special case of (2), where the operator $\mathcal{A}$ is the identity operator, and thus the convergence analysis for (12) will provide useful insight for the problem (2). We will analyze the local convergence of Riemannian gradient descent (RGD) to solve the factorization problem (12) and explore how the convergence speed and requirements depend on the properties of $\mathcal{X}^{\star}$, such as the tensor order. We will then extend the analysis to the sensing problem (2) in the next section.

Specifically, we utilize the following (hybrid) RGD

$$L(X_i^{(t+1)}) = \mathrm{Retr}_{L(X_i)}\left(L(X_i^{(t)}) - \frac{\mu}{\overline{\sigma}^2(\mathcal{X}^{\star})}\mathcal{P}_{\mathrm{T}_{L(X_i)}\mathrm{St}}\big(\nabla_{L(X_i)}f(X_1^{(t)},\ldots,X_N^{(t)})\big)\right), i \in [N-1], \tag{13}$$

$$L(X_N^{(t+1)}) = L(X_N^{(t)}) - \mu\nabla_{L(X_N)}f(X_1^{(t)},\ldots,X_N^{(t)}), \tag{14}$$

where $\mathcal{P}_{\mathrm{T}_{L(X_i)}\mathrm{St}}$ denotes the projection onto the tangent space of the Stiefel manifold at the point $L(X_i)$, as defined in (9), such that $\mathcal{P}_{\mathrm{T}_{L(X_i)}\mathrm{St}}\big(\nabla_{L(X_i)}f(X_1^{(t)},\ldots,X_N^{(t)})\big)$ is the Riemannian gradient of the objective function f with respect to the i-th factors $L(X_i)$. The detailed expression of the gradients $\nabla_{L(X_i)}f(X_1^{(t)},\ldots,X_N^{(t)})$ is presented in Appendix B (see (73)). We update the factor X_N using gradient descent in (14) since there is no constraint on this factor. For simplicity, we still refer to the updates in (13) and (14) as RGD. Note that we use discrepant step sizes between $L(X_i)$ and $L(X_N)$ in the proposed RGD algorithm in order to balance the convergence of those factors as they have different energies; $\|L(X_i)\|^2 = 1, i \in [N-1]$ and $\|R(X_N)\|^2 = \sigma_1^2(\mathcal{X}^{\langle N-1\rangle}) \leq \overline{\sigma}^2(\mathcal{X})$ in each iteration. To simplify the analysis, we employ a discrepant learning rate ratio, i.e., $\overline{\sigma}^2(\mathcal{X}^{\star})$, to balance the two sets of factors. We note that other choices of discrepant learning rate ratio are also effective in practice, such as $\|\mathcal{X}^{(t)}\|_F^2$, which can be efficiently computed for TT-format tensors (Oseledets, 2011). In addition, when the ground-truth tensor $\mathcal{X}^{\star}$ is unknown a prior in practice, we can instead use the information of either $\mathcal{X}^{(0)}$ or $\mathcal{X}^{(t)}$ to balance the learning rate, given that the iterates will remain close to the target tensor as guaranteed by the following analysis.

Local convergence of RGD algorithm According to Section 2, each TT format tensor has an equivalent left-orthogonal form. Let $\{\boldsymbol{X}_i^\star\}$ be the left-orthogonal form factors for a minimal TT decomposition of $\mathcal{X}^\star$. These factors are utilized exclusively for the analysis and are not required for the algorithm. Recall the distance between the two set of factors $\{\boldsymbol{X}_i\}$ and $\{\boldsymbol{X}_i^\star\}$ as given in (6).

We now establish a local linear convergence guarantee for the RGD algorithm in (13) and (14).

Theorem 2 *Consider a low-TT-rank tensor $\mathcal{X}^\star$ with ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$. Suppose that the RGD in (13) and (14) is initialized with $\{\boldsymbol{X}_i^{(0)}\}$ satisfying*

$$\text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{72(N^2-1)(N+1+\sum_{i=2}^{N-1} r_i)}, \tag{15}$$

and the step size $\mu \leq \frac{1}{9N-5}$. Then, the iterates $\{\boldsymbol{X}_i^{(t)}\}_{t\geq 0}$ generated by RGD will converge linearly to $\{\boldsymbol{X}_i^\star\}$ (up to rotation):

$$\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq \left(1 - \frac{1}{64(N+1+\sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)}\mu\right) \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}). \tag{16}$$

Note that Theorem 2 only establishes local convergence. Since our primary objective is to gain insight into local convergence, and initialization is not our focus, we will omit discussions related to obtaining a valid initialization for this factorization problem. When we address the TT sensing problem in the next section, we will present approaches for finding a suitable initialization. Due to the presence of non-global critical points, linear convergence for first-order methods is likely to be attainable only within a certain region. Moreover, products of multiple (more than two) matrices often lead to the emergence of high-order saddle points or even spurious local minima that are distant from the global minima (Vidal et al., 2022). Relying only on the gradient information is not sufficient to circumvent these high-order saddle points. Therefore, this paper primarily focuses on local convergence.

Remarkably, both terms $O(\frac{\underline{\sigma}^2(\mathcal{X}^\star)}{N^3\bar{r}})$ and $O(\frac{1}{N^2\bar{r}\kappa^2(\mathcal{X}^\star)})$ in the initialization requirement (15) and convergence rate (16) only decay polynomially rather than exponentially in terms of the tensor order N. The detailed proof of Theorem 2 is provided in Appendix B. Below, we provide a high-level overview of the proof.

Proof sketch We focus on establishing an error contraction inequality that characterizes the error $\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$ based on the previous iterate. Utilizing the error metric defined in (6), we define the best rotation matrices to align $\{\boldsymbol{X}_i^{(t)}\}$ and $\{\boldsymbol{X}_i^\star\}$ as

$$(\boldsymbol{R}_1^{(t)}, \ldots, \boldsymbol{R}_{N-1}^{(t)}) = \underset{\substack{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}, \\ i \in [N-1]}}{\arg\min} \sum_{i=1}^{N-1} \overline{\sigma}^2(\mathcal{X}^\star) \|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 + \|L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2, \tag{17}$$

where $L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)$ is defined in (5). We now expand $\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$ as

$$
\begin{aligned}
&\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \\
&= \sum_{i=1}^N \gamma_i \left\| L(\boldsymbol{X}_i^{(t+1)}) - L_{\boldsymbol{R}^{(t+1)}}(\boldsymbol{X}_i^\star) \right\|_F^2 \leq \sum_{i=1}^N \gamma_i \left\| L(\boldsymbol{X}_i^{(t+1)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star) \right\|_F^2 \\
&\leq \sum_{i=1}^N \gamma_i \left\| L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star) - \frac{\mu}{\gamma_i} \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 \\
&= \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) + \sum_{i=1}^N \frac{\mu^2}{\gamma_i} \left\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 \\
&\quad - 2\mu \sum_{i=1}^N \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\rangle, \qquad (18)
\end{aligned}
$$

where to simplify the expression, we define $\gamma_i = \begin{cases} \bar{\sigma}^2(\mathcal{X}^\star), & i \in [N-1] \\ 1, & i = N \end{cases}$ and a projection operator for the last factor as $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_N)}\mathrm{St}} = \mathcal{I}$ such that $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_N)}\mathrm{St}}(\nabla_{L(\boldsymbol{X}_N)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})) = \nabla_{L(\boldsymbol{X}_N)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})$. We note that the second inequality follows from the nonexpansiveness property described in Lemma 17 of Appendix A.

The remainder of the proof is to quantify the last two terms in (18) to ensure the decay of the distance. On the one hand, we can obtain an upper bound of the second term in (18) as

$$
\sum_{i=1}^N \frac{1}{\gamma_i} \left\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 \leq \frac{9N-5}{4} \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2, \qquad (19)
$$

which ensures a bounded Riemannian gradient when the iterates converge to the target solution. On the other hand, under the assumption that $\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\sigma^2(\mathcal{X}^\star)}{72(N^2-1)(N+1+\sum_{i=2}^{N-1} r_i)}$, we can lower bound the third term in (18) as

$$
\begin{aligned}
&\sum_{i=1}^N \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\rangle \\
&\geq \frac{1}{128(N+1+\sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)} \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) + \frac{1}{8} \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2, \qquad (20)
\end{aligned}
$$

which implies that the negative direction of the Riemannian gradient points toward the optimal factors.

Plugging (20) and (19) into (18) yields the convergence of the factors

$$
\begin{aligned}
\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) &\leq \left(1 - \frac{1}{64(N+1+\sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)}\mu\right) \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\quad + \left(\frac{9N-5}{4}\mu^2 - \frac{\mu}{4}\right) \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 \\
&\leq \left(1 - \frac{1}{64(N+1+\sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)}\mu\right) \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), (21)
\end{aligned}
$$

where the last line uses the fact that the step size $\mu \leq \frac{1}{9N-5}$.

Connection to matrix and Tucker tensor factorization approaches   There have been numerous studies on nonconvex matrix estimation (Tu et al., 2016; Wang et al., 2017; Jin et al., 2017; Zhu et al., 2018; Li et al., 2020b; Ma et al., 2021; Tong et al., 2021a) and Tucker tensor estimation (Tong et al., 2021b; Xia and Yuan, 2019; Han et al., 2022). However, we note that most of the theoretical analyses developed for the matrix case cannot be directly extended to the TT factorization approach since the orthonormal constraint is not applied there. On the other hand, the highly unbalanced nature of orthonormal matrices and a core tensor in the Tucker tensor make it more likely that the theoretical analysis in the Tucker factorization estimation can be applied to the TT factorization estimation. In the Tucker tensor estimation, the introduction of an approximately orthonormal structure has led to the development of a regularized gradient descent algorithm (Han et al., 2022), which has been demonstrated effectively to achieve a linear convergence rate. However, the results presented in (Han et al., 2022) primarily pertain to order-3 Tucker tensors. When extended to high-order tensors, both the theoretical convergence rate and the initial conditions are likely to deteriorate exponentially with respect to the order N. One reason for this deterioration is that the factor matrices are only guaranteed to be approximately rather than exactly orthogonal in each iteration, which may lead to a high condition number of the product of multiple approximately orthogonal matrices in the theoretical analysis. This issue is addressed in our approach by strictly enforcing orthonormality of the factors.

4. Low-rank Tensor-train Sensing

In this section, we consider the problem of recovering a low-TT-rank tensor $\mathcal{X}^\star$ from its linear measurements

$$\boldsymbol{y} = \mathcal{A}(\mathcal{X}^\star) = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \langle \mathcal{A}_1, \mathcal{X}^\star \rangle \\ \vdots \\ \langle \mathcal{A}_m, \mathcal{X}^\star \rangle \end{bmatrix} \in \mathbb{R}^m, \tag{22}$$

where $\mathcal{A}(\mathcal{X}^\star) : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ is a linear map modeling the measurement process. This problem appears in many applications such as quantum state tomography (Lidiak et al., 2022; Qin et al., 2024), neuroimaging analysis (Zhou et al., 2013; Li and Zhang, 2017), 3D imaging (Guo et al., 2011), high-order interaction pursuit (Hao et al., 2020), and more.

To enable the recovery of the low-TT-rank tensor $\mathcal{X}^\star$ from its linear measurements, the sensing operator is required to satisfy certain properties. One desirable property is the following Restricted Isometry Property (RIP), which has been widely studied and popularized in the compressive sensing literature (Donoho, 2006; Candès et al., 2006; Candès and Wakin, 2008; Recht et al., 2010), and has been extended for structured tensors (Grotheer et al., 2021; Rauhut et al., 2017; Qin et al., 2024).

Definition 3 *(Restricted Isometry Property (Rauhut et al., 2017)) A linear operator* $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N}$ $\to \mathbb{R}^m$ *is said to satisfy the* $\bar{r}$*-restricted isometry property ($\bar{r}$-RIP) with constant* $\delta_{\bar{r}}$ *if*

$$(1 - \delta_{\bar{r}})\|\mathcal{X}\|_F^2 \leq \frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_2^2 \leq (1 + \delta_{\bar{r}})\|\mathcal{X}\|_F^2, \tag{23}$$

holds for any TT format tensors $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ *with TT ranks* $\boldsymbol{r} = (r_1, \ldots, r_{N-1}), r_i \leq \bar{r}$.

In words, the RIP ensures a stable embedding for TT format tensors and guarantees that the energy $\|\mathcal{A}(\mathcal{X})\|_2^2$ is proportional to $\|\mathcal{X}\|_F^2$. The RIP can often be attained by randomly selecting measurement operators from a specific distribution, with subgaussian measurement ensembles serving as a common example.

Definition 4 *(Subgaussian measurement ensembles (Biermé and Lacaux, 2015)) A real random variable X is called L-subgaussian if there exists a constant $L > 0$ such that $\mathbb{E}\, e^{tX} \leq e^{L^2 t^2/2}$ holds for all $t \in \mathbb{R}$. Typical examples include the Gaussian random variable and the Bernoulli random variable. We say that $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ is an L-subgaussian measurement ensemble if all the elements of $\mathcal{A}_k, k = 1, \ldots, m$ are independent L-subgaussian random variables with mean zero and variance one.*

The following result shows that the RIP holds with high probability for L-subgaussian measurement ensembles.

Theorem 5 *((Rauhut et al., 2017, Theorem 4),(Qin et al., 2024, Theorem 2)) Suppose that the linear map $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ is an L-subgaussian measurement ensemble. Let $\delta_{\overline{r}} \in (0, 1)$ denote a positive constant. Then, with probability at least $1 - \epsilon$, $\mathcal{A}$ satisfies the $\overline{r}$-RIP as in (23) for any TT format tensors $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ with TT ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1}), r_i \leq \overline{r}$, given that*

$$m \geq C \cdot \frac{1}{\delta_{\overline{r}}^2} \cdot \max\left\{ N\overline{d}\overline{r}^2 \log(N\overline{r}), \log(1/\epsilon) \right\}, \tag{24}$$

where $\overline{d} = \max_i d_i$ and C is a universal constant depending only on L.

Theorem 5 ensures the RIP for L-subgaussian measurement ensembles with a number of measurements m only scaling linearly, rather than exponentially, with respect to the tensor order N. When the RIP holds, then for any two distinct TT format tensors $\mathcal{X}_1, \mathcal{X}_2$ with TT ranks smaller than $\overline{r}$, noting that $\mathcal{X}_1 - \mathcal{X}_2$ is also a TT format tensor according to (3), we have distinct measurements since

$$\frac{1}{m}\|\mathcal{A}(\mathcal{X}_1) - \mathcal{A}(\mathcal{X}_2)\|_2^2 = \frac{1}{m}\|\mathcal{A}(\mathcal{X}_1 - \mathcal{X}_2)\|_2^2 \geq (1 - \delta_{2\overline{r}})\|\mathcal{X}_1 - \mathcal{X}_2\|_F^2,$$

which guarantees the possibility of exact recovery. We will now examine the convergence of RGD for solving the factorized optimization problem given by (2).

4.1 Exact Recovery with Linear Convergence

We start by restating the factorized approach in (2) that minimizes the discrepancy between the measurements $\boldsymbol{y}$ and the linear mapping of the estimated low-TT-rank tensor $\mathcal{X}$ as

$$\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}, \\ i \in [N]}} g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) = \frac{1}{2m}\|\mathcal{A}([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]) - \boldsymbol{y}\|_2^2, \tag{25}$$

$$\text{s.\,t.}\;\; L^\top(\boldsymbol{X}_i)L(\boldsymbol{X}_i) = \mathbf{I}_{r_i}, i \in [N-1].$$

As for (25), we solve the above optimization problem over the Stiefel manifold by the following RGD:

$$L(\boldsymbol{X}_i^{(t+1)}) = \text{Retr}_{L(\boldsymbol{X}_i)}\left(L(\boldsymbol{X}_i^{(t)}) - \frac{\mu}{\overline{\sigma}^2(\mathcal{X}^\star)}\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}}\big(\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\big)\right), i \in [N-1], \tag{26}$$

$$L(\boldsymbol{X}_N^{(t+1)}) = L(\boldsymbol{X}_N^{(t)}) - \mu\nabla_{L(\boldsymbol{X}_N)}g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}), \tag{27}$$

where expressions for the gradients are given in Appendix D (see (93)). As discussed in Section 3, our primary focus lies in examining the local convergence of the RGD algorithm. Before presenting the local convergence, we first discuss and analyze a spectral initialization approach, which serves as an initial value for the optimization algorithm.

Spectral Initialization   To provide a good initialization for the RGD algorithm, we apply the following spectral initialization method:

$$\mathcal{X}^{(0)} = \text{SVD}_r^{tt}\left(\frac{1}{m}\sum_{k=1}^{m} y_k \mathcal{A}_k\right), \tag{28}$$

where $\text{SVD}_r^{tt}(\cdot)$ is the TT-SVD algorithm (Oseledets, 2011) that can efficiently find an approximately optimal TT approximation to any tensor. This spectral initialization approach has been widely employed for various inverse problems (Lu and Li, 2020), such as phase retrieval (Candès et al., 2015; Luo et al., 2019), low-rank matrix recovery (Ma et al., 2021; Tong et al., 2021a), and structured tensor recovery (Cai et al., 2019; Tong et al., 2021b; Han et al., 2022). Here, when $\mathcal{A}$ satisfies the RIP condition, we can ensure that the initialization $\mathcal{X}^{(0)}$ is close to $\mathcal{X}^\star$.

Theorem 6 *When $\mathcal{A}$ satisfies the $3\bar{r}$-RIP for TT format tensors with a constant $\delta_{3\bar{r}}$, the spectral initialization generated by* (28) *satisfies*

$$\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \le \delta_{3\bar{r}}(1 + \sqrt{N-1})\|\mathcal{X}^\star\|_F. \tag{29}$$

The proof is provided in Appendix C. Referring to Theorem 6, it can be observed that by ensuring a sufficiently small $\delta_{3\bar{r}}$, we can always find a suitable initialization within a desired distance to the ground truth.

Exact recovery with linear convergence of RGD   Again, our analysis utilizes the left-orthogonal form factors $\{X_i^\star\}$ of a minimal TT decomposition of $\mathcal{X}^\star$, although the factors are not required for implementing the RGD algorithm in (26) and (27). We now establish a local linear convergence guarantee for RGD.

Theorem 7 *Consider a low-TT-rank tensor $\mathcal{X}^\star$ with ranks $r = (r_1, \ldots, r_{N-1})$. Assume $\mathcal{A}$ obeys the $(N+3)\bar{r}$-RIP with a constant $\delta_{(N+3)\bar{r}} \le \frac{4}{15}$ and $\bar{r} = \max_i r_i$. Given $y = \mathcal{A}(\mathcal{X}^\star)$, let $\{X_i^{(t)}\}_{t\ge 0}$ be the iterates generated by the RGD algorithm in (26) and (27). Suppose the algorithm is initialized with $\{X_i^{(0)}\}$ satisfying*

$$\text{dist}^2(\{X_i^{(0)}\}, \{X_i^\star\}) \le \frac{(4 - 15\delta_{(N+3)\bar{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{8(N + 1 + \sum_{i=2}^{N-1} r_i)(57N^2 + 393N - 450)} \tag{30}$$

and uses step size $\mu \le \frac{4-15\delta_{(N+3)\bar{r}}}{10(9N-5)(1+\delta_{(N+3)\bar{r}})^2}$. Then, the iterates $\{X_i^{(t)}\}_{t\ge 0}$ converge linearly to $\{X_i^\star\}$ (up to rotation):

$$\text{dist}^2(\{X_i^{(t+1)}\}, \{X_i^\star\}) \le \left(1 - \frac{4 - 15\delta_{(N+3)\bar{r}}}{320(N + 1 + \sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)}\mu\right)\text{dist}^2(\{X_i^{(t)}\}, \{X_i^\star\}). \tag{31}$$

13

The proof is given in Appendix D. We omit the overview of the proof as it shares a similar structure to the one in Section 3 for the TT factorization problem, with the key difference being the involvement of the sensing operator and the utilization of the RIP. Our results demonstrate that, similar to the findings presented in Theorem 2, when the initial condition $\text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^{\star}\}) \leq O(\frac{\sigma^2(\mathcal{X}^{\star})}{N^3 \overline{r}})$ is satisfied during the initial stage, RGD exhibits a linear convergence rate of $1 - O(\frac{1}{N^2 \overline{r} \kappa^2(\mathcal{X}^{\star})})$. Notably, both the convergence rate and the initial requirement depend on the tensor order N only polynomially rather than exponentially. Note that the RIP is required to hold for TT format tensors with rank $(N+3)\overline{r}$; this is because the analysis involves a summation of $N+3$ TT format tensors of rank $\overline{r}$ as in (18) and (20). However, the requirement of $(N+3)\overline{r}$-RIP may perhaps be improved to $2\overline{r}$-RIP (or higher, such as $4\overline{r}$-RIP, but independent of N) by using an alternative approach to bound the term (97) in Appendix D. A thorough investigation of this improvement is left for future work. To utilize the spectral initialization, we can invoke Lemma 1 to rewrite (30) in terms of the tensors, i.e., $\|\mathcal{X}^{(0)} - \mathcal{X}^{\star}\|_F^2 \leq \frac{(4 - 15\delta_{(N+3)\overline{r}})\sigma^2(\mathcal{X}^{\star})}{64(57N^2 + 393N - 450)(N + 1 + \sum_{i=2}^{N-1} r_i)^2 \kappa^2(\mathcal{X}^{\star})} \lesssim O(\frac{\sigma^2(\mathcal{X}^{\star})}{N^4 \overline{r}^2 \kappa^2(\mathcal{X}^{\star})})$; see Lemma 16 in Appendix A for the details. Thus, Theorem 6 ensures that spectral initialization satisfies the requirement (30), provided that $\delta_{3\overline{r}} \lesssim \frac{\sigma^2(\mathcal{X}^{\star})}{N^{\frac{5}{2}} \overline{r} \|\mathcal{X}^{\star}\|_F^2}$.

Corollary 8 *Consider a low-TT-rank tensor $\mathcal{X}^{\star}$ with ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$. Assume that $\mathcal{A}$ obeys the $(N+3)\overline{r}$-RIP with constants satisfying $\delta_{3\overline{r}} \lesssim \frac{\sigma^2(\mathcal{X}^{\star})}{N^{\frac{5}{2}} \overline{r} \|\mathcal{X}^{\star}\|_F^2}$ and $\delta_{(N+3)\overline{r}} \leq \frac{4}{15}$. When utilizing the spectral initialization and the step size $\mu \leq \frac{4 - 15\delta_{(N+3)\overline{r}}}{10(9N-5)(1 + \delta_{(N+3)\overline{r}})^2}$, RGD converges linearly to a global minimum as in (31).*

When the linear map $\mathcal{A}$ is an L-subgaussian measurement ensemble, using (24), the RIP requirement is satisfied when $m \gtrsim \frac{N^6 \overline{d} \overline{r}^3 \|\mathcal{X}^{\star}\|_F^2 \kappa^2(\mathcal{X}^{\star}) \log(N\overline{r})}{\sigma^2(\mathcal{X}^{\star})}$. This requires the number of measurements m grows only polynomially in N; however, the order is suboptimal due to the following reasons: (i) TT-SVD, used in the spectral initialization, only provides a quasi-optimal low-rank TT approximation—since computing the optimal low-rank TT approximation is NP-hard (Hillar and Lim, 2013), (ii) the requirement of $(N+3)\overline{r}$-RIP in the convergence analysis, and (iii) the relationship between the distances of the factors and of the entire tensors may not be tight in Lemma 1.

Special case: Matrix sensing When $N = 2$, the tensor becomes a matrix and the TT decomposition simplifies to matrix factorization. In this case, the recovery problem reduces to the matrix sensing problem for the rank-r matrix $\boldsymbol{X}^{\star}$. Our Theorem 7 ensures a local linear convergence for RGD with rate of convergence $1 - \frac{(4 - 15\delta_{5r})\sigma_r^2(\boldsymbol{X}^{\star})}{960\sigma_1^2(\boldsymbol{X}^{\star})}\mu$ when $\mathcal{A}$ satisfies the $5r$-RIP (Candès and Plan, 2011) with constant $\delta_{5r} \leq \frac{4}{15}$. Note that the result can be improved to only requiring $3r$-RIP by using Lemma 14 in the analysis of cross terms. In comparison, the work (Tu et al., 2016) establishes local linear convergence for gradient descent (GD) solving a regularized problem with rate of convergence $1 - \frac{4\sigma_r^2(\boldsymbol{X}^{\star})}{25\sigma_1^2(\boldsymbol{X}^{\star})}\mu$, given that the sensing operator $\mathcal{A}$ satisfies the $6r$-RIP with constant $\delta_{6r} \leq \frac{1}{6}$. Without relying on any regularizer to balance the two factors, the work (Ma et al., 2021) also establishes local linear convergence for GD with rate of convergence $(1 - \frac{\sigma_r(\boldsymbol{X}^{\star})}{50\sigma_1(\boldsymbol{X}^{\star})}\mu)^2$ when the sensing operator $\mathcal{A}$ satisfies the $2r$-RIP with constant $\delta_{2r} \leq c_1$ (where c_1 is a constant). We can observe that the convergence guarantee for RGD is similar to that of GD in the context of matrix sensing.
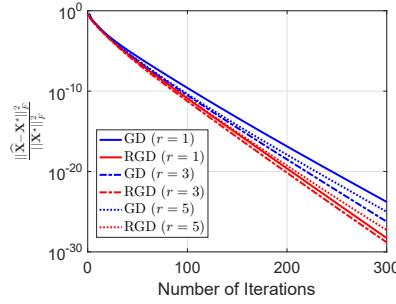
Figure 2: Convergence analysis of GD and RGD with $m = 150r$.

We conduct a numerical experiment to compare the performance of the RGD algorithm with GD (Ma et al., 2021) for matrix sensing. A ground truth matrix $\boldsymbol{X}^\star \in \mathbb{R}^{30 \times 20}$ of rank r is generated by first creating a random Gaussian matrix with i.i.d. entries from a normal distribution, followed by computing a best rank-r approximation. The step size for both the GD and RGD algorithms is set to $\mu = 0.6$. For each experimental setting, we set $m = 150r$, conduct 20 Monte Carlo trials, and then take the average over the 20 trials to report the results. Figure 2 shows that both GD and RGD achieve a similar linear convergence rate.

Comparison with IHT (Rauhut et al., 2017) and Riemannian gradient descent on the fixed-rank manifold (Budzinskiy and Zamarashkin, 2021) To the best of our knowledge, our work is the first to offer a convergence guarantee for directly solving the factorization approach for low-TT-rank in recovery problems. Two iterative algorithms, iterative hard thresholding (IHT) (Rauhut et al., 2017) and RGD on the TT manifold (Budzinskiy and Zamarashkin, 2021), have been studied with local convergence guarantees. Specifically, the IHT algorithm (GD with TT-SVD truncation) has been proven to converge linearly with a rate of convergence $\frac{a}{4}$ ($a \in (0,1)$), but relies on an unverified perturbation bound of TT-SVD, expressed as $\|\text{SVD}_r^{tt}(\mathcal{X}^{(t)}) - \mathcal{X}^{(t)}\|_F \leq (1 + \frac{a^2}{17(1+\sqrt{1+\delta_{3r}}\|\mathcal{A}\|)^2})\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F$ where $\mathcal{X}^{(t)}$ represents the iterate after the gradient update in the t-th iteration. As mentioned in (Rauhut et al., 2017), $\|\mathcal{A}\|$ may increase exponentially in terms of N, imposing a very strong requirement on the optimality of TT-SVD. Viewing all the TT format tensors of fixed rank as an embedded manifold in $\mathbb{R}^{d_1 \times \cdots \times d_N}$ (Holtz et al., 2012), the work (Budzinskiy and Zamarashkin, 2021) introduces an RGD algorithm on the embedded manifold with TT-SVD retraction to approximate the projection of the estimated tensor onto the embedded manifold. It establishes local linear convergence with the rate $(1 + \sqrt{N-1})(\frac{2\delta_{3\bar{r}}}{1-\delta_{3\bar{r}}} + \frac{2}{1-\delta_{3\bar{r}}} \frac{\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F}{\underline{\sigma}(\mathcal{X}^\star)})$. This holds under the conditions $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F^2 \lesssim O(\frac{\sigma^2(\mathcal{X}^\star)}{N^2})$, $3\bar{r}$-RIP with RIP constant $\delta_{3\bar{r}} \leq \frac{1}{3+2\sqrt{N-1}}$, and an unverified condition $\|\mathcal{A}^*\mathcal{A}\| \leq C$ (where C is a constant). Additionally, as mentioned in (Rauhut et al., 2017), the Riemannian gradient in the RGD depends on the curvature information at $\mathcal{X}^\star$ of the embedded manifold, which is often unknown a priori.

Note that both algorithms require the estimation of the entire tensor $\mathcal{X}$ in each iteration and rely on performing the TT-SVD to project the iterates back to the TT format, which demands a substantial amount of storage memory and could be computationally expensive for high-order tensors. In contrast, the factorization approach avoids the need to compute the entire tensor in each iteration.

15

Specifically, we can employ a tensor contraction operation (Cichocki, 2014) to efficiently compute the gradient without explicitly computing the tensor $\mathcal{X}$, such as for the term[4] $\langle \mathcal{A}_k, [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N] \rangle$. Intuitively, this is the same as in the matrix case where we can efficiently compute $\langle \boldsymbol{A}, \boldsymbol{uv}^\top \rangle$ as $\langle \boldsymbol{A}^\top \boldsymbol{u}, \boldsymbol{v} \rangle$ without the need of computing $\boldsymbol{uv}^\top$ for any $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}, \boldsymbol{u} \in \mathbb{R}^{d_1}, \boldsymbol{v} \in \mathbb{R}^{d_2}$.

4.2 Recovery Guarantee for Noisy TT Format Tensor Sensing

In practice, measurements are often noisy, either due to additive noise or the probabilistic nature of the measurements, as seen in quantum state tomography (Qin et al., 2024), which causes statistical error in the measurements and can also be modeled as additive noise. In this subsection, we will extend our analysis to TT sensing with noisy measurements of the form

$$\widehat{\boldsymbol{y}} = \mathcal{A}(\mathcal{X}^\star) + \boldsymbol{\epsilon} \in \mathbb{R}^m, \tag{32}$$

where the noise vector $\boldsymbol{\epsilon}$ is assumed to be a Gaussian random vector with a mean of zero and a covariance of $\gamma^2 \mathbf{I}_m$. Similar to (25), we attempt to estimate the target low-TT-rank tensor $\mathcal{X}^\star$ by solving the following constrained factorized optimization problem:

$$\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ i \in [N]}} G(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) = \frac{1}{2m} \|\mathcal{A}([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]) - \widehat{\boldsymbol{y}}\|_2^2,$$

$$\text{s.t. } L^\top(\boldsymbol{X}_i) L(\boldsymbol{X}_i) = \mathbf{I}_{r_i}, i \in [N-1]. \tag{33}$$

Spectral Initialization In the midst of a noisy environment, we can still employ the spectral initialization method to obtain an appropriate initialization $\mathcal{X}^{(0)}$; that is

$$\mathcal{X}^{(0)} = \text{SVD}_{\boldsymbol{r}}^{tt} \left(\frac{1}{m} \sum_{k=1}^m \widehat{y}_k \mathcal{A}_k \right), \tag{34}$$

which is guaranteed to be close to the ground-truth $\mathcal{X}^\star$ when the operator $\mathcal{A}$ satisfies the RIP.

Theorem 9 *Assume that the operator $\mathcal{A}$ satisfies the $3\overline{r}$-RIP for TT format tensors with constant $\delta_{3\overline{r}}$ and that the additive noise vector $\boldsymbol{\epsilon}$ is randomly generated from the distribution $\mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_m)$. Then with probability at least $1 - 2e^{-c_1 N \overline{d} \overline{r}^2 \log N}$, the spectral initialization in (34) satisfies*

$$\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \le (1 + \sqrt{N-1}) \left(\delta_{3\overline{r}} \|\mathcal{X}^\star\|_F + \frac{c_2 \overline{r} \sqrt{(1 + \delta_{3\overline{r}}) N \overline{d} \log N}}{\sqrt{m}} \gamma \right), \tag{35}$$

where c_1, c_2 are positive constants, $\overline{r} = \max_{i=1}^{N-1} r_i$ and $\overline{d} = \max_{i=1}^N d_i$.

The proof is provided in Appendix E. Compared to the noiseless case, (35) for the noisy scenario includes an additional term caused by the additive noise in the measurement, which only scales polynomially with N thanks to the concise structure of the TT factorization.

---

4. Specifically, this term $\langle \mathcal{A}_k, [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N] \rangle$ can be efficiently evaluated as $\mathcal{A}_k \times_1^1 \boldsymbol{X}_1 \times_{N,1}^{1,2} \boldsymbol{X}_2 \times_{N-1,1}^{1,2} \cdots \times_{2,1}^{1,2} \boldsymbol{X}_N$, where we reshape $\boldsymbol{X}_1$ and $\boldsymbol{X}_N$ as matrices of size $d_1 \times r_1$ and $r_{N-1} \times d_N$, respectively. Here, the tensor contraction operation $\mathcal{A}_k \times_1^1 \boldsymbol{X}_1$ results in a new tensor $\mathcal{B}$ of size $d_2 \times d_3 \times \cdots \times d_N \times r_1$, with the $(s_2, s_3, \ldots, s_{N+1})$-th entry being $\sum_{s_1} \mathcal{A}_k(s_1, \ldots, s_N) \boldsymbol{X}_1(s_1, s_{N+1})$. Likewise, $\mathcal{B} \times_{N,1}^{1,2} \boldsymbol{X}_2$ results in a new tensor of size $d_3 \times d_4 \times \cdots \times d_N \times r_2$, with the $(s_3, \ldots, s_N, s_{N+2})$-th entry being $\sum_{s_2, s_{N+1}} \mathcal{B}(s_2, s_3, \ldots, s_{N+1}) \boldsymbol{X}_2(s_{N+1}, s_2, s_{N+2})$.

Local convergence of RGD algorithm   The following result ensures that, given an appropriate initialization, RGD will converge to the target low-TT-rank tensor up to a certain distance that is proportional to the noise level.

Theorem 10 *Consider a low-TT-rank tensor $\mathcal{X}^\star$ with ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$. Assume that $\mathcal{A}$ obeys the $(N+3)\overline{r}$-RIP with a constant $\delta_{(N+3)\overline{r}} \leq \frac{7}{30}$, where $\overline{r} = \max_i r_i$. Suppose that the RGD algorithm in (26) and (27) is initialized with $\{\boldsymbol{X}_i^{(0)}\}$ satisfying*

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{8(N + 1 + \sum_{i=2}^{N-1} r_i)(129N^2 + 7231N - 7360)} \tag{36}$$

and uses the step size $\mu \leq \frac{7 - 30\delta_{(N+3)\overline{r}}}{20(9N-5)(1+\delta_{(N+3)\overline{r}})^2}$. Then, with probability at least $1 - 2Ne^{-\Omega(N\overline{d}\overline{r}^2 \log N)} - 2e^{-\Omega(N^3\overline{d}\overline{r}^2 \log N)}$, the iterates $\{\boldsymbol{X}_i^{(t)}\}_{t \geq 0}$ generated by RGD satisfy

$$\begin{aligned}
\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq{}& \left(1 - \frac{7 - 30\delta_{(N+3)\overline{r}}}{1280(N + 1 + \sum_{i=2}^{N-1} r_i)\kappa^2(\mathcal{X}^\star)}\mu\right)^{t+1} \mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \\
&+ O\left(\frac{(N+\mu)(N + 1 + \sum_{i=2}^{N-1} r_i)(1 + \delta_{(N+3)\overline{r}})N^2\overline{d}\overline{r}^2(\log N)\kappa^2(\mathcal{X}^\star)\gamma^2}{m(7 - 30\delta_{(N+3)\overline{r}})}\right)
\end{aligned} \tag{37}$$

as long as $m \geq C\frac{N^5\overline{d}\overline{r}^3(\log N)\kappa^2(\mathcal{X}^\star)\gamma^2}{\underline{\sigma}^2(\mathcal{X}^\star)}$ with a universal constant C and $\overline{d} = \max_i d_i$.

The proof is given in Appendix F. Theorem 10 provides a similar guarantee to that in Theorem 7 for noisy measurements and shows that once the initial condition is satisfied, RGD converges at a linear rate to the target solution, up to a statistical error due to the additive noise. Notably, the second term in (37) scales linearly in terms of the variance γ^2, and polynomially in terms of the tensor order N. This is due to the presence of a polynomial number of degrees of freedom in the TT format, denoted as $O(N\overline{d}\overline{r}^2)$, which effectively mitigates the impact of noise. Similar to the discussion following Theorem 7, the requirement on the initialization (36) can be rewritten as $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F^2 \lesssim O(\frac{\underline{\sigma}^2(\mathcal{X}^\star)}{N^4\overline{r}^2\kappa^2(\mathcal{X}^\star)})$. Using Theorem 9 and Theorem 5, for an L-subgaussian measurement ensemble, the requirement for initialization and local convergence can be met by using the number of measurements $m \gtrsim \frac{N^6\overline{d}\overline{r}^3\kappa^2(\mathcal{X}^\star)(\|\mathcal{X}^\star\|_F^2 \log(N\overline{r}) + N\overline{r}\log N\gamma^2)}{\underline{\sigma}^2(\mathcal{X}^\star)}$, which compared with the noiseless case has an additional factor that scales with noise level γ^2. The recovery guarantee (37) scales optimally with the noise level γ^2, but is suboptimal with respect to the tensor order N when compared to the following minimax lower bound from (Qin and Zhu, 2024, Theorem 3).

Corollary 11 *Consider the tensor sensing problem in (32) for TT format tensors $\mathcal{X} \in \mathbb{X}_{\boldsymbol{r}} = \{\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N} : \boldsymbol{r} = (r_1, \ldots, r_{N-1})\}$. Suppose $\min_{i=1}^{N-1} r_i \geq C$ for some absolute constant C, and each element of $\{\mathcal{A}_i\}_i$ and $\boldsymbol{\epsilon}$ in (32) respectively are drawn independently from the distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \gamma^2)$. Then, it holds that*

$$\inf_{\widehat{\mathcal{X}}} \sup_{\mathcal{X} \in \mathbb{X}_{\boldsymbol{r}}} \mathbb{E}\|\widehat{\mathcal{X}} - \mathcal{X}\|_F \gtrsim \sqrt{\frac{\sum_{j=1}^N d_j r_{j-1} r_j}{m}}\gamma. \tag{38}$$

The above minimax lower bound demonstrates the potential for improving the recovery guarantee in Theorem 10, which together with Lemma 1 states that $\|\mathcal{X}^{(t)} - \mathcal{X}^{\star}\|_F \leq O\left(\frac{N^5 \overline{d} \overline{r}^3 (\log N) \kappa^2(\mathcal{X}^{\star}) \gamma^2}{m} \right)$ for sufficiently large t. Similar to the discussion right after Corollary 8, the suboptimal dependence with tensor order N arise from the following factors: (i) the requirement of $(N+3)\overline{r}$-RIP in the convergence analysis, and (ii) the exchange between the distances of the factors and the entire tensors. We leave the improvement as the subject of future work.

5. Numerical Experiments

In this section, we conduct numerical experiments to evaluate the performance of the RGD algorithm for tensor train sensing and completion. In all the experiments, we generate an order-N ground truth tensor $\mathcal{X}^{\star} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ in TT format with ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$ by first generating a random Gaussian tensor with i.i.d. entries from the normal distribution, and then using the sequential SVD algorithm to obtain a TT format tensor, which is finally normalized to unit Frobenius norm, i.e., $\|\mathcal{X}^{\star}\|_F = 1$. To simplify the selection of parameters, we set $d = d_1 = \cdots = d_N$ and $r = r_1 = \cdots = r_{N-1}$. For the RGD algorithm in (26) and (27), we set $\mu = 0.5$ to compute factors. To avoid the high computational complexity associated with $\overline{\sigma}^2(\mathcal{X}^{\star})$, we replace it with $\|\mathcal{X}^{\star}\|_F^2$ in the RGD. For each experimental setting, we conduct 20 Monte Carlo trials and then take the average over the 20 trials to report the results.

5.1 TT Format Tensor Sensing

We first consider the tensor sensing problem by generating each measurement operator $\mathcal{A}_i, i = 1, \ldots, m$ as a random tensor with i.i.d. entries drawn from the standard normal distribution. We then obtain noisy measurements $\widehat{y}_i = \langle \mathcal{A}_i, \mathcal{X}^{\star} \rangle + \epsilon_i$, where the noise ϵ is drawn from a Gaussian distribution with a mean of zero and a variance of γ^2.
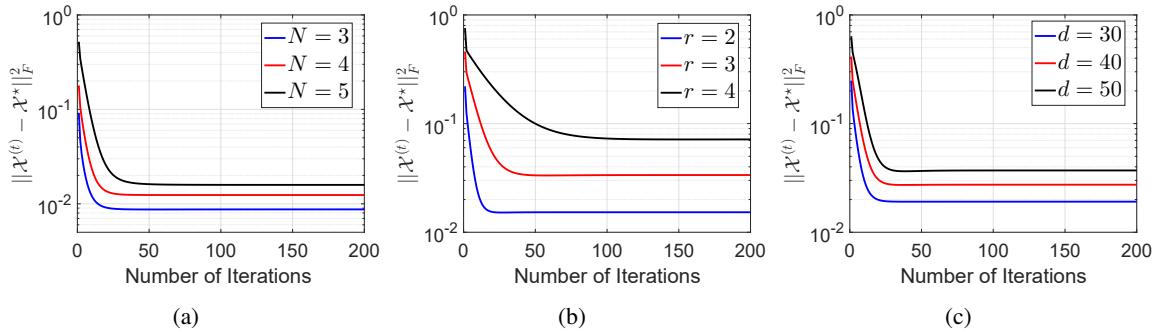

Figure 3: Convergence of RGD for TT format tensor sensing (a) for different N with $d = 10$, $r = 2$, $m = 1000$, and $\gamma^2 = 0.1$, (b) for different r with $d = 50$, $N = 3$, $m = 3000$, and $\gamma^2 = 0.1$, (c) for different d with $N = 3$, $r = 2$, $m = 1500$, and $\gamma^2 = 0.1$.

Convergence of RGD We first display the convergence of RGD in terms of the tensor, denoted as $\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2$, for different settings in Figure 3. We observe rapid convergence of RGD across all cases shown in Figure 3. Furthermore, the plots reveal the following trends when a fixed number of measurements m is maintained, while the values of N, r, and d increase: (i) the recovery error at the initialization using spectral methods increases, (ii) RGD converges more slowly, and (iii) RGD converges to a solution with larger recovery error. These observations align with our theoretical findings as presented in Theorem 9 for spectral initialization and Theorem 10 for the convergence guarantee of RGD. Since RGD converges relatively fast, as demonstrated in Figure 3, in the following experiments we run RGD for $T = 500$ iterations to obtain the estimated tensor $\widehat{\mathcal{X}}$.

(a)   (b)   (c)
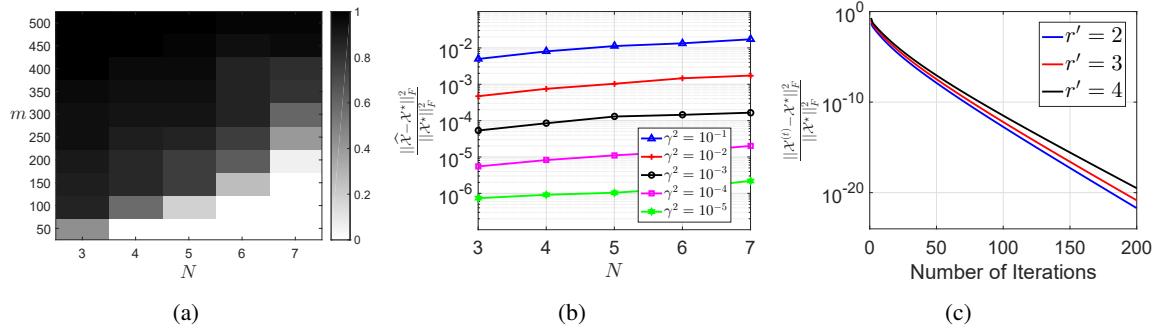
Figure 4: Performance comparison of RGD for TT format tensor sensing (a) for different N and m with $d = 4$, $r = 2$, and $\gamma^2 = 0$, (b) for different N and γ^2 with $d = 4$, $r = 2$, and $m = 500$, (c) for overparameterized tensor sensing where $N = 3$, $d = 30$, $r = 2$, $\gamma^2 = 0$, and $m = 5000$.

Exact recovery with clean measurements In the following two sets of experiments, we fix $d = 4$ and $r = 2$, and evaluate the performance for varying tensor order N. In the case of clean measurements, we conduct experiments for different N and number of measurements m, and we say a recovery is successful if $\|\widehat{\mathcal{X}} - \mathcal{X}^\star\|_F \leq 10^{-5}$. We conduct 100 independent trials to evaluate the success rate for each pair of N and m. The result is displayed in Figure 4(a). We can observe from Figure 4(a) that the number of measurements m needed to ensure exact recovery only scales linearly rather than exponentially in terms of the order N, consistent with the findings in Theorem 5. This relationship is indeed an improvement over the polynomial dependence of m on N required to guarantee Theorem 6 and Theorem 7. This also suggests the potential for refining our analysis in future work.

Stable recovery with noisy measurements In the case of noisy measurements, we fix the number of measurements $m = 500$ and vary the tensor order N and noise level γ^2. Figure 4(b) shows that the performance of RGD remains stable as N increases, with recovery error in the curves increasing polynomially. This behavior aligns with the findings outlined in Theorem 10. In addition, the recovery error scales roughly linearly with respect to the noise level γ^2, consistent with the second term in (37).

Recovery with over-parameterization   In the fourth experiment, we consider the case where the true rank r is unknown a priori, and we use an estimated rank r' in RGD. Figure 4(c) shows the convergence rates of RGD across various values of r' for the setting with clean measurements. While our current theory is only established when the rank is exactly specified (i.e., $r' = r$), we also observe linear convergence when the rank is over-specified (i.e., $r' > r$), albeit with a slightly slower rate of convergence as r' increases. A theoretical study for this overparameterized case is a topic for future research.

5.2 TT Format Tensor Completion

We now consider the problem of tensor completion, with the goal of reconstructing the entire tensor $\mathcal{X}^\star$ based on a subset of its entries. Specifically, let Ω denote the indices of m observed entries and let $\mathcal{P}_\Omega$ denote the corresponding measurement operator that produces the observed measurements. Then, our measurements $\widehat{\boldsymbol{y}}$ are obtained by

$$\widehat{\boldsymbol{y}} = \mathcal{P}_\Omega(\mathcal{X}^\star) + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ denotes the possible additive noise with each entry being independently drawn from a Gaussian distribution with a mean of zero and a variance of γ^2. Throughout the experiments, we assume the m observed entries are uniformly sampled. As the measurement operator $\mathcal{P}_\Omega$ can be viewed as a special instance of the linear map $\mathcal{A}$ in (22), tensor completion can be regarded as a special case of tensor sensing. Thus, as in (33), we attempt to recover the underlying tensor by solving the following constrained factorized optimization problem

$$\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ i \in [N]}} G(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) = \frac{1}{2}\|\mathcal{P}_\Omega([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]) - \widehat{\boldsymbol{y}}\|_2^2,$$

$$\text{s.t.}\quad L^\top(\boldsymbol{X}_i)L(\boldsymbol{X}_i) = \mathbf{I}_{r_i}, i \in [N-1].$$

We solve this problem using the RGD algorithm in (26) and (27). In addition, we employ a modified spectral initialization approach proposed in (Cai et al., 2022) with incoherent factors and guaranteed performance. We note that, in general, the measurement operator $\mathcal{P}_\Omega$ in tensor completion does not satisfy the RIP condition (Rauhut et al., 2015). Therefore, our theory may not be directly applicable in this context. Here, we only present numerical results of RGD for tensor completion, deferring the theoretical analysis to future work.

Convergence of RGD   Continuing with the same experiment conducted in tensor sensing, we begin by demonstrating the convergence rate of RGD in tensor completion under various settings. The results presented in Figures 5(a)-5(c) reveal a noticeable trend: as the values of N, r, and d increase, similar to the observations made in tensor sensing, we witness a degradation in the convergence rate, recovery error, and the estimated initialization. Additionally, it is important to emphasize that this consistency in degradation across different parameters reinforces the similarities between tensor completion and tensor sensing when employing RGD.

Exact recovery with clean measurements   In the second experiment, we set $d = 4$ and $r = 2$, and then assess the performance across various tensor orders N. When dealing with clean measurements, we perform experiments using different combinations of N and the number of samples m.

(a)   (b)   (c)
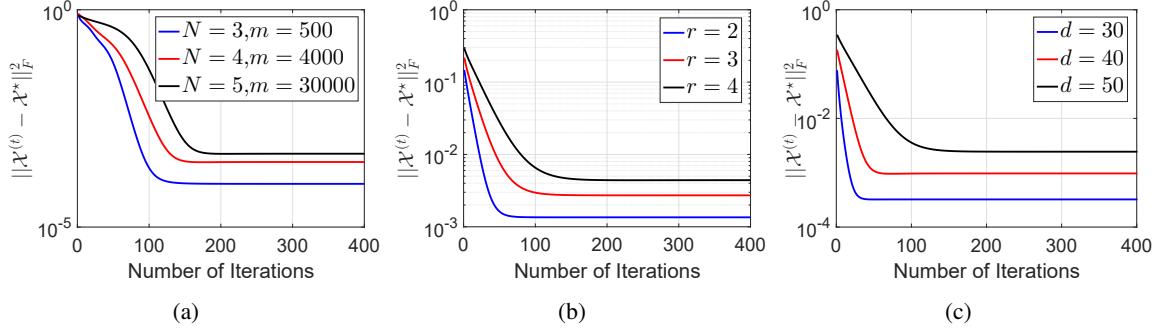
Figure 5: Convergence of RGD for TT format tensor completion (a) for different N and m with $d = 10$, $r = 2$, and $\gamma^2 = 10^{-6}$, (b) for different r with $d = 50$, $N = 3$, $m = 35000$, and $\gamma^2 = 10^{-6}$, (c) for different d with $N = 3$, $r = 2$, $m = 20000$, and $\gamma^2 = 10^{-6}$.

(a)   (b)
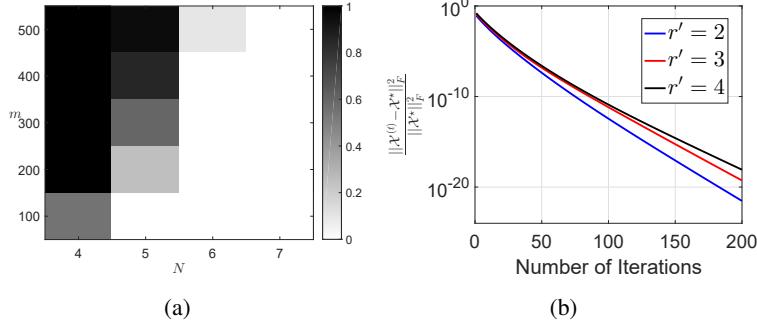
Figure 6: Performance comparison of RGD for TT format tensor completion, (a) for different N and m with $d = 4$, $r = 2$, and $\gamma^2 = 0$, (note that for $N = 4$ in (a), $m = 256$ has been chosen when $m \geq 300$), (b) for overparameterized tensor completion where $N = 3$, $d = 30$, $r = 2$, and $m = 20000$.

A successful recovery by RGD is defined as $\|\widehat{\mathcal{X}} - \mathcal{X}^\star\|_F \leq 10^{-5}$. For each pair of N and m, we conduct 100 independent trials to evaluate the success rate. Note that random initialization is employed here because the sequential second-order moment method (Cai et al., 2022) is likely to fail when the number of measurements m is relatively small compared to the total number of entries of the tensor. The results are presented in Figure 6(a). It is evident that the number of samples needed for successful recovery does not exhibit a polynomial relationship with N. This discovery aligns with the theoretical result in (Cai et al., 2022) that requires the number of samples m to increase exponentially with N.

Recovery with over-parameterization In the third experiment, we conclude by assessing the performance of RGD with overparameterized rank. This evaluation involves varying pre-defined values of r', as illustrated in Figure 6(b). Notably, the results indicate a clear trend as observed in

Figure 4(c) for the sensing problem: the convergence rate diminishes as r' increases, aligning with the observations from the tensor sensing scenario.

6. Conclusion and Outlook

In this paper, we study the tensor factorization approach for recovering low-TT-rank tensors from limited numbers of linear measurements. To avoid the ambiguity and to facilitate theoretical analysis, we optimize over the left-orthogonal TT format which enforces orthonormality among all the factors except for the last one. To ensure the orthonormal structure, we utilize the Riemannian gradient descent (RGD) algorithm for optimizing those factors over the Stiefel manifold. When the sensing operator obeys the RIP, we show that with an appropriate initialization which can be achieved by spectral initialization, RGD converges to the target solution at a linear rate. In the presence of measurement noise, RGD produces a stable recovery with error proportional to the noise level and scaling only polynomially in terms of the tensor order. Our findings support the growing evidence for using the factorization approach for low-TT-rank tensors and adopting local search algorithms such as gradient descent for solving the corresponding factorized optimization problems.

Extension to tensor factorization approach without orthonormal constraints   In this work, we use the left-orthogonal form to avoid scaling ambiguity among the tensor factors and to facilitate theoretical analysis. This approach is in line with other works on matrix and tensor factorization that use regularizers to balance the factors (Tu et al., 2016; Han et al., 2022; Cai et al., 2019; Tong et al., 2021b). However, both empirical observations (Zhu et al., 2018) and theoretical results (Ma et al., 2021; Li et al., 2020a) have shown that such regularizers are not necessary for the convergence of gradient descent (GD) in matrix factorization problems. Similarly, recent work (Jameson et al., 2024) has empirically demonstrated that GD can efficiently solve the TT factorization problem without orthogonal constraints or regularizers on the factors, as applied in quantum state tomography. Extending the analysis to provide a theoretical justification for this approach will be of interest.

Extension to other TT applications   An important area for future work is the analysis of the convergence properties of RGD in TT completion. Due to the random sampling process, the incoherence condition (Cai et al., 2022) plays a crucial role in ensuring the even distribution of energy across the entries of the tensor. Although experimental results have shown a linear convergence rate for RGD, there is a theoretical challenge in guaranteeing the nonexpansiveness property when applying both the orthonormal structure and incoherence condition simultaneously. Additionally, unlike the tensor itself, the TT rank is often unknown beforehand in practical scenarios. Building upon recent research efforts (Stöger and Soltanolkotabi, 2021; Jiang et al., 2022; Ding et al., 2022; Xu et al., 2023), a possible extension of our analysis is to consider overparameterized low-rank tensor recovery. This extension would involve investigating the convergence and error analysis of RGD in this context.

Acknowledgments

Appendix A. Technical tools used in the proofs

In this section, we introduce a new operation related to the multiplication of submatrices within the left unfolding matrices $L(\boldsymbol{X}_i) = \begin{bmatrix} \boldsymbol{X}_i(1) \\ \vdots \\ \boldsymbol{X}_i(d_i) \end{bmatrix} \in \mathbb{R}^{(r_{i-1}d_i)\times r_i}, i \in [N]$. For simplicity, we will only consider the case $d_i = 2$, but extending to the general case is straightforward.

Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix}$ be two block matrices, where $\boldsymbol{A}_i \in \mathbb{R}^{r_1 \times r_2}$ and $\boldsymbol{B}_i \in \mathbb{R}^{r_2 \times r_3}$ for $i = 1, 2$. We introduce the notation $\overline{\otimes}$ to represent the Kronecker product between submatrices in the two block matrices, as an alternative to the standard Kronecker product based on element-wise multiplication. Specifically, we define $\boldsymbol{A}\overline{\otimes}\boldsymbol{B}$ as

$$\boldsymbol{A}\overline{\otimes}\boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \overline{\otimes} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1\boldsymbol{B}_1 \\ \boldsymbol{A}_2\boldsymbol{B}_1 \\ \boldsymbol{A}_1\boldsymbol{B}_2 \\ \boldsymbol{A}_2\boldsymbol{B}_2 \end{bmatrix}. \tag{39}$$

Then we establish the following useful result.

Lemma 12 *For any matrices $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix}$, where $\boldsymbol{A}_i \in \mathbb{R}^{r_1 \times r_2}$ and $\boldsymbol{B}_i \in \mathbb{R}^{r_2 \times r_3}$, the following inequalities hold*

$$\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|_F, \tag{40}$$

$$\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|. \tag{41}$$

In particular, when $r_3 = 1$, (40) becomes

$$\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|_2. \tag{42}$$

Proof Using the relation $\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F^2 = \text{trace}((\boldsymbol{A}\overline{\otimes}\boldsymbol{B})^\top(\boldsymbol{A}\overline{\otimes}\boldsymbol{B}))$ gives

$$\begin{aligned} \|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F^2 &= \text{trace}((\boldsymbol{A}\overline{\otimes}\boldsymbol{B})^\top(\boldsymbol{A}\overline{\otimes}\boldsymbol{B})) \\ &= \text{trace}(\boldsymbol{B}_1^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_1 + \boldsymbol{B}_2^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_2) \\ &\leq \|\boldsymbol{B}_1\|_F^2\|\boldsymbol{A}\|^2 + \|\boldsymbol{B}_2\|_F^2\|\boldsymbol{A}\|^2 \\ &= \|\boldsymbol{A}\|^2\|\boldsymbol{B}\|_F^2, \end{aligned} \tag{43}$$

where the first inequality utilizes the result that $\text{trace}(\boldsymbol{C}\boldsymbol{D}) \leq \|\boldsymbol{C}\| \text{trace}(\boldsymbol{D})$ holds for any two PSD matrices $\boldsymbol{C}, \boldsymbol{D}$ (see (Zhu et al., 2021, Lemma 7)).

Likewise, by connecting the spectral norms between $\boldsymbol{A}$ and $(\boldsymbol{A}\overline{\otimes}\boldsymbol{B})^\top(\boldsymbol{A}\overline{\otimes}\boldsymbol{B})$, we have

$$
\begin{aligned}
\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|^2 &= \lambda_{\max}((\boldsymbol{A}\overline{\otimes}\boldsymbol{B})^\top(\boldsymbol{A}\overline{\otimes}\boldsymbol{B})) \\
&= \lambda_{\max}(\boldsymbol{B}_1^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_1 + \boldsymbol{B}_2^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_2) \\
&= \max_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \boldsymbol{B}_1^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_1\boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{B}_2^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{B}_2\boldsymbol{u} \\
&\leq \max_{\|\boldsymbol{u}\|_2=1} \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{u}^\top \boldsymbol{B}_1^\top \boldsymbol{B}_1\boldsymbol{u} + \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{u}^\top \boldsymbol{B}_2^\top \boldsymbol{B}_2\boldsymbol{u} \\
&= \max_{\|\boldsymbol{u}\|_2=1} \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})(\boldsymbol{u}^\top \boldsymbol{B}_1^\top \boldsymbol{B}_1\boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{B}_2^\top \boldsymbol{B}_2\boldsymbol{u}) \\
&= \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})\lambda_{\max}(\boldsymbol{B}^\top \boldsymbol{B}) \\
&= \|\boldsymbol{A}\|^2\|\boldsymbol{B}\|^2.
\end{aligned}
\tag{44}
$$

■

The inequality (40) can be viewed as a generalization of the result $\|\boldsymbol{C}\boldsymbol{D}\|_F \leq \|\boldsymbol{C}\| \cdot \|\boldsymbol{D}\|_F$ for any two matrices $\boldsymbol{C}, \boldsymbol{D}$ of appropriate sizes. However, unlike the matrix product case which also satisfies $\|\boldsymbol{C}\boldsymbol{D}\|_F \leq \|\boldsymbol{C}\|_F \cdot \|\boldsymbol{D}\|$, $\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|$ does not always hold. To upper bound $\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F$ with the spectral norm of $\boldsymbol{B}$, we will instead use $\|\boldsymbol{A}\overline{\otimes}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F \cdot \|\boldsymbol{B}\|_F \leq \mathrm{rank}(\boldsymbol{B})\|\boldsymbol{A}\|_F \cdot \|\boldsymbol{B}\|$. This discrepancy will account for the term $\sum_{i=2}^{N-1} r_i$ in the subsequent Lemma 16.

Applying Lemma 12 to the left-orthogonal TT format tensor $\mathcal{X}^\star = [\boldsymbol{X}_1^\star, \ldots, \boldsymbol{X}_N^\star]$ gives the following useful results:

$$
\|\mathcal{X}^\star\|_F = \|\mathrm{vec}(\mathcal{X}^\star)\|_2 = \|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^\star)\|_2 = \|L(\boldsymbol{X}_N^\star)\|_2,
\tag{45}
$$

$$
\|L(\boldsymbol{X}_i^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_j^\star)\| \leq \Pi_{l=i}^{j}\|L(\boldsymbol{X}_l^\star)\| = 1, \ \ i \leq j, \ \ \forall i,j \in [N-1],
\tag{46}
$$

$$
\|L(\boldsymbol{X}_i^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_j^\star)\|_F \leq \Pi_{l=i}^{j-1}\|L(\boldsymbol{X}_l^\star)\|\|L(\boldsymbol{X}_j^\star)\|_F = \sqrt{r_j}, \ \ i \leq j, \ \ \forall i,j \in [N-1].
\tag{47}
$$

In addition, according to (39), each row in $L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_i^\star)$ can be represented as

$$
\begin{aligned}
&(L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_i^\star))(s_1\cdots s_i,:) \\
&= (L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_i^\star))(s_1 + d_1(s_2-2) + \cdots + d_1\cdots d_{i-1}(s_i-1),:) \\
&= \boldsymbol{X}_1(s_1)\cdots\boldsymbol{X}_i(s_i).
\end{aligned}
\tag{48}
$$

Next, we provide some useful lemmas in terms of products of matrices.

Lemma 13 *For any $\boldsymbol{A}_i, \boldsymbol{A}_i^\star \in \mathbb{R}^{r_{i-1}\times r_i}, i \in [N]$, we have*

$$
\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star = \sum_{i=1}^{N}\boldsymbol{A}_1^\star\cdots\boldsymbol{A}_{i-1}^\star(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}\cdots\boldsymbol{A}_N.
\tag{49}
$$

Proof

The result can be obtained by summing up the following equations:

$$
\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2\cdots\boldsymbol{A}_N = \boldsymbol{A}_1^\star\boldsymbol{A}_2\cdots\boldsymbol{A}_N
$$

$$
\boldsymbol{A}_1^\star\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star(\boldsymbol{A}_2 - \boldsymbol{A}_2^\star)\boldsymbol{A}_3\cdots\boldsymbol{A}_N = \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\boldsymbol{A}_3\cdots\boldsymbol{A}_N
$$

$$
\vdots
$$

$$
\boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_{N-1}^\star\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_{N-1}^\star(\boldsymbol{A}_N - \boldsymbol{A}_N^\star) = \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star.
$$

Lemma 14 *For any $\boldsymbol{A}_i, \boldsymbol{A}_i^\star \in \mathbb{R}^{r_{i-1} \times r_i}, i \in [N]$, we have*

$$\boldsymbol{A}_1^\star \cdots \boldsymbol{A}_N^\star - \boldsymbol{A}_1 \ldots \boldsymbol{A}_{N-1} \boldsymbol{A}_N + \sum_{i=1}^N \boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1} \cdots \boldsymbol{A}_N$$

$$= \sum_{i=1}^{N-1} \sum_{j=i+1}^N \boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}^\star \cdots \boldsymbol{A}_{j-1}^\star(\boldsymbol{A}_j - \boldsymbol{A}_j^\star)\boldsymbol{A}_{j+1} \cdots \boldsymbol{A}_N, \quad (50)$$

where the right-hand side of (50) contains a total of $\frac{N(N-1)}{2}$ terms.

Proof We first rewrite the term $(\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2 \cdots \boldsymbol{A}_N$ as

$$(\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2 \cdots \boldsymbol{A}_N = (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2^\star \cdots \boldsymbol{A}_N^\star + (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)(\boldsymbol{A}_2 \cdots \boldsymbol{A}_N - \boldsymbol{A}_2^\star \cdots \boldsymbol{A}_N^\star)$$

$$= (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2^\star \cdots \boldsymbol{A}_N^\star + (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\left(\sum_{j=2}^N \boldsymbol{A}_2^\star \cdots \boldsymbol{A}_{j-1}^\star(\boldsymbol{A}_j - \boldsymbol{A}_j^\star)\boldsymbol{A}_{j+1} \cdots \boldsymbol{A}_N \right), \quad (51)$$

where the second line uses Lemma 13 for expanding the difference $\boldsymbol{A}_2 \cdots \boldsymbol{A}_N - \boldsymbol{A}_2^\star \cdots \boldsymbol{A}_N^\star$. We can apply the same approach for $i = 2, \ldots, N$ to get

$$\boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1} \cdots \boldsymbol{A}_N$$

$$= \boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}^\star \cdots \boldsymbol{A}_N^\star$$

$$+ \boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\left(\sum_{j=i+1}^N \boldsymbol{A}_{i+1}^\star \cdots \boldsymbol{A}_{j-1}^\star(\boldsymbol{A}_j - \boldsymbol{A}_j^\star)\boldsymbol{A}_{j+1} \cdots \boldsymbol{A}_N \right). \quad (52)$$

Noting that the sum of the second terms in the right-hand side of (51) and (52) equals the right-hand side of (50), we complete the proof by checking the rest of the terms:

$$\boldsymbol{A}_1^\star \boldsymbol{A}_2^\star \cdots \boldsymbol{A}_N^\star - \boldsymbol{A}_1 \cdots \boldsymbol{A}_{N-1} \boldsymbol{A}_N + \sum_{i=1}^N \boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}^\star \cdots \boldsymbol{A}_N^\star = 0, \quad (53)$$

which follows from Lemma 13.

∎

Lemma 15 *((Cai et al., 2022; Han et al., 2022)) For any two matrices $\boldsymbol{X}, \boldsymbol{X}^\star$ with rank r, let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and $\boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ respectively represent the compact singular value decompositions (SVDs) of $\boldsymbol{X}$ and $\boldsymbol{X}^\star$. Supposing that $\boldsymbol{R} = \arg\min_{\widetilde{\boldsymbol{R}} \in \mathbb{O}^{r \times r}} \|\boldsymbol{U} - \boldsymbol{U}^\star\widetilde{\boldsymbol{R}}\|_F$, we have*

$$\|\boldsymbol{U} - \boldsymbol{U}^\star\boldsymbol{R}\|_F \leq \frac{2\|\boldsymbol{X} - \boldsymbol{X}^\star\|_F}{\sigma_r(\boldsymbol{X}^\star)}. \quad (54)$$

Lemma 16 *For any two TT format tensors $\mathcal{X}$ and $\mathcal{X}^\star$ with ranks $\boldsymbol{r} = (r_1, \ldots, r_{N-1})$. Let $\{\boldsymbol{X}_i\}$ and $\{\boldsymbol{X}_i^\star\}$ be the corresponding left-orthogonal form factors. Assume $\overline{\sigma}^2(\mathcal{X}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$. Then we have*

$$\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \geq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{8(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)} \operatorname{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}), \tag{55}$$

$$\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \leq \frac{9N}{4} \operatorname{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}), \tag{56}$$

where $\operatorname{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\})$ is defined in (6).

Proof By the definition of the i-th unfolding of the tensor $\mathcal{X}$, we have

$$\mathcal{X}^{\langle i \rangle} = \boldsymbol{X}^{\leq i} \boldsymbol{X}^{\geq i+1}, \tag{57}$$

where each row of the left part $\boldsymbol{X}^{\leq i}$ and each column of the right part $\boldsymbol{X}^{\geq i+1}$ can be represented as

$$\boldsymbol{X}^{\leq i}(s_1 \cdots s_i, :) = \boldsymbol{X}_1(s_1) \cdots \boldsymbol{X}_i(s_i), \tag{58}$$

$$\boldsymbol{X}^{\geq i+1}(:, s_{i+1} \cdots s_N) = \boldsymbol{X}_{i+1}(s_{i+1}) \cdots \boldsymbol{X}_N(s_N). \tag{59}$$

According to (Cai et al., 2022, Lemma 1), the left part in the left-orthogonal TT format satisfies $\boldsymbol{X}^{\leq i\top} \boldsymbol{X}^{\leq i} = \boldsymbol{I}_{r_i}, i \in [N - 1]$. Furthermore, based on the analysis of (Cai et al., 2022) stated in Lemma 15, we have

$$\max_{i=1,\ldots,N-1} \|\boldsymbol{X}^{\leq i} - \boldsymbol{X}^{\star \leq i} \boldsymbol{R}_i\|_F \leq \frac{2\|\mathcal{X} - \mathcal{X}^\star\|_F}{\underline{\sigma}(\mathcal{X}^\star)}, \tag{60}$$

where $\boldsymbol{R}_i = \arg\min_{\widetilde{\boldsymbol{R}}_i \in \mathbb{O}^{r_i \times r_i}} \|\boldsymbol{X}^{\leq i} - \boldsymbol{X}^{\star \leq i} \widetilde{\boldsymbol{R}}_i\|_F$.

By the definition of $L(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_N^\star)$ in (48), we have

$$\boldsymbol{X}^{\leq i} - \boldsymbol{X}^{\star \leq i} \boldsymbol{R}_i = L(\boldsymbol{X}_1) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \tag{61}$$

which together with the above equation gives

$$\|L(\boldsymbol{X}_1) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 \leq \frac{4\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)}, i \in [N - 1]. \tag{62}$$

We now use this result to upper bound $\|L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2$ for each i. First, setting $i = 1$ in the above equation directly yields

$$\|L(\boldsymbol{X}_1) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\|_F^2 \leq \frac{4\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)}. \tag{63}$$

With (47) and (62), we can obtain the result for $i = 2$ as

$$\begin{aligned}
&\|L(\boldsymbol{X}_2) - L_{\boldsymbol{R}}(\boldsymbol{X}_2^\star)\|_F^2 \\
&= \|L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L(\boldsymbol{X}_2) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_2^\star)\|_F^2 \\
&= \|L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L(\boldsymbol{X}_2) - L(\boldsymbol{X}_1) \overline{\otimes} L(\boldsymbol{X}_2) + L(\boldsymbol{X}_1) \overline{\otimes} L(\boldsymbol{X}_2) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_2^\star)\|_F^2 \\
&\leq 2\|L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L(\boldsymbol{X}_2) - L(\boldsymbol{X}_1) \overline{\otimes} L(\boldsymbol{X}_2)\|_F^2 + 2\|L(\boldsymbol{X}_1) \overline{\otimes} L(\boldsymbol{X}_2) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_2^\star)\|_F^2 \\
&\leq 2\|L(\boldsymbol{X}_2)\|_F^2 \|L(\boldsymbol{X}_1) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\|_F^2 + 2\|L(\boldsymbol{X}_1) \overline{\otimes} L(\boldsymbol{X}_2) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_2^\star)\|_F^2 \\
&\leq \frac{(8r_2 + 8)\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)}.
\end{aligned} \tag{64}$$

A similar derivation also gives

$$\|L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 \le \frac{(8r_i + 8)\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)}, \quad i = 3, \ldots, N-1. \tag{65}$$

Finally, we bound the term for $i = N$ as follows:

$$
\begin{aligned}
& \|L(\boldsymbol{X}_N) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2 \\
= \ & \|L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{N-1}^\star)\overline{\otimes}(L(\boldsymbol{X}_N) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star))\|_2^2 \\
= \ & \|L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{N-1}^\star)\overline{\otimes}L(\boldsymbol{X}_N) - L(\boldsymbol{X}_1)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1})\overline{\otimes}L(\boldsymbol{X}_N) \\
& +L(\boldsymbol{X}_1)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1})\overline{\otimes}L(\boldsymbol{X}_N) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{N-1}^\star)\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2 \\
\le \ & 2\|R(\boldsymbol{X}_N)\|^2\|L(\boldsymbol{X}_1)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1}) - L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{N-1}^\star)\|_F^2 + 2\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \\
\le \ & \frac{18\overline{\sigma}^2(\mathcal{X}^\star)\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)} + 2\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \\
\le \ & \frac{20\overline{\sigma}^2(\mathcal{X}^\star)\|\mathcal{X} - \mathcal{X}^\star\|_F^2}{\underline{\sigma}^2(\mathcal{X}^\star)}, \tag{66}
\end{aligned}
$$

where $R(\boldsymbol{X}_N)$ is the right unfolding matrix of $\boldsymbol{X}_N$ and the first inequality follows because $\|\boldsymbol{A}_1\overline{\otimes}L(\boldsymbol{X}_N) - \boldsymbol{A}_2\overline{\otimes}L(\boldsymbol{X}_N)\|_F = \|(\boldsymbol{A}_1 - \boldsymbol{A}_2)R(\boldsymbol{X}_N)\|_F \le \|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F\|R(\boldsymbol{X}_N)\| = \|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F\sigma_1(\mathcal{X}^{\langle N-1\rangle})$ with $\boldsymbol{A}_1 = L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{N-1}^\star)$ and $\boldsymbol{A}_2 = L(\boldsymbol{X}_1)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1})$.

Combining (64), (65) and (66) together gives

$$\text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \le \frac{8(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)}\|\mathcal{X} - \mathcal{X}^\star\|_F^2. \tag{67}$$

On the other hand, invoking Lemma 13 gives

$$
\begin{aligned}
\|\mathcal{X} - \mathcal{X}^\star\|_F^2 &= \|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{i-1}^\star)\overline{\otimes}(L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star))\overline{\otimes}L(\boldsymbol{X}_{i+1})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N)\|_F^2 \\
&\le N\left(\frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star)\|_F^2 + \|L(\boldsymbol{X}_N) - L(\boldsymbol{X}_N^\star)\|_F^2\right) \\
&\le \frac{9N}{4}\text{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}). \tag{68}
\end{aligned}
$$

where the first inequality follows $\|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}(L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star))\overline{\otimes}L(\boldsymbol{X}_{i+1})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N)\|_F = \|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}(L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star))\mathcal{X}^{\ge i+1}\|_F \le \|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}(L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star))\|_F\|\mathcal{X}^{\ge i+1}\| \le \|L(\boldsymbol{X}_1^\star)\|\cdots\|L(\boldsymbol{X}_{i-1}^\star)\|\|L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^\star)\|_F\overline{\sigma}(\mathcal{X})$. ∎

Lemma 17 *((Li et al., 2021, Lemma 1)) Let $\boldsymbol{X} \in St(n, r)$ and $\boldsymbol{\xi} \in T_{\boldsymbol{X}}St$ be given. Consider the point $\boldsymbol{X}^+ = \boldsymbol{X} + \boldsymbol{\xi}$. Then, the polar decomposition-based retraction $Retr_{\boldsymbol{X}}(\boldsymbol{X}^+) = \boldsymbol{X}^+(\boldsymbol{X}^{+\top}\boldsymbol{X}^+)^{-\frac{1}{2}}$ satisfies*

$$\|Retr_{\boldsymbol{X}}(\boldsymbol{X}^+) - \overline{\boldsymbol{X}}\|_F \le \|\boldsymbol{X}^+ - \overline{\boldsymbol{X}}\|_F = \|\boldsymbol{X} + \boldsymbol{\xi} - \overline{\boldsymbol{X}}\|_F, \ \forall \overline{\boldsymbol{X}} \in St(n, r). \tag{69}$$

Appendix B. Proof of Theorem 2 in Tensor-train Factorization

Proof Before proving Theorem 2, we first present a useful property for the factors $L(\boldsymbol{X}_i^{(t)})$. Due to the retraction, the factors $L(\boldsymbol{X}_i^{(t)}), i \in [N-1]$ are always orthonormal. Assuming that

$$\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \le \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{72(N^2-1)(N+1+\sum_{i=2}^{N-1} r_i)}, \tag{70}$$

which is true for $t = 0$ and will be proved later for $t \ge 1$, we obtain that

$$\sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) = \|\mathcal{X}^{(t) \ge i+1}\|^2 \le \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 2\|\boldsymbol{R}_i^\top \mathcal{X}^{\star \ge i+1}\|^2 + 2\|\mathcal{X}^{(t) \ge i+1} - \boldsymbol{R}_i^\top \mathcal{X}^{\star \ge i+1}\|^2$$

$$\le 2\overline{\sigma}^2(\mathcal{X}^\star) + \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 2\|\mathcal{X}^{(t)\langle i \rangle} - \mathcal{X}^{\star \langle i \rangle} + \mathcal{X}^{\star \langle i \rangle} - \mathcal{X}^{(t) \le i} \boldsymbol{R}_i^\top \mathcal{X}^{\star \ge i+1}\|^2$$

$$\le 2\overline{\sigma}^2(\mathcal{X}^\star) + 4\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 4\|\boldsymbol{R}_i^\top \mathcal{X}^{\star \ge i+1}\|^2 \|\mathcal{X}^{(t) \le i} - \mathcal{X}^{\star \le i} \boldsymbol{R}_i\|_F^2$$

$$\le 2\overline{\sigma}^2(\mathcal{X}^\star) + \left(4 + \frac{16\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)}\right) \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2$$

$$\le 2\overline{\sigma}^2(\mathcal{X}^\star) + \frac{45N\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)} \text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \le \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1]. \tag{71}$$

where each element of $\mathcal{X}^{(t) \le i}$ and $\mathcal{X}^{(t) \ge i+1}$ are respectively defined in (58) and (59). The fourth and last lines respectively follow (60) and (56). Note that $\overline{\sigma}^2(\mathcal{X}^{(t)}) = \max_{i=1}^{N-1} \sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) \le \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$.

We now prove the decay of $\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$. First recall from (18) that

$$\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\le \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) - 2\mu \sum_{i=1}^{N} \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)} \text{St}} \left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$+ \mu^2 \left(\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\|\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)} \text{St}} \left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\|_F^2 + \left\|\nabla_{L(\boldsymbol{X}_N)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\|_2^2\right), \tag{72}$$

where to unify the notation for all i, we define the projection onto the tangent space for $i = N$ as $\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_N)} \text{St}} = \mathcal{I}$ since there is no constraint on the N-th factor. Note that the gradient $\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})$ is defined as

$$\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) = \begin{bmatrix} \nabla_{\boldsymbol{X}_i(1)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \\ \vdots \\ \nabla_{\boldsymbol{X}_i(d_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \end{bmatrix}, \tag{73}$$

where the gradient with respect to each factor $\boldsymbol{X}_i(s_i)$ can be computed as

$$\nabla_{\boldsymbol{X}_i(s_i)} f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) = \sum_{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_N} \Big((\mathcal{X}^{(t)}(s_1, \ldots, s_N) - \mathcal{X}^\star(s_1, \ldots, s_N)) \cdot$$

$$\boldsymbol{X}_{i-1}^{(t)}(s_{i-1})^\top \cdots \boldsymbol{X}_1^{(t)}(s_1)^\top \boldsymbol{X}_N^{(t)}(s_N)^\top \cdots \boldsymbol{X}_{i+1}^{(t)}(s_{i+1})^\top\Big).$$

Note that computing this gradient only requires $\mathcal{X}^\star$ and does not rely on the knowledge of the factors in a TT decomposition of $\mathcal{X}^\star$.

The following is to bound the second and third terms in (72), respectively.

Upper bound of the third term in (72) We first define three matrices for $i \in [N]$ as follows:

$$
\begin{aligned}
\boldsymbol{D}_1(i) &= \left[\boldsymbol{X}_{i-1}^{(t)}(1)^\top \cdots \boldsymbol{X}_1^{(t)}(1)^\top \quad \cdots \quad \boldsymbol{X}_{i-1}^{(t)}(d_{i-1})^\top \cdots \boldsymbol{X}_1^{(t)}(d_1)^\top \right] \\
&= L^\top (\boldsymbol{X}_{i-1}^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L^\top (\boldsymbol{X}_1^{(t)}) \in \mathbb{R}^{r_i \times (d_1 \cdots d_{i-1})},
\end{aligned}
\tag{74}
$$

$$
\boldsymbol{D}_2(i) = \begin{bmatrix} \boldsymbol{X}_N^{(t)}(1)^\top \cdots \boldsymbol{X}_{i+1}^{(t)}(1)^\top \\ \vdots \\ \boldsymbol{X}_N^{(t)}(d_N)^\top \cdots \boldsymbol{X}_{i+1}^{(t)}(d_{i+1})^\top \end{bmatrix} \in \mathbb{R}^{(d_{i+1} \cdots d_N) \times r_i},
\tag{75}
$$

where we note that $\boldsymbol{D}_1(1) = 1$ and $\boldsymbol{D}_2(N) = 1$. Moreover, for each $s_i \in [d_i]$, we define matrix $\boldsymbol{E}(s_i) \in \mathbb{R}^{(d_1 \cdots d_{i-1}) \times (d_{i+1} \cdots d_N)}$ whose $(s_1 \cdots s_{i-1}, s_{i+1} \cdots s_N)$-th element is given by

$$
\boldsymbol{E}(s_i)(s_1 \cdots s_{i-1}, s_{i+1} \cdots s_N) = \boldsymbol{X}^{(t)}(s_1, \dots, s_i, \dots, s_N) - \boldsymbol{X}^\star(s_1, \dots, s_i, \dots, s_N).
$$

Based on the aforementioned notations, we bound $\left\| \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F^2$ by

$$
\begin{aligned}
\left\| \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F^2 &= \sum_{s_i=1}^{d_i} \left\| \nabla_{\boldsymbol{X}_i(s_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F^2 \\
&= \sum_{s_i=1}^{d_i} \| \boldsymbol{D}_1(i) \boldsymbol{E}(s_i) \boldsymbol{D}_2(i) \|_F^2 \\
&\leq \sum_{s_i=1}^{d_i} \| L^\top (\boldsymbol{X}_{i-1}^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L^\top (\boldsymbol{X}_1^{(t)}) \|^2 \| \boldsymbol{D}_2(i) \|^2 \| \boldsymbol{E}(s_i) \|_F^2 \\
&\leq \| L(\boldsymbol{X}_1^{(t)}) \|^2 \cdots \| L(\boldsymbol{X}_{i-1}^{(t)}) \|^2 \| \mathcal{X}^{(t) \geq i+1} \|^2 \| \mathcal{X}^{(t)} - \mathcal{X}^\star \|_F^2 \\
&\leq \begin{cases} \frac{9 \overline{\sigma}^2 (\mathcal{X}^\star)}{4} \| \mathcal{X}^{(t)} - \mathcal{X}^\star \|_F^2, & i \in [N-1], \\ \| \mathcal{X}^{(t)} - \mathcal{X}^\star \|_F^2, & i = N, \end{cases}
\end{aligned}
\tag{76}
$$

where we use (46), $\| \boldsymbol{D}_2(i) \| = \| (\boldsymbol{D}_2(i))^\top \| = \| \mathcal{X}^{(t) \geq i+1} \| = \sigma_1(\mathcal{X}^{(t) \langle i \rangle}) \leq \frac{3 \overline{\sigma}(\mathcal{X}^\star)}{2}$ and $\sum_{s_i=1}^{d_i} \| \boldsymbol{E}(s_i) \|_F^2 = \| \mathcal{X}^{(t)} - \mathcal{X}^\star \|_F^2$ in the second inequality.

Using (76), we obtain an upper bound for the third term in (72) as

$$
\begin{aligned}
& \frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}} \left(\nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 + \left\| \nabla_{L(\boldsymbol{X}_N)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_2^2 \\
&\leq \frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\| \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F^2 + \left\| \nabla_{L(\boldsymbol{X}_N)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_2^2 \\
&\leq \frac{9N - 5}{4} \| \mathcal{X}^{(t)} - \mathcal{X}^\star \|_F^2,
\end{aligned}
\tag{77}
$$

where the first inequality follows from the fact that for any matrix $\boldsymbol{B} = \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}}(\boldsymbol{B}) + \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}}^\perp (\boldsymbol{B})$ where $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}}(\boldsymbol{B})$ and $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}}^\perp (\boldsymbol{B})$ are orthogonal, we have $\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}}(\boldsymbol{B}) \|_F^2 \leq \| \boldsymbol{B} \|_F^2$.

29

Lower bound of the second term in (72)   We first expand the second term in (72) as follows:

$$\sum_{i=1}^{N}\left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$= \sum_{i=1}^{N}\left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}), \nabla_{L(\boldsymbol{X}_i)}f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\rangle - T_1$$

$$= \left\langle L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^{\star})\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^{\star}), L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)})\right.$$

$$\left. - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^{\star})\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^{\star}) + \boldsymbol{h}^{(t)}\right\rangle - T_1, \tag{78}$$

where we define

$$\boldsymbol{h}^{(t)} = L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^{\star})\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^{\star}) - L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1}^{(t)})\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^{\star})$$

$$+ \sum_{i=1}^{N-1} L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{i-1}^{(t)})\overline{\otimes}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))\overline{\otimes}L(\boldsymbol{X}_{i+1}^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)}), \tag{79}$$

and

$$T_1 = \sum_{i=1}^{N-1}\left\langle \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star})), \nabla_{L(\boldsymbol{X}_i)}f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\rangle. \tag{80}$$

Recalling the definition for orthogonal complement projection in (10), we can rewrite the term $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(\cdot)$ as

$$\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))$$

$$= \frac{1}{2}L(\boldsymbol{X}_i^{(t)})\left((L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))^{\top}L(\boldsymbol{X}_i^{(t)}) + L^{\top}(\boldsymbol{X}_i^{(t)})(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))\right)$$

$$= \frac{1}{2}L(\boldsymbol{X}_i^{(t)})\left(2\mathbf{I}_{r_i} - L_{\boldsymbol{R}^{(t)}}^{\top}(\boldsymbol{X}_i^{\star})L(\boldsymbol{X}_i^{(t)}) - L^{\top}(\boldsymbol{X}_i^{(t)})L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star})\right)$$

$$= \frac{1}{2}L(\boldsymbol{X}_i^{(t)})\left((L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))^{\top}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))\right). \tag{81}$$

To derive the lower bound of (78), we first utilize Lemma 14 to obtain the upper bound on $\|\boldsymbol{h}^{(t)}\|_2^2$ as follows:

$$\|\boldsymbol{h}^{(t)}\|_2^2 = \|\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{i-1}^{(t)})\overline{\otimes}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star}))\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_{i+1}^{\star})\overline{\otimes}$$

$$\cdots\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{j-1}^{\star})\overline{\otimes}(L(\boldsymbol{X}_j^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_j^{\star}))\overline{\otimes}L(\boldsymbol{X}_{j+1}^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)})\|_2^2$$

$$\leq \frac{N(N-1)}{2}\left(\sum_{j=1}^{N-1}\|L(\boldsymbol{X}_j^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_j^{\star})\|_F^2 \cdot\right.$$

$$\left(\sum_{i=j+1}^{N-1}\frac{9\overline{\sigma}^2(\mathcal{X}^{\star})}{4}\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^{\star})\|_F^2 + \|L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^{\star})\|_2^2\right)\right)$$

30

$$\leq \frac{9N(N-1)}{8} \sum_{i=1}^{N-1} \|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2 \mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\leq \frac{9N(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)} \mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), \tag{82}$$

where (47) and $\|\boldsymbol{A}\overline{\otimes}L(\boldsymbol{X}_{j+1}^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)})\|_2 = \|\boldsymbol{A}\mathcal{X}^{(t)\geq j+1}\|_F \leq \|\mathcal{X}^{(t)\geq j+1}\|\|\boldsymbol{A}\|_F \leq \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}\|\boldsymbol{A}\|_F$ with $\boldsymbol{A} = L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{i-1}^{(t)})\overline{\otimes}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star))\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_{i+1}^\star)\overline{\otimes}\cdots$ $\overline{\otimes}L_{\boldsymbol{R}}(\boldsymbol{X}_{j-1}^\star)\overline{\otimes}(L(\boldsymbol{X}_j^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_j^\star))$ are used in the first inequality. We then establish the upper bound of T_1 as follows:

$$T_1 \leq \sum_{i=1}^{N-1} \frac{1}{2}\|L(\boldsymbol{X}_i^{(t)})\|\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2 \left\|\nabla_{L(\boldsymbol{X}_i)}f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\|_F$$

$$\leq \sum_{i=1}^{N-1} \frac{3\overline{\sigma}(\mathcal{X}^\star)}{4}\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2 \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F$$

$$\leq \frac{1}{4}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \frac{9(N-1)\overline{\sigma}^2(\mathcal{X}^\star)}{16} \sum_{i=1}^{N-1} \|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^4$$

$$\leq \frac{1}{4}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \frac{9(N-1)}{16\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), \tag{83}$$

where the second inequality follows from (76).

Now plugging (82) and (83) into (78) gives

$$\sum_{i=1}^{N} \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}f(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$\geq \frac{1}{2}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{1}{2}\|\boldsymbol{h}^{(t)}\|_2^2 - \frac{1}{4}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{9(N-1)}{16\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\geq \frac{1}{4}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{9N(N-1)}{16\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) - \frac{9(N-1)}{16\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\geq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{128(N+1+\sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) + \frac{1}{8}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2, \tag{84}$$

where the first and second inequalities follow from (83) and (82), and we utilize Lemma 16 along with the initial condition $\mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{72(N^2-1)(N+1+\sum_{i=2}^{N-1} r_i)}$ in the last line.

Contraction Taking (77) and (84) into (72), we have

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq \left(1 - \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{64(N+1+\sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right)\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$$

$$+ \left(\frac{9N-5}{4}\mu^2 - \frac{\mu}{4}\right)\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2$$

$$\leq \left(1 - \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{64(N+1+\sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right)\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), \tag{85}$$

where we assume $\mu \leq \frac{1}{9N-5}$ in the last line.

31

Proof of (70) We now prove (70) by induction. First note that (90) holds for $t = 0$. We now assume it holds at $t = t'$, which implies that $\sigma_1^2(\mathcal{X}^{(t')\langle i \rangle}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1]$. By invoking (85), we have $\text{dist}^2(\{\boldsymbol{X}_i^{(t'+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$. Consequently, (70) also holds at $t = t' + 1$. By induction, we can conclude that (70) holds for all $t \geq 0$. This completes the proof. ∎

Appendix C. Proof of Theorem 6 in Spectral Initialization

We first provide one useful lemma. As an immediate consequence of the RIP, the inner product between two low-rank TT format tensors is also nearly preserved if $\mathcal{A}$ satisfies the RIP.

Lemma 18 *((Candès and Plan, 2011; Rauhut et al., 2017)) Suppose that $\mathcal{A}$ obeys the $2\overline{r}$-RIP with a constant $\delta_{2\overline{r}}$. Then for any left-orthogonal TT formats $\mathcal{X}_1, \mathcal{X}_2 \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ of rank at most $\overline{r}$, one has*

$$\left| \frac{1}{m} \langle \mathcal{A}(\mathcal{X}_1), \mathcal{A}(\mathcal{X}_2) \rangle - \langle \mathcal{X}_1, \mathcal{X}_2 \rangle \right| \leq \delta_{2\overline{r}} \|\mathcal{X}_1\|_F \|\mathcal{X}_2\|_F, \tag{86}$$

or equivalently,

$$\left| \left\langle \left(\frac{1}{m} \mathcal{A}^* \mathcal{A} - \mathcal{I} \right)(\mathcal{X}_1), \mathcal{X}_2 \right\rangle \right| \leq \delta_{2\overline{r}} \|\mathcal{X}_1\|_F \|\mathcal{X}_2\|_F, \tag{87}$$

where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$ and is defined as $\mathcal{A}^*(\boldsymbol{x}) = \sum_{i=1}^m x_i \mathcal{A}_i$.

Proof [Proof of Theorem 6] Before analyzing the spectral initialization, we first define the following restricted Frobenius norm for any tensor $\mathcal{H} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$:

$$
\begin{aligned}
\|\mathcal{H}\|_{F,\overline{r}} &= \max_{i \in [N-1]} \sqrt{\sum_{j=1}^{r_i} \sigma_j^2(\mathcal{H}^{\langle i \rangle})} \\
&= \max_{\substack{\boldsymbol{V}_i \in \mathbb{R}^{d_{i+1} \cdots d_N \times r_i}, \\ \boldsymbol{V}_i \boldsymbol{V}_i^\top = \mathbf{I}_{r_i}, i \in [N-1]}} \|\mathcal{H}^{\langle i \rangle} \boldsymbol{V}_i\|_F \\
&= \max_{\substack{\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N}, \|\mathcal{X}\|_F \leq 1, \\ \text{rank}(\mathcal{X}) = (r_1, \ldots, r_{N-1})}} \langle \mathcal{H}, \mathcal{X} \rangle, \tag{88}
\end{aligned}
$$

where $\text{rank}(\mathcal{X})$ denotes the TT ranks of $\mathcal{X}$. Similar forms for the matrix case are provided in (Zhang et al., 2021a; Tong, 2022). We now upper bound $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F$ as

$$
\begin{aligned}
& \|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \\
&= \left\| \text{SVD}_{\boldsymbol{r}}^{tt} \left(\frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k \right) - \mathcal{X}^\star \right\|_{F,2\overline{r}} \\
&\leq \left\| \text{SVD}_{\boldsymbol{r}}^{tt} \left(\frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k \right) - \frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k \right\|_{F,2\overline{r}} + \left\| \frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k - \mathcal{X}^\star \right\|_{F,2\overline{r}} \\
&\leq \sqrt{N-1} \left\| \text{opt}_{\boldsymbol{r}} \left(\frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k \right) - \frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k \right\|_{F,2\overline{r}} + \left\| \frac{1}{m} \sum_{k=1}^m y_k \mathcal{A}_k - \mathcal{X}^\star \right\|_{F,2\overline{r}}
\end{aligned}
$$

$$\leq (1 + \sqrt{N-1}) \left\| \frac{1}{m} \sum_{k=1}^{m} y_k \mathcal{A}_k - \mathcal{X}^\star \right\|_{F, 2\overline{r}}$$

$$= (1 + \sqrt{N-1}) \max_{\substack{\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_N}, \|\mathcal{Z}\|_F \leq 1, \\ \mathrm{rank}(\mathcal{Z}) = (2r_1, \ldots, 2r_{N-1})}} \left| \left\langle \left(\frac{1}{m} \mathcal{A}^* \mathcal{A} - \mathcal{I} \right)(\mathcal{X}^\star), \mathcal{Z} \right\rangle \right|$$

$$\leq \delta_{3\overline{r}}(1 + \sqrt{N-1}) \|\mathcal{X}^\star\|_F, \tag{89}$$

where $\mathrm{opt}_{\boldsymbol{r}}(\frac{1}{m} \sum_{k=1}^{m} y_k \mathcal{A}_k)$ is the best TT-approximation of ranks $\boldsymbol{r}$ to $\frac{1}{m} \sum_{k=1}^{m} y_k \mathcal{A}_k$ in the Frobenius norm, the second inequality utilizes the quasi-optimality property of TT-SVD projection (Oseledets, 2011), the third inequality follows because the definition of $\mathrm{opt}_{\boldsymbol{r}}(\cdot)$ and $\mathcal{X}^\star$ has ranks $\boldsymbol{r}$, and the last uses (87). ∎

Appendix D. Proof of Theorem 7 in Tensor-train Sensing

Proof The proof follows a similar approach to that for Theorem 2 in Appendix B. We first present useful properties for the factors $L(\boldsymbol{X}_i^{(t)})$. Due to the retraction, $L(\boldsymbol{X}_i^{(t)}), i \in [N-1]$ are always orthonormal. Assuming that

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{(4 - 15\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{8(N + 1 + \sum_{i=2}^{N-1} r_i)(57N^2 + 393N - 450)}, \tag{90}$$

which is true for $t = 0$ and will be proved later for $t \geq 1$, based on the derivation of (71), we have

$$\sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) = \|\mathcal{X}^{(t)\geq i+1}\|^2 \leq 2\overline{\sigma}^2(\mathcal{X}^\star) + \left(4 + \frac{16\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)} \right) \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2$$

$$\leq 2\overline{\sigma}^2(\mathcal{X}^\star) + \frac{45N\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)} \mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1]. \tag{91}$$

Note that $\overline{\sigma}^2(\mathcal{X}^{(t)}) = \max_{i=1}^{N-1} \sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$.

We now prove the decay of the distance based on the result $\|\mathcal{X}^{(t)\geq i+1}\|^2 \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1]$. First recall (18):

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\leq \mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) - 2\mu \sum_{i=1}^{N} \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}} \left(\nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \right) \right\rangle$$

$$+ \mu^2 \left(\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}} \left(\nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 + \left\| \nabla_{L(\boldsymbol{X}_N)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \right\|_2^2 \right). \tag{92}$$

Note that the gradient is defined as

$$\nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) = \begin{bmatrix} \nabla_{\boldsymbol{X}_i(1)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \\ \vdots \\ \nabla_{\boldsymbol{X}_i(d_i)} g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \end{bmatrix}, \tag{93}$$

33

where the gradient with respect to each factor $\boldsymbol{X}_i(s_i)$ can be obtained as

$$\nabla_{\boldsymbol{X}_i(s_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) = \frac{1}{m} \sum_{k=1}^m (\langle \mathcal{A}_k, \mathcal{X}^{(t)} \rangle - y_k) \sum_{s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_N} \Big(\mathcal{A}_k(s_1, \dots, s_N) \cdot$$
$$\boldsymbol{X}_{i-1}^{(t)}(s_{i-1})^\top \cdots \boldsymbol{X}_1^{(t)}(s_1)^\top \boldsymbol{X}_N^{(t)}(s_N)^\top \cdots \boldsymbol{X}_{i+1}^{(t)}(s_{i+1})^\top \Big).$$

Upper bound of the third term in (92) Using the RIP, we begin by quantifying the difference in the gradients of g and f through

$$\left\| \nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) - \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F$$
$$= \max_{\substack{\boldsymbol{H}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ \|\boldsymbol{H}_i\|_F \le 1}} \langle \nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) - \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}), L(\boldsymbol{H}_i) \rangle$$
$$= \max_{\substack{\boldsymbol{H}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ \|\boldsymbol{H}_i\|_F \le 1}} \langle (\frac{1}{m} \mathcal{A}^* \mathcal{A} - \mathcal{I})(\mathcal{X}^{(t)} - \mathcal{X}^\star), [\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{H}_i, \dots, \boldsymbol{X}_N^{(t)}] \rangle$$
$$\le \delta_{3\bar{r}} \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F \|[\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{H}_i, \dots, \boldsymbol{X}_N^{(t)}]\|_F$$
$$\le \begin{cases} \frac{3\bar{\sigma}(\mathcal{X}^\star)}{2} \delta_{3\bar{r}} \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F, & i \in [N-1], \\ \delta_{3\bar{r}} \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F, & i = N, \end{cases} \tag{94}$$

where the first inequality follows (87) and the second inequality uses (47) and $\|[\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{H}_i, \dots, \boldsymbol{X}_N^{(t)}]\|_F = \|(L(\boldsymbol{X}_1^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{H}_i)) \mathcal{X}^{(t) \ge i+1}\|_F \le \|L(\boldsymbol{X}_1^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{H}_i)\|_F \|\mathcal{X}^{(t) \ge i+1}\|$. This together with the upper bound for $\left\| \nabla_{L(\boldsymbol{X}_i)} f(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F$ in (76) gives

$$\left\| \nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F \le \begin{cases} \frac{3\bar{\sigma}(\mathcal{X}^\star)}{2} (1 + \delta_{3\bar{r}}) \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F, & i \in [N-1], \\ (1 + \delta_{3\bar{r}}) \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F, & i = N. \end{cases} \tag{95}$$

Plugging this into the third term in (92) and following the same analysis of (77), we can obtain

$$\frac{1}{\bar{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\| \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)} \mathrm{St}} \left(\nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right) \right\|_F^2 + \left\| \nabla_{L(\boldsymbol{X}_N)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_2^2$$
$$\le \frac{1}{\bar{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\| \nabla_{L(\boldsymbol{X}_i)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_F^2 + \left\| \nabla_{L(\boldsymbol{X}_N)} g(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}) \right\|_2^2$$
$$\le \frac{9N - 5}{2} (1 + \delta_{3\bar{r}})^2 \|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2. \tag{96}$$

Lower bound of the second term in (92)   We first expand the second term of (92) as follows:

$$\sum_{i=1}^{N}\left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$= \sum_{i=1}^{N}\left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\rangle - T_2$$

$$= \frac{1}{m}\sum_{k=1}^{m}\langle\boldsymbol{a}_k, L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star)\rangle\langle\boldsymbol{a}_k,\boldsymbol{h}^{(t)}\rangle$$

$$+ \frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(t)} - \mathcal{X}^\star)\|_2^2 - T_2, \tag{97}$$

where $\boldsymbol{a}_k = \mathrm{vec}(\mathcal{A}_k)$ and T_2 is defined as

$$T_2 = \sum_{i=1}^{N-1}\left\langle\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)), \nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\rangle. \tag{98}$$

T_2 can be upper bonded by

$$T_2 \le \frac{1}{2}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)})\|\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2\left\|\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F$$

$$\le \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F$$

$$\le \frac{1}{10}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \frac{45(N-1)\overline{\sigma}^2(\mathcal{X}^\star)}{8}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^4$$

$$\le \frac{1}{10}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \frac{45(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\}), \tag{99}$$

where the second inequality follows (95) with $\delta_{3\overline{r}} = 1$. We now plug this into (97) to get

$$\sum_{i=1}^{N}\left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$\ge (1 - \delta_{2\overline{r}})\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \langle L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star),\boldsymbol{h}^{(t)}\rangle$$

$$- \delta_{(N+3)\overline{r}}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F\|\boldsymbol{h}^{(t)}\|_2 - \frac{1}{10}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{45(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})$$

$$\ge \left(\frac{9}{10} - \delta_{(N+3)\overline{r}}\right)\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{1 + \delta_{(N+3)\overline{r}}}{2}(\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \|\boldsymbol{h}^{(t)}\|_2^2)$$

$$- \frac{45(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})$$

$$\ge \frac{4 - 15\delta_{(N+3)\overline{r}}}{20}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{19}{30}\|\boldsymbol{h}^{(t)}\|_2^2 - \frac{45(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})$$

$$\ge \frac{(4 - 15\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{640(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\}) + \frac{4 - 15\delta_{(N+3)\overline{r}}}{40}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2, \tag{100}$$

where we utilize (99), Theorem 5, and Lemma 18 in the first inequality. Note that according to the definition of $\boldsymbol{h}^{(t)}$ in (79), it has TT ranks at most $((N+1)r_1, \ldots, (N+1)r_{N-1})$. Therefore, with the TT format $L(\boldsymbol{X}_1^{(t)})\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star)$, which has TT ranks $(2r_1, \ldots, 2r_{N-1})$, the measurement operator $\mathcal{A}$ needs to satisfy the $(N+3)\overline{r}$-RIP which is assumed. The third inequality follows because $\delta_{2\overline{r}} \leq \delta_{(N+3)\overline{r}} \leq \frac{4}{15}$. The last line utilizes Lemma 16, (82), and the initial condition $\text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{(4-15\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{8(N+1+\sum_{i=2}^{N-1}r_i)(57N^2+393N-450)}$.

Contraction Taking (96) and (100) into (92), we can get

$$
\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})
$$
$$
\leq \left(1 - \frac{(4-15\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{320(N+1+\sum_{i=2}^{N-1}r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right)\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})
$$
$$
+\left(\frac{9N-5}{2}(1+\delta_{3\overline{r}})^2\mu^2 - \frac{4-15\delta_{(N+3)\overline{r}}}{20}\mu\right)\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2
$$
$$
\leq \left(1 - \frac{(4-15\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{320(N+1+\sum_{i=2}^{N-1}r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right)\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), \tag{101}
$$

where we use $\mu \leq \frac{4-15\delta_{(N+3)\overline{r}}}{10(9N-5)(1+\delta_{(N+3)\overline{r}})^2}$ in the last line.

Proof of (90) This can be proved by using the same induction argument in (70) together with $\delta_{(N+3)\overline{r}} \leq \frac{4}{15}$. This completes the proof.

∎

Appendix E. Proof of Theorem 9 for Noisy Spectral Initialization

Proof

Recalling the definition of $\|\cdot\|_{F,\overline{r}}$ in (88), we follow the same approach in (89) to quantify $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F$:

$$
\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F
$$
$$
= \left\|\text{SVD}_{\boldsymbol{r}}^{tt}\left(\frac{1}{m}\sum_{k=1}^m (y_k + \epsilon_k)\mathcal{A}_k\right) - \mathcal{X}^\star\right\|_{F,2\overline{r}}
$$
$$
\leq (1+\sqrt{N-1})\left\|\frac{1}{m}\sum_{k=1}^m (y_k + \epsilon_k)\mathcal{A}_k - \mathcal{X}^\star\right\|_{F,2\overline{r}}
$$
$$
\leq (1+\sqrt{N-1})\left\|\frac{1}{m}\sum_{k=1}^m y_k\mathcal{A}_k - \mathcal{X}^\star\right\|_{F,2\overline{r}} + (1+\sqrt{N-1})\left\|\frac{1}{m}\sum_{k=1}^m \epsilon_k\mathcal{A}_k\right\|_{F,2\overline{r}}
$$
$$
\leq \delta_{3\overline{r}}(1+\sqrt{N-1})\|\mathcal{X}^\star\|_F + (1+\sqrt{N-1})\left\|\frac{1}{m}\sum_{k=1}^m \epsilon_k\mathcal{A}_k\right\|_{F,2\overline{r}}. \tag{102}
$$

Next, we will use an ϵ-net and a covering argument to bound the second term in the last line:

$$\left\|\frac{1}{m}\sum_{k=1}^{m}\epsilon_k\mathcal{A}_k\right\|_{F,2\overline{r}} = \max_{\substack{\mathcal{H}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{H}\|_F\leq 1,\\\mathrm{rank}(\mathcal{H})=(2r_1,\ldots,2r_{N-1})}}\left\langle\frac{1}{m}\sum_{k=1}^{m}\epsilon_k\mathcal{A}_k,\mathcal{H}\right\rangle = \max_{\substack{\mathcal{H}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{H}\|_F\leq 1,\\\mathrm{rank}(\mathcal{H})=(2r_1,\ldots,2r_{N-1})}}\frac{1}{m}\langle\boldsymbol{\epsilon},\mathcal{A}(\mathcal{H})\rangle. \quad (103)$$

To begin, according to (Zhang et al., 2018), for each $i\in[N-1]$, we can construct an ϵ-net $\{L(\boldsymbol{H}_i^{(1)}),\ldots,L(\boldsymbol{H}_i^{(n_i)})\}$ with the covering number $n_i\leq(\frac{4+\epsilon}{\epsilon})^{d_ir_{i-1}r_i}$ for the set of factors $\{L(\boldsymbol{H}_i)\in\mathbb{R}^{d_ir_{i-1}\times r_i}:\|L(\boldsymbol{H}_i)\|\leq 1\}$ such that

$$\sup_{L(\boldsymbol{H}_i):\|L(\boldsymbol{H}_i)\|\leq 1}\min_{p_i\leq n_i}\|L(\boldsymbol{H}_i)-L(\boldsymbol{H}_i^{(p_i)})\|\leq\epsilon. \quad (104)$$

Similarly, we can construct an ϵ-net $\{L(\boldsymbol{H}_N^{(1)}),\ldots,L(\boldsymbol{H}_N^{(n_N)})\}$ with the covering number $n_N\leq(\frac{2+\epsilon}{\epsilon})^{d_Nr_{N-1}}$ for $\{L(\boldsymbol{H}_N)\in\mathbb{R}^{d_Nr_{N-1}\times 1}:\|L(\boldsymbol{H}_N)\|_2\leq 1\}$ such that

$$\sup_{L(\boldsymbol{H}_N):\|L(\boldsymbol{H}_N)\|_2\leq 1}\min_{p_N\leq n_N}\|L(\boldsymbol{H}_N)-L(\boldsymbol{H}_N^{(p_N)})\|_2\leq\epsilon. \quad (105)$$

Therefore, we can construct an ϵ-net $\{\mathcal{H}^{(1)},\ldots,\mathcal{H}^{(n_1\cdots n_N)}\}$ with covering number

$$\Pi_{i=1}^{N}n_i\leq\left(\frac{4+\epsilon}{\epsilon}\right)^{d_1r_1+\sum_{i=2}^{N-1}d_ir_{i-1}r_i+d_Nr_{N-1}}\leq\left(\frac{4+\epsilon}{\epsilon}\right)^{N\overline{d}\overline{r}^2}$$

(where $\overline{r}=\max_{i=1}^{N-1}r_i$ and $\overline{d}=\max_{i=1}^{N}d_i$) for any TT format tensors $\mathcal{H}=[\boldsymbol{H}_1,\ldots,\boldsymbol{H}_N]\in\mathbb{R}^{d_1\times\cdots\times d_N}$ with TT ranks $(r_1,\ldots,r_{N-1})$.

Denote by T the value of (103), i.e.,

$$[\widetilde{\boldsymbol{H}}_1,\ldots,\widetilde{\boldsymbol{H}}_N]=\mathop{\arg\max}_{\substack{L(\boldsymbol{H}_i)\in\mathbb{R}^{2d_ir_{i-1}\times 2r_i}\\\|L(\boldsymbol{H}_i)\|\leq 1,i\in[N-1]\\\|L(\boldsymbol{H}_N)\|_2\leq 1}}\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\boldsymbol{H}_1,\ldots,\boldsymbol{H}_N]\rangle, \quad (106)$$

$$T:=\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\widetilde{\boldsymbol{H}}_1,\ldots,\widetilde{\boldsymbol{H}}_N]\rangle. \quad (107)$$

Using $\mathcal{I}$ to denote the index set $[n_1]\times\cdots\times[n_N]$, then according to the construction of the ϵ-net, there exists $p=(p_1,\ldots,p_N)\in\mathcal{I}$ such that

$$\|L(\widetilde{\boldsymbol{H}}_i)-L(\boldsymbol{H}_i^{(p_i)})\|\leq\epsilon,\ i\in[N-1]\ \text{ and }\ \|L(\widetilde{\boldsymbol{H}}_N)-L(\boldsymbol{H}_N^{(p_N)})\|_2\leq\epsilon. \quad (108)$$

Now taking $\epsilon=\frac{1}{2N}$ gives

$$T=\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_N^{(p_N)}]\rangle+\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\widetilde{\boldsymbol{H}}_1,\ldots,\widetilde{\boldsymbol{H}}_N]-[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_N^{(p_N)}]\rangle$$

$$=\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_N^{(p_N)}]\rangle+\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,\sum_{a_1=1}^{N}[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_{a_1}^{(p_{a_1})}-\widetilde{\boldsymbol{H}}_{a_1},\ldots,\widetilde{\boldsymbol{H}}_N]\rangle$$

$$\leq\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_N^{(p_N)}]\rangle+N\epsilon T$$

$$=\frac{1}{m}\sum_{k=1}^{m}\langle\epsilon_k\mathcal{A}_k,[\boldsymbol{H}_1^{(p_1)},\ldots,\boldsymbol{H}_N^{(p_N)}]\rangle+\frac{T}{2}, \quad (109)$$

where the second line uses Lemma 13 to rewrite $[\widetilde{\boldsymbol{H}}_1, \ldots, \widetilde{\boldsymbol{H}}_N] - [\boldsymbol{H}_1^{(p_1)}, \ldots, \boldsymbol{H}_N^{(p_N)}]$ into a sum of N terms.

Note that when conditioned on $\{\mathcal{A}_k\}_{k=1}^m$, for any fixed $\mathcal{H}^{(p)} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, $\frac{1}{m}\langle \boldsymbol{\epsilon}, \mathcal{A}(\mathcal{H}^{(p)})\rangle$ has a normal distribution with zero mean and variance $\frac{\gamma^2 \|\mathcal{A}(\mathcal{H}^{(p)})\|_2^2}{m^2}$, which implies that

$$\mathbb{P}\left(\frac{1}{m}|\langle \boldsymbol{\epsilon}, \mathcal{A}(\mathcal{H}^{(p)})\rangle| \geq t | \{\mathcal{A}_k\}_{k=1}^m \right) \leq e^{-\frac{m^2 t^2}{2\gamma^2 \|\mathcal{A}(\mathcal{H}^{(p)})\|_2^2}}. \tag{110}$$

Furthermore, under the event $F := \{\mathcal{A} \text{ satisfies } 2\overline{r}\text{-RIP with constant } \delta_{2\overline{r}}\}$, which implies that $\frac{1}{m}\|\mathcal{A}(\mathcal{H}^{(p)})\|_2^2 \leq (1 + \delta_{2\overline{r}})\|\mathcal{H}^{(p)}\|_F^2$, plugging this together with the fact $\|\mathcal{H}^{(p)}\|_F \leq 1$ into the above further gives

$$\mathbb{P}\left(\frac{1}{m}|\langle \boldsymbol{\epsilon}, \mathcal{A}(\mathcal{H}^{(p)})\rangle| \geq t | F \right) \leq e^{-\frac{mt^2}{2(1+\delta_{2\overline{r}})\gamma^2}}. \tag{111}$$

We now apply this tail bound to (109) and get

$$\mathbb{P}(T \geq t | F) \leq \mathbb{P}\left(\max_{p_1, \ldots, p_n} \frac{1}{m}\sum_{k=1}^m \langle \epsilon_k \mathcal{A}_k, [\boldsymbol{H}_1^{(p_1)}, \ldots, \boldsymbol{H}_N^{(p_N)}]\rangle \geq \frac{t}{2} | F \right)$$

$$\leq \left(\frac{4+\epsilon}{\epsilon}\right)^{4N\overline{d}\overline{r}^2} e^{-\frac{mt^2}{8(1+\delta_{2\overline{r}})\gamma^2}} \leq e^{-\frac{mt^2}{8(1+\delta_{2\overline{r}})\gamma^2} + c_1 N\overline{d}\overline{r}^2 \log N}, \tag{112}$$

where c_1 is a constant and based on the assumption in (109), $\frac{4+\epsilon}{\epsilon} = \frac{4+\frac{1}{2N}}{\frac{1}{2N}} = 8N + 1$.

Hence, we can take $t = \frac{c_2 \overline{r}\sqrt{(1+\delta_{2\overline{r}})N\overline{d}(\log N)}}{\sqrt{m}}\gamma$ with a constant c_2 and further derive

$$\mathbb{P}\left(T \leq \frac{c_2 \overline{r}\sqrt{(1+\delta_{2\overline{r}})N\overline{d}(\log N)}}{\sqrt{m}}\gamma\right)$$

$$\geq \mathbb{P}\left(T \leq \frac{c_2 \overline{r}\sqrt{(1+\delta_{2\overline{r}})N\overline{d}(\log N)}}{\sqrt{m}}\gamma \cap F\right)$$

$$\geq P(F)\mathbb{P}\left(T \leq \frac{c_2 \overline{r}\sqrt{(1+\delta_{2\overline{r}})N\overline{d}(\log N)}}{\sqrt{m}}\gamma | F\right)$$

$$\geq (1 - e^{-c_3 N\overline{d}\overline{r}^2 \log N})(1 - e^{-c_4 N\overline{d}\overline{r}^2 \log N}) \geq 1 - 2e^{-c_5 N\overline{d}\overline{r}^2 \log N}, \tag{113}$$

where $c_i, i = 3, 4, 5$ are constants. Note that $P(F)$ is obtained via Theorem 5 by setting ϵ in (24) to be $e^{-c_3 N\overline{d}\overline{r}^2 \log N}$.

Combing (102) and (113), we finally obtain that with probability $1 - 2e^{-c_5 N\overline{d}\overline{r}^2 \log N}$,

$$\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \leq (1 + \sqrt{N-1})\left(\delta_{3\overline{r}}\|\mathcal{X}^\star\|_F + \frac{c_2 \overline{r}\sqrt{(1+\delta_{3\overline{r}})N\overline{d}\log N}}{\sqrt{m}}\gamma\right), \tag{114}$$

where $\delta_{2\overline{r}} \leq \delta_{3\overline{r}}$ is used.

$\blacksquare$

Appendix F. Proof of Theorem 10 for Noisy TT Format Tensor Sensing

Proof By the same analysis in the beginning of Appendix D, we can get that $L(\boldsymbol{X}_i^{(t)}), i \in [N-1]$ are orthonormal matrices and $\sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) = \|\mathcal{X}^{(t) \geq i+1}\|^2 \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1], t \geq 0$ by assuming

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\sigma^2(\mathcal{X}^\star)}{180N}, \tag{115}$$

which will be proved later. Now recall (18):

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$$

$$\leq \mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) - 2\mu \sum_{i=1}^N \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\rangle$$

$$+ \mu^2 \left(\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \left\|\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right)\right\|_F^2 + \left\|\nabla_{L(\boldsymbol{X}_N)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\|_2^2\right), \tag{116}$$

where the gradient with respect to each factor $\boldsymbol{X}_i(s_i)$ can be computed as

$$\nabla_{\boldsymbol{X}_i(s_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) = \frac{1}{m} \sum_{k=1}^m (\langle \mathcal{A}_k, \mathcal{X}^{(t)} \rangle - y_k - \epsilon_k) \sum_{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_N} \left(\mathcal{A}_k(s_1, \ldots, s_N) \cdot \right.$$

$$\left. \boldsymbol{X}_{i-1}^{(t)}(s_{i-1})^\top \cdots \boldsymbol{X}_1^{(t)}(s_1)^\top \boldsymbol{X}_N^{(t)}(s_N)^\top \cdots \boldsymbol{X}_{i+1}^{(t)}(s_{i+1})^\top\right).$$

Upper bound of the third term in (116)   To upper bound $\|\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\|_F$, we first analyze the difference in the gradient caused by noise, using the same analysis in (113). Specifically, with the same ϵ-net argument in (103),

$$\left\|\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) - \nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)})\right\|_F$$

$$= \max_{\substack{\boldsymbol{H}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ \|\boldsymbol{H}_i\|_F \leq 1}} \left\langle \nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) - \nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}), L(\boldsymbol{H}_i)\right\rangle$$

$$= \max_{\substack{\boldsymbol{H}_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i} \\ \|\boldsymbol{H}_i\|_F \leq 1}} \left\langle \frac{1}{m} \sum_{k=1}^m \epsilon_k \mathcal{A}_k, [\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_{i-1}^{(t)}, \boldsymbol{H}_i, \boldsymbol{X}_{i+1}^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}]\right\rangle$$

$$\leq \begin{cases} \dfrac{c_i \overline{r} \sqrt{(1+\delta_{3\overline{r}})N\overline{d}(\log N)}\gamma \overline{\sigma}(\mathcal{X}^\star)}{\sqrt{m}}, & i = 1, \ldots, N-1, \\ \dfrac{c_N \overline{r} \sqrt{(1+\delta_{3\overline{r}})N\overline{d}(\log N)}\gamma}{\sqrt{m}}, & i = N, \end{cases}$$

where the last inequality holds with probability at least $1 - 2Ne^{-\Omega(N\overline{d}\overline{r}^2 \log N)}$ with $c_i, i \in [N]$ being positive constants, and is derived by using (47) that $\|[\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{H}_i, \ldots, \boldsymbol{X}_N^{(t)}]\|_F = \|(L(\boldsymbol{X}_1^{(t)})\overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{H}_i))\mathcal{X}^{(t) \geq i+1}\|_F \leq \|L(\boldsymbol{X}_1^{(t)})\overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{H}_i)\|_F \|\mathcal{X}^{(t) \geq i+1}\| \leq \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}, i \in [N-1]$ and $\|[\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_{N-1}^{(t)}, \boldsymbol{H}_N]\|_F = \|\boldsymbol{H}_N\|_F \leq 1$.

This together with the bound for $\left\|\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F$ in (95) gives

$$\left\|\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F$$

$$\leq \left\|\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F + \left\|\nabla_{L(\boldsymbol{X}_i)}g(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)}) - \nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F$$

$$\leq \begin{cases} \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}(1+\delta_{3\overline{r}})\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F + \frac{c_i\overline{r}\sqrt{(1+\delta_{3\overline{r}})N\overline{d}(\log N)}\gamma\overline{\sigma}(\mathcal{X}^\star)}{\sqrt{m}}, & i=1,\ldots,N-1, \\ (1+\delta_{3\overline{r}})\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F + \frac{c_N\overline{r}\sqrt{(1+\delta_{3\overline{r}})N\overline{d}(\log N)}\gamma}{\sqrt{m}}, & i=N. \end{cases} \tag{117}$$

We now plug the above into the third term in (113) to get

$$\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)}\sum_{i=1}^{N-1}\left\|\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\left(\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right)\right\|_F^2 + \left\|\nabla_{L(\boldsymbol{X}_N)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_2^2$$

$$\leq \frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)}\sum_{i=1}^{N-1}\left\|\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F^2 + \left\|\nabla_{L(\boldsymbol{X}_N)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_2^2$$

$$\leq \frac{9N-5}{2}(1+\delta_{3\overline{r}})^2\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F^2 + O\left(\frac{(1+\delta_{3\overline{r}})N^2\overline{d}\overline{r}^2(\log N)\gamma^2}{m}\right). \tag{118}$$

Lower bound of the second term in (116)   To apply the same approach as in (100) for establishing a lower bound for the second term in (116), we first need to establish upper bounds for two terms involving noise. To begin, following the derivation of (99), we can get

$$\sum_{i=1}^{N-1}\left\langle \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^\perp(L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)), \nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\rangle$$

$$\leq \frac{1}{2}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)})\|\|L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2\left\|\nabla_{L(\boldsymbol{X}_i)}G(\boldsymbol{X}_1^{(t)},\ldots,\boldsymbol{X}_N^{(t)})\right\|_F$$

$$\leq \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F$$

$$+\sum_{i=1}^{N-1}\frac{c_i\overline{r}\sqrt{(1+\delta_{3\overline{r}})N\overline{d}(\log N)}\gamma\overline{\sigma}(\mathcal{X}^\star)}{\sqrt{m}}\|L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^2$$

$$\leq \frac{1}{20}\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F^2 + 46(N-1)\overline{\sigma}^2(\mathcal{X}^\star)\sum_{i=1}^{N-1}\|L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)\|_F^4$$

$$+\sum_{i=1}^{N-1}\frac{c_i^2(1+\delta_{3\overline{r}})N\overline{d}\overline{r}^2(\log N)\gamma^2}{16m}$$

$$\leq \frac{1}{20}\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F^2 + \frac{46(N-1)}{\overline{\sigma}^2(\mathcal{X}^\star)}\mathrm{dist}^4(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\}) + O\left(\frac{(1+\delta_{3\overline{r}})N^2\overline{d}\overline{r}^2(\log N)\gamma^2}{m}\right), \tag{119}$$

where the second inequality uses (117). In addition, recalling the notations of $\boldsymbol{a}_k = \text{vec}(\mathcal{A}_k)$ and $\boldsymbol{h}^{(t)}$ defined in (79), then with probability $1 - 2e^{-\Omega(N^3 \overline{dr}^2 \log N)}$, we have

$$
\frac{1}{m} \sum_{k=1}^{m} \langle \epsilon_k \boldsymbol{a}_k, L(\boldsymbol{X}_1^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star) + \boldsymbol{h}^{(t)} \rangle
$$

$$
\leq \frac{C(N+3)\overline{r}\sqrt{(1 + \delta_{(N+3)\overline{r}})N\overline{d}(\log N)\gamma}}{\sqrt{m}} \| L(\boldsymbol{X}_1^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_N^{(t)})
$$

$$
- L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star) + \boldsymbol{h}^{(t)} \|_F
$$

$$
\leq \frac{5C^2(1 + \delta_{(N+3)\overline{r}})N(N+3)^2\overline{dr}^2(\log N)\gamma^2}{m} + \frac{1}{10}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 + \frac{1}{10}\|\boldsymbol{h}^{(t)}\|_F^2
$$

$$
\leq \frac{5C^2(1 + \delta_{(N+3)\overline{r}})N(N+3)^2\overline{dr}^2(\log N)\gamma^2}{m} + \frac{1}{10}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2
$$

$$
+ \frac{9N(N-1)}{80\overline{\sigma}^2(\mathcal{X}^\star)} \text{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}), \tag{120}
$$

where the first inequality follows the same ϵ-net argument used in (103) and the fact that the TT ranks of the second term in the cross term is $((N+3)r_1, \ldots, (N+3)r_{N-1})$, and the last inequality uses (82).

Using (119), we can proceed with the analysis similar to (100) to obtain the following derivation

$$
\sum_{i=1}^{N} \left\langle L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}} \left(\nabla_{L(\boldsymbol{X}_i)} G(\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_N^{(t)}) \right) \right\rangle
$$

$$
\geq (\frac{9}{20} - \frac{3\delta_{(N+3)\overline{r}}}{2})\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{1 + \delta_{(N+3)\overline{r}}}{2}\|\boldsymbol{h}^{(t)}\|_F^2 - \frac{46(N-1)}{\overline{\sigma}^2(\mathcal{X}^\star)}\text{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})
$$

$$
- \frac{1}{m} \sum_{k=1}^{m} \langle \epsilon_k \boldsymbol{a}_k, L(\boldsymbol{X}_1^{(t)}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star) + \boldsymbol{h}^{(t)} \rangle
$$

$$
- O\left(\frac{(1 + \delta_{3\overline{r}})N^2\overline{dr}^2(\log N)\gamma^2}{m} \right)
$$

$$
\geq \frac{7 - 30\delta_{(N+3)\overline{r}}}{40}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \frac{129N^2 + 7231N - 7360}{160\overline{\sigma}^2(\mathcal{X}^\star)}\text{dist}^4(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})
$$

$$
- O\left(\frac{(1 + \delta_{(N+3)\overline{r}})N^3\overline{dr}^2(\log N)\gamma^2}{m} \right)
$$

$$
\geq \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{1280(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) + \frac{7 - 30\delta_{(N+3)\overline{r}}}{80}\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2
$$

$$
- O\left(\frac{(1 + \delta_{(N+3)\overline{r}})N^3\overline{dr}^2(\log N)\gamma^2}{m} \right), \tag{121}
$$

where we use $\delta_{(N+3)\overline{r}} \leq \frac{7}{30}$, (120) and (82) in the second inequality, and the last line follows Lemma 16 and the initial condition $\text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{8(N + 1 + \sum_{i=2}^{N-1} r_i)(129N^2 + 7231N - 7360)}$.

41

Contraction   Taking (118) and (121) into (116), with probability $1-2Ne^{-\Omega(N\overline{dr}^2 \log N)}-2e^{-\Omega(N^3\overline{dr}^2 \log N)}$, we can get

$$
\begin{aligned}
\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) &\leq \left(1 - \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{1280(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right) \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\quad + \left(\frac{9N - 5}{2}(1 + \delta_{3\overline{r}})^2\mu^2 - \frac{7 - 30\delta_{(N+3)\overline{r}}}{40}\mu\right)\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 \\
&\quad + O\left(\frac{(1 + \delta_{(N+3)\overline{r}})N^2\overline{dr}^2(\log N)\gamma^2}{m}(\mu N + \mu^2)\right) \\
&\leq \left(1 - \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{1280(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right) \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\quad + O\left(\frac{(1 + \delta_{(N+3)\overline{r}})N^2\overline{dr}^2(\log N)\gamma^2}{m}(\mu N + \mu^2)\right),
\end{aligned}
\tag{122}
$$

where we use $\mu \leq \frac{7 - 30\delta_{(N+3)\overline{r}}}{20(9N - 5)(1 + \delta_{(N+3)\overline{r}})^2}$ in the last line. By induction, this further implies that

$$
\begin{aligned}
&\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\leq \left(1 - \frac{(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}{1280(N + 1 + \sum_{i=2}^{N-1} r_i)\overline{\sigma}^2(\mathcal{X}^\star)}\mu\right)^{t+1} \text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\quad + O\left(\frac{(N + \mu)(N + 1 + \sum_{i=2}^{N-1} r_i)(1 + \delta_{(N+3)\overline{r}})N^2\overline{dr}^2(\log N)\overline{\sigma}^2(\mathcal{X}^\star)\gamma^2}{m(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}\right).
\end{aligned}
\tag{123}
$$

Proof of (115)   We can prove it by induction as used in the proof of (70). First note that (90) holds for $t = 0$. We now assume it holds for all $t \leq t'$, which implies that $\sigma_1^2(\mathcal{X}^{(t')\langle i \rangle}) = \|\mathcal{X}^{(t')\geq i+1}\|^2 \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N - 1]$. By invoking (123), we have

$$
\begin{aligned}
&\text{dist}^2(\{\boldsymbol{X}_i^{(t'+1)}\}, \{\boldsymbol{X}_i^\star\}) \\
&\leq \text{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) + O\left(\frac{(N + \mu)(N + 1 + \sum_{i=2}^{N-1} r_i)(1 + \delta_{(N+3)\overline{r}})N^2\overline{dr}^2(\log N)\overline{\sigma}^2(\mathcal{X}^\star)\gamma^2}{m(7 - 30\delta_{(N+3)\overline{r}})\underline{\sigma}^2(\mathcal{X}^\star)}\right) \\
&\leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{180N},
\end{aligned}
$$

as long as $m \geq C\frac{N^5\overline{dr}^3(\log N)\overline{\sigma}^2(\mathcal{X}^\star)\gamma^2}{\underline{\sigma}^4(\mathcal{X}^\star)}$ with a universal constant C. Consequently, (115) also holds at $t = t' + 1$. By induction, we can conclude that (115) holds for all $t \geq 0$. This completes the proof. $\blacksquare$

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Evrim Acar and Bülent Yener. Unsupervised multiway data analysis: A literature survey. *IEEE transactions on knowledge and data engineering*, 21(1):6–20, 2008.

Johann A Bengua, Ho N Phien, Hoang Duong Tuan, and Minh N Do. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Transactions on Image Processing*, 26(5):2466–2479, 2017.

Hermine Biermé and Céline Lacaux. Modulus of continuity of some conditionally sub-gaussian fields, application to stable random fields. *Bernoulli*, 21(3):1719–1759, 2015.

Rasmus Bro. Parafac. Tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

Stanislav Budzinskiy and Nikolai Zamarashkin. Tensor train completion: local recovery guarantees via riemannian optimization. *arXiv preprint arXiv:2110.03975*, 2021.

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.

Jian-Feng Cai, Jingyang Li, and Dong Xia. Provable tensor-train format tensor completion by riemannian optimization. *Journal of Machine Learning Research*, 23(123):1–77, 2022.

Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Andrzej Cichocki. Tensor networks for big data analytics and large-scale optimization problems. *arXiv preprint arXiv:1407.3124*, 2014.

Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163, 2015.

Vin De Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

Lijun Ding, Zhen Qin, Liwei Jiang, Jinxin Zhou, and Zhihui Zhu. A validation approach to overparameterized matrix and image recovery. *arXiv preprint arXiv:2209.10675*, 2022.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3):e1201, 2017.

Rachel Grotheer, Shuang Li, Anna Ma, Deanna Needell, and Jing Qin. Iterative hard thresholding for low cp-rank tensor models. *Linear and Multilinear Algebra*, pages 1–17, 2021.

Weiwei Guo, Irene Kotsia, and Ioannis Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2011.

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.

Botao Hao, Anru R Zhang, and Guang Cheng. Sparse and low-rank tensor estimation via cubic sketchings. In *International Conference on Artificial Intelligence and Statistics*, pages 1319–1330. PMLR, 2020.

Johan Håstad. Tensor rank is np-complete. In *Automata, Languages and Programming: 16th International Colloquium Stresa, Italy, July 11–15, 1989 Proceedings 16*, pages 451–460. Springer, 1989.

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

Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. On manifolds of tensors of fixed tt-rank. *Numerische Mathematik*, 120(4):701–731, 2012.

Victoria Hore, Ana Vinuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100, 2016.

Casey Jameson, Zhen Qin, Alireza Goldar, Michael B Wakin, Zhihui Zhu, and Zhexuan Gong. Optimal quantum state tomography with local informationally complete measurements. *arXiv preprint arXiv:2408.07115*, 2024.

Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*, 2022.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. *arXiv preprint arXiv:1711.00811*, 2017.

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

Maxim A Kuznetsov and Ivan V Oseledets. Tensor train spectral method for learning of hidden markov models (hmm). *Computational Methods in Applied Mathematics*, 19(1):93–99, 2019.

Jose I Latorre. Image compression and entanglement. *arXiv preprint quant-ph/0510031*, 2005.

Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

Shuang Li, Qiuwei Li, Zhihui Zhu, Gongguo Tang, and Michael B Wakin. The global geometry of centralized and distributed low-rank matrix recovery without regularization. *IEEE Signal Processing Letters*, 27:1400–1404, 2020a.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020b.

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

Alexander Lidiak, Casey Jameson, Zhen Qin, Gongguo Tang, Michael B Wakin, Zhihui Zhu, and Zhexuan Gong. Quantum state tomography with tensor train cross approximation. *arXiv preprint arXiv:2207.06397*, 2022.

Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional non-convex estimation. *Information and Inference: A Journal of the IMA*, 9(3):507–541, 2020.

Wangyu Luo, Wael Alghamdi, and Yue M Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356, 2019.

Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021.

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. *Advances in neural information processing systems*, 32, 2019.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

Georgii S Novikov, Maxim E Panov, and Ivan V Oseledets. Tensor-train density estimation. In *Uncertainty in artificial intelligence*, pages 1321–1331. PMLR, 2021.

Matthias Ohliger, Vincent Nesme, and Jens Eisert. Efficient and feasible state tomography of quantum many-body systems. *New Journal of Physics*, 15(1):015024, 2013.

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

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74. PMLR, 2017.

Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, and Javier Tejedor. Exploiting low-rank tensor-train deep neural networks based on riemannian gradient descent with illustrations of speech processing. *arXiv preprint arXiv:2203.06031*, 2022.

Zhen Qin and Zhihui Zhu. Computational and statistical guarantees for tensor-on-tensor regression with tensor train decomposition. *arXiv preprint arXiv:2406.06002*, 2024.

Zhen Qin, Casey Jameson, Zhexuan Gong, Michael B Wakin, and Zhihui Zhu. Quantum state tomography for matrix product density operators. *IEEE Transactions on Information Theory*, 70 (7):5030–5056, 2024.

Holger Rauhut, Reinhold Schneider, and Željka Stojanac. Tensor completion in hierarchical tensor representations. In *Compressed sensing and its applications*, pages 419–450. Springer, 2015.

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

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of physics*, 326(1):96–192, 2011.

Nicholas D Sidiropoulos, Georgios B Giannakis, and Rasmus Bro. Blind PARAFAC receivers for DS-CDMA systems. *IEEE Transactions on Signal Processing*, 48(3):810–823, 2000.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Age K Smilde, Paul Geladi, and Rasmus Bro. *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons, 2005.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

Edwin Stoudenmire and David J Schwab. Supervised learning with tensor networks. *Advances in neural information processing systems*, 29, 2016.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Compressing recurrent neural network with tensor train. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4451–4458. IEEE, 2017.

Tian Tong. *Scaled gradient methods for ill-conditioned low-rank matrix and tensor estimation*. PhD thesis, Carnegie Mellon University, 2022.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22:150–1, 2021a.

Tian Tong, Cong Ma, Ashley Prater-Bennette, Erin Tripp, and Yuejie Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *arXiv preprint arXiv:2104.14526*, Nov. 2021b.

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.

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

Frank Verstraete and J Ignacio Cirac. Matrix product states represent ground states faithfully. *Physical review b*, 73(9):094423, 2006.

Frank Verstraete, Valentin Murg, and J Ignacio Cirac. Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems. *Advances in physics*, 57(2):143–224, 2008.

René Vidal, Zhihui Zhu, and Benjamin D Haeffele. Optimization landscape of neural networks. *Mathematical Aspects of Deep Learning*, page 200, 2022.

Andong Wang, Xulin Song, Xiyin Wu, Zhihui Lai, and Zhong Jin. Latent schatten tt norm for tensor completion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2922–2926. IEEE, 2019a.

Junli Wang, Guangshe Zhao, Dingheng Wang, and Guoqi Li. Tensor completion using low-rank tensor train decomposition by riemannian optimization. In *2019 Chinese Automation Congress (CAC)*, pages 3380–3384. IEEE, 2019b.

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

Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. Tensor completion by alternating minimization under the tensor train (tt) model. *arXiv preprint arXiv:1609.05587*, 2016.

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

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. *arXiv preprint arXiv:2302.01186*, 2023.

Yinchong Yang, Denis Krompass, and Volker Tresp. Tensor-train recurrent neural networks for video classification. In *International Conference on Machine Learning*, pages 3891–3900. PMLR, 2017.

Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. Long-term forecasting using tensor-train rnns. *Arxiv*, 2017.

Longhao Yuan, Chao Li, Danilo Mandic, Jianting Cao, and Qibin Zhao. Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9151–9158, 2019a.

Longhao Yuan, Qibin Zhao, Lihua Gui, and Jianting Cao. High-order tensor completion via gradient-based optimization under tensor train format. *Signal Processing: Image Communication*, 73:53–61, 2019b.

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021a.

Jing Zhang, Xiaoli Ma, Jun Qi, and Shi Jin. Designing tensor-train deep neural networks for time-varying mimo channel estimation. *IEEE Journal of Selected Topics in Signal Processing*, 15(3): 759–773, 2021b.

Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. A tensor-train deep computation model for industry informatics big data feature learning. *IEEE Transactions on Industrial Informatics*, 14(7):3197–3204, 2018.

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

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

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.