# Robust Low-Rank Tensor Train Recovery

Zhen Qin and Zhihui Zhu, *Member, IEEE*

***Abstract*—Tensor train (TT) decomposition represents an $N$-order tensor using $O(N)$ matrices (i.e., factors) of small dimensions, achieved through products among these factors. Due to its compact representation, TT decomposition has found wide applications, including various tensor recovery problems in signal processing and quantum information. In this paper, we study the problem of reconstructing a TT format tensor from measurements that are contaminated by outliers with arbitrary values. Given the vulnerability of smooth formulations to corruptions, we use an $\ell_1$ loss function to enhance robustness against outliers. We first establish the $\ell_1/\ell_2$-restricted isometry property (RIP) for Gaussian measurement operators, demonstrating that the information in the TT format tensor can be preserved using a number of measurements that grows linearly with $N$. We also prove the sharpness property for the $\ell_1$ loss function optimized over TT format tensors. Building on the $\ell_1/\ell_2$-RIP and sharpness property, we then propose two complementary methods to recover the TT format tensor from the corrupted measurements: the projected subgradient method (PSubGM), which optimizes over the entire tensor, and the factorized Riemannian subgradient method (FRSubGM), which optimizes directly over the factors. Compared to PSubGM, the factorized approach FRSubGM significantly reduces the memory cost at the expense of a slightly slower convergence rate. Nevertheless, we show that both methods, with diminishing step sizes, converge linearly to the ground-truth tensor given an appropriate initialization, which can be obtained by a truncated spectral method. To the best of our knowledge, this is the first work to provide a theoretical analysis of the robust TT recovery problem and to demonstrate that TT-format tensors can be robustly recovered even when a certain fraction of measurements are arbitrarily corrupted. We conduct various numerical experiments to demonstrate the effectiveness of the two methods in robust TT recovery.**

***Index Terms*—Tensor-train decomposition, robust tensor recovery, $\ell_1/\ell_2$-RIP, sharpness, projected subgradient method, factorized Riemannian subgradient method, linear convergence.**

## I. Introduction

Tensor recovery has been widely investigated in many areas, such as signal processing and machine learning [1], [2], communication [3], quantum physics [4], [5], [6], chemometrics [7], [8], genetic engineering [9], and so on. One fundamental task is to recover a tensor $\mathcal{X}^\star \in \mathbb{R}^{d_1\times\cdots\times d_N}$ from highly incomplete, sometimes even corrupted, observations $\boldsymbol{y} = \{y_k\}_{k=1}^m$ given by

$$\boldsymbol{y} = \mathcal{A}(\mathcal{X}^\star) + \boldsymbol{s} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \langle \mathcal{A}_1, \mathcal{X}^\star\rangle + s_1 \\ \vdots \\ \langle \mathcal{A}_m, \mathcal{X}^\star\rangle + s_m \end{bmatrix} \in \mathbb{R}^m, \quad (1)$$

where $\mathcal{A}(\mathcal{X}^\star): \mathbb{R}^{d_1\times\cdots\times d_N} \to \mathbb{R}^m$ is a linear observation operator that models the measurement process and $\boldsymbol{s}\in\mathbb{R}^m$ represents an outlier vector, wherein only a small fraction of its entries (referred to as outliers) have arbitrary magnitudes but their locations are unknown a prior, while the remaining entries are zero. In practical scenarios, outliers are frequently encountered in sensing or regression models [4], [5], [10], [11], [12], [13], [14], [15], [16], stemming from various factors such as sensor malfunctions and malicious attacks. For instance, in quantum state tomography, imperfections during quantum state preparation can randomly generate unwarranted outlier quantum states, which subsequently lead to outliers during the measurement operation [14], [17], [18].

Even in the absence of outliers, the recovery from (1) remains ill-posed due to the curse of dimensionality, which arises from the exponential storage complexity of $\mathcal{X}^\star$ with respect to $N$. Therefore, it is often advantageous (and even necessary) to employ certain tensor decomposition models to compactly represent the full tensor. One commonly used model is the tensor train (TT) decomposition [19], which expresses the $(s_1,\ldots,s_N)$-th element of $\mathcal{X}^\star$ as the following matrix product form [19]

$$\mathcal{X}^\star(s_1,\ldots,s_N) = \boldsymbol{X}_1^\star(:,s_1,:)\boldsymbol{X}_2^\star(:,s_2,:)\cdots\boldsymbol{X}_N^\star(:,s_N,:), \quad (2)$$

where tensor factors $\boldsymbol{X}_i^\star\in\mathbb{R}^{r_{i-1}\times d_i\times r_i}, i\in[N]$ with $r_0 = r_N = 1$. The dimensions $\operatorname{rank}(\mathcal{X}^\star) = (r_1,\ldots,r_{N-1})$ of such a decomposition are called the TT ranks[1] of $\mathcal{X}^\star$. 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 $i$ so that the total number of parameters in the tensor factors $\{\boldsymbol{X}_i^\star\}$ is much smaller than the number of entries in $\mathcal{X}^\star$. We refer to any tensor for which such a low-rank TT decomposition exists as a *low-TT-rank* tensor. To simplify the notation, we may also use $[\boldsymbol{X}_1^\star,\ldots,\boldsymbol{X}_N^\star]$ as the compact form of $\mathcal{X}^\star$.

Compared to the other two commonly used tensor decompositions—canonical polyadic (CP) [20] and Tucker [21]

Received 21 October 2024; revised 8 March 2025; accepted 12 April 2025. Date of publication 16 April 2025; date of current version 13 June 2025. This work was supported by NSF under Grant CCF-2241298 and Grant ECCS-2409701. The associate editor coordinating the review of this article and approving it for publication was Xiao Fu. *(Corresponding author: Zhihui Zhu.)*

The authors are with the Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43201 USA (e-mail: qin.660@osu.edu; zhu.3440@osu.edu).

Digital Object Identifier 10.1109/TSP.2025.3561121

[1]Any tensor can be decomposed in the TT format (2) with sufficiently large TT ranks [19, Theorem 2.1]. Indeed, there always exists a TT decomposition with $r_i \le \min\{\Pi_{j=1}^i d_j, \Pi_{j=i+1}^N d_j\}$ for any $i\ge 1$.

decompositions—TT decomposition strikes a balance between the advantages of both approaches[2]. The number of parameters of TT decomposition is $O(N\overline{d}\overline{r}^2)$ with $\overline{d} = \max_i d_i$ and $\overline{r} = \max_i r_i$, not growing exponentially with the tensor order as the CP decomposition. Furthermore, similar to the Tucker decomposition, the TT decomposition can be approximately computed using an SVD-based algorithm, called the tensor train SVD (TT-SVD), with a guaranteed accuracy [19]. See [24] for a detailed description. Consequently, TT decomposition has been widely applied to tensor recovery across various fields, including quantum tomography [5], neuroimaging [13], facial model refinement [25], and the distinction of its attributes [26], longitudinal relational data analysis [27], and forecasting tasks [28].

### A. Our Goals and Main Results

In this paper, we study the robust recovery problem in (1), where the underlying tensor $\mathcal{X}^\star$ has low TT ranks. We refer to this as the robust TT recovery problem. To handle outliers in the measurements, we employ a robust $\ell_1$ loss function together with the TT format and solve the following problem:

$$\min_{\substack{\mathcal{X} \in^{d_1\times\cdots\times d_N} \\ \mathrm{rank}(\mathcal{X}) = (r_1,\dots,r_{N-1})}} f(\mathcal{X}) = \frac{1}{m}\|\mathcal{A}(\mathcal{X}) - \boldsymbol{y}\|_1. \tag{3}$$

Compared to the conventional least-squares ($\ell_2$) loss, the $\ell_1$ loss function[3] is more robust against outliers and has been widely adopted in robust signal recovery problems [14], [16], [34], [35], [36], [37], [38], [39]. However, the combination of the $\ell_1$ loss function and TT decomposition makes the problem (3) highly nonsmooth and nonconvex. Our goal is to study its optimality conditions and develop optimization algorithms with guaranteed performance.

Note that measurements should satisfy certain properties to enable robust recovery from corrupted measurements. *Our first contribution is to study the stable embedding of low-TT-rank tensors by establishing the following $\ell_1/\ell_2$-restricted isometry property ($\ell_1/\ell_2$-RIP[4]) without outliers for $\mathcal{A}$*, which has been introduced previously in the context of low-rank matrix/Tucker tensor recovery [16], [36], [41], [42], [43] and covariance estimation [44]. This mixed-norm approximate isometry evaluates the signal strength before and after projection using different metrics: the input is measured in terms of the Frobenius norm, and the output is measured in terms of the $\ell_1$ norm. Specifically, we say $\mathcal{A}$ satisfies rank-$\overline{r}$ $\ell_1/\ell_2$-RIP if there exits a constant $\delta_{\overline{r}} \in (0, \sqrt{2/\pi})$ such that

$$(\sqrt{2/\pi} - \delta_{\overline{r}})\|\mathcal{X}\|_F \le \frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1 \le (\sqrt{2/\pi} + \delta_{\overline{r}})\|\mathcal{X}\|_F \tag{4}$$

holds for all low-TT-rank tensors with ranks $(r_1, \dots, r_{N-1}), r_i \le \overline{r}$. We show that Gaussian measurement operators $\mathcal{A}$, where $\mathcal{A}_1, \dots, \mathcal{A}_m$ have independent and identically distributed (i.i.d.) standard Gaussian entries, satisfies $\ell_1/\ell_2$-RIP (4) with high probability as long as $m \ge \Omega(N\overline{d}\overline{r}^2 \log N/\delta_{\overline{r}}^2)$ with $\overline{d} = \max_i d_i$. This implies that robust TT recovery is possible using a number of measurements that only scale (approximately) linearly with regard to *N*. With the $\ell_1/\ell_2$-RIP property, we show that the robust loss function in (3) satisfies the sharpness property [16], [36], [43], [45], [46], [47]: for any low-TT-rank tensors $\mathcal{X}$ with TT ranks $r_i \le \overline{r}$, it holds that

$$\begin{aligned}\frac{1}{m}\|\mathcal{A}(\mathcal{X} - \mathcal{X}^\star) - \boldsymbol{s}\|_1 - \frac{1}{m}\|\boldsymbol{s}\|_1 \\ \ge ((1 - 2p_s)\sqrt{2/\pi} - \delta_{\overline{r}})\|\mathcal{X} - \mathcal{X}^\star\|_F,\end{aligned} \tag{5}$$

where $p_s \in [0, 0.5 - 0.5\delta_{2\overline{r}}/\sqrt{2/\pi}]$ represents the fraction of outliers in $\boldsymbol{y}$, i.e., $p_s = \|\boldsymbol{s}\|_0/m$. Since (3) optimizes only over low-TT-rank tensors, (5) needs to hold only for these tensors; as such, a similar condition for Tucker tensors is also referred to as *restricted sharpness* in [16], [43]. The sharpness condition (5) implies that $\mathcal{X}^\star$ is the unique global minimum, with the loss function increasing as the variable $\mathcal{X}$ deviates from $\mathcal{X}^\star$.

*Our second contribution is to propose two complementary iterative algorithms for solving* (3). Building on insights from [40], we first introduce a projected subgradient method (PSubGM). This method optimizes the entire tensor in each iteration and employs the TT-SVD to project the iterates back to the TT format. Under the sharpness property, we establish a robust regularity condition (RRC) for the objective function (3). We show that the PSubGM algorithm, with appropriate initialization and diminishing step sizes, achieves a linear convergence rate. Remarkably, PSubGM can precisely recover the ground-truth tensor $\mathcal{X}^\star$ even in the presence of outliers.

A potential drawback of PSubGM when handling high-order tensors is that it requires storing the full estimated tensor $\mathcal{X}$ and performing TT-SVD at each iteration, which becomes impractical for large *N*, such as in quantum state tomography involving hundreds of qubits [5]. To address this issue, instead of optimizing directly over the tensor $\mathcal{X}$, we employ the factorization approach that optimizes over the factors $\{\boldsymbol{X}_i\}_{i\ge 1}$ which can significantly reduce the memory cost. Specifically, we consider the following optimization problem:

$$\begin{aligned}&\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1}\times d_i \times r_i}, \\ i \in [N]}} \frac{1}{m}\|\mathcal{A}([\boldsymbol{X}_1, \dots, \boldsymbol{X}_N]) - \boldsymbol{y}\|_1, \\ &\text{s.t.} \sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}, \quad i \in [N-1].\end{aligned} \tag{6}$$

The additional constraints $\sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}$ are introduced to reduce the scaling ambiguity of the factors

[2] In general, finding the optimal CP decomposition for high-order tensors can be computationally difficulty [22], [23], while the Tucker decomposition becomes inapplicable for high-order tensors due to the number of parameters scaling exponentially with the tensor order.

[3] Various smoothed version of $\ell_1$, including Huber, Welsch, hybrid ordinary-$\ell_p$, hybrid ordinary-Welsch, and hybrid ordinary-Cauchy [29], [30], [31], have been proposed to suppress outliers. Additionally, the truncated-quadratic function [32], [33] has been employed to clip large values. However, these functions introduce additional hyperparameters that may require careful tuning in practice. In this work, we primarily focus on the $\ell_1$ norm and prove the convergence of subgradient methods. However, we believe that our analysis can also be extended to other functions, which we leave for future research.

[4] $\ell_1/\ell_2$ -RIP differs from the $\ell_2/\ell_2$-RIP [5], [40], which examines the relationship between $\|\mathcal{A}(\mathcal{X})\|_2^2$ and $\|\mathcal{X}\|_F^2$.

TABLE I

COMPARISON OF IHT/FRGD FOR SOLVING TT RECOVERY WITH NOISELESS MEASUREMENTS AND PSUBGM/FRSUBGM FOR THE CORRUPTED MEASUREMENTS. HERE $p_s$ DENOTES THE FRACTION OF OUTLIERS IN THE MEASUREMENTS. THE UPPER BOUND OF INITIALIZATION IS EXPRESSED IN TERMS OF $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F$. THE CONVERGENCE RATES OF IHT, PSUBGM AND FRGD, FRSUBGM ARE RESPECTIVELY ANALYZED CONCERNING $\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2$ AND $\text{DIST}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$ DEFINED IN (22). $\delta_{\overline{r}} \in (0,1)$ IS A CONSTANT IN STANDARD $\ell_2/\ell_2$-RIP (SEE [5, THEOREM 2]), WHILE $\delta_{\overline{r}} \in (0,1)$ IS A CONSTANT IN $\ell_1/\ell_2$-RIP. $c$ IS A UNIVERSAL CONSTANT

| Algorithm | Outlier | Initialization Requirement | Rate of Convergence | RIP Condition |
|---|---|---|---|---|
| IHT [40], [48] | × | $\frac{c\underline{\sigma}(\mathcal{X}^\star)}{600N}$ | $(1+c)\big(1-\frac{(1-\delta'_{4\overline{r}})^2}{2(1+\delta'_{4\overline{r}})^2}\big) \leq 1$ | $\delta'_{4\overline{r}} \leq \frac{1-\sqrt{2c/(1+c)}}{1+\sqrt{2c/(1+c)}}$ |
| FRGD [49] | × | $O\big(\frac{\underline{\sigma}^2(\mathcal{X}^\star)}{\overline{r}N^2\overline{\sigma}(\mathcal{X}^\star)}\big)$ | $1-O\big(\frac{(4-15\delta'_{(N+3)\overline{r}})^2}{(1+\delta'_{(N+3)\overline{r}})^2N^2\overline{r}\kappa^2(\mathcal{X}^\star)}\big)$ | $\delta'_{(N+3)\overline{r}} \leq \frac{4}{15}$ |
| PSubGM | √ | $\frac{c\underline{\sigma}(\mathcal{X}^\star)}{600N}$ | $(1+c)\big(1-\frac{3((1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}})^2}{4(\sqrt{2/\pi}+\delta_{2\overline{r}})^2}\big) \leq 1$ | $\delta_{2\overline{r}} \leq \frac{1-2p_s-\sqrt{4c/(3+3c)}}{1+\sqrt{4c/(3+3c)}}\sqrt{\frac{2}{\pi}}$ |
| FRSubGM | √ | $O\big(\frac{\underline{\sigma}^2(\mathcal{X}^\star)}{\overline{r}N^2\overline{\sigma}(\mathcal{X}^\star)}\big)$ | $1-O\big(\frac{((1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}})^2}{N^2\overline{r}(\sqrt{2/\pi}+\delta_{(N+1)\overline{r}})^2\kappa^2(\mathcal{X}^\star)}\big)$ | $\delta_{(N+1)\overline{r}} \leq (1-2p_s)\sqrt{\frac{2}{\pi}}$ |

[49]. The orthogonality constraints can be viewed as Stiefel manifolds of Riemannian space, so we utilize a factorized Riemannian subgradient method (FRSubGM) on the Stiefel manifold to optimize (6). We show that the objective function (6) also satisfies a Riemannian RRC, and prove that the FRSubGM algorithm, with an appropriate initialization and a diminishing step size, converges to the ground-truth tensor $\mathcal{X}^\star$ at a linear rate. Finally, we present a guaranteed truncated spectral initialization as a valid starting point, ensuring linear convergence for both the PSubGM and FRSubGM algorithms.

While the proposed PSubGM and FRSubGM are inspired by their smooth counterparts, iterative hard thresholding (IHT) [40] and factorized Riemannian gradient descent (FRGD) [49], which were developed to solve the smooth $\ell_2$ loss function, the analysis of subgradient methods is generally more challenging. For instance, subgradient methods with a constant step size may fail to converge to a critical point of a nonsmooth function, even when the function is convex [50], [51], [52]. To ensure convergence, a diminishing step size is required. Moreover, obtaining a good starting point is more difficult due to the presence of outliers. In Table I, we summarize the convergence results for PSubGM and FRSubGM and compare them with previous results on tensor recovery in the absence of outliers, specifically the IHT [40], [48] and factorized Riemannian gradient descent (FRGD) [49] that solve problems similar to (3) and the factorized problem (6), with the objective function being changed to a smooth $\ell_2$ loss function. We observe that PSubGM and FRSubGM achieve a similar linear convergence rate as their smooth counterparts, demonstrating that the outliers[5] can be handled as easily as in the noiseless case by the nonconvex optimization approaches. The convergence rate of IHT/PSubGM primarily hinges on the RIP constant, with a potential decay $(1+c)$ (where $c$ is a universal constant) owing to the expansiveness of the TT-SVD. Conversely, the convergence rate of FRGD/FRSubGM relies not only on the RIP constant but also on factors like $N$, $\overline{r}$, and $\mathcal{X}^\star$, which could impede the convergence speed. Our work extends the literature [35], [36], [37], [47], demonstrating that nonsmooth nonconvex optimization can be solved as efficiently as its smooth counterpart.

[5]In our paper, we focus on applications involving outliers, such as sparse noise, impulse noise, structured noise, and heavy-tailed noise. However, our methods remain applicable to scenarios with dense noise, where they achieve performance comparable to methods based on $\ell_2$ loss.

## B. Related Works

Theoretical analyses and algorithmic designs for robust low-rank matrix recovery via nonsmooth optimization have been extensively studied in [36], [39], [47], [53], [54]. A notable advantage of nonsmooth formulations is the enhanced robustness to adversarial outliers, achieved through a simple algorithmic design—the low-rank factors are updated in essentially the same manner, irrespective of the presence of outliers. However, existing theoretical frameworks for asymmetric matrix factorization cannot be extended to robust high-order tensor recovery, as the additional regularization terms introduced to balance the factors may not generalize to multiple tensor factors.

For tensor recovery from a limited number of measurements, most existing theoretical work and algorithmic designs have predominantly focused on developing optimization algorithms for either the noiseless case or the presence of Gaussian noise. Typically, a smooth loss function, such as the residual sum of squares ($\ell_2$ loss), is employed. Variants of projected gradient descent (PGD) algorithms, including IHT [40], [55], [56] and Riemannian gradient descent on the fixed-rank manifold [57], [58], have been studied for operating on the entire tensor with guaranteed convergence and performance. However, direct optimization over the tensor $\mathcal{X}$ poses a challenge due to its exponentially large memory requirements in terms of $N$. To address this storage issue, factorization approaches [49], [59], [60] have been developed to optimize the factors of a tensor decomposition.

In contrast, tensor recovery from measurements corrupted by outliers has been less studied. Recently, the work [16] introduced a scaled gradient method for exact recovery of order-3 Tucker from corrupted measurements. This method is proven to have a fast convergence rate that is independent of the condition number of the ground-truth tensor. However, the convergence analysis, as well as the $\ell_1/\ell_2$-RIP and sharpness properties in [16], are only established for order-3 Tucker tensors. It remains unclear whether the algorithm scales with respect to parameters such as the tensor order $N$, dimension $\overline{d}$, and Tucker rank. To the best of our knowledge, there is a lack of analysis and algorithmic design with guaranteed convergence for robust TT recovery.

### C. Notation

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 factors in the TT format, bold lowercase letters (e.g., $\boldsymbol{y}$) to denote vectors, and italic letters (e.g., $y$) to denote scalar quantities. Elements of matrices and tensors are denoted in parentheses, as in Matlab notation. For example, $\mathcal{X}(i_1, i_2, i_3)$ denotes the element in position $(i_1, i_2, i_3)$ of the order-3 tensor $\mathcal{X}$. The inner product of $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ and $\mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ can be denoted as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{s_1=1}^{d_1} \cdots \sum_{s_N=1}^{d_N} \mathcal{A}(s_1, \dots, s_N) \mathcal{B}(s_1, \dots, s_N)$. 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 $(s_1, \dots, s_N)$-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 $\mathcal{X}$. $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_F$ respectively represent the spectral norm and Frobenius norm of $\boldsymbol{X}$. $\sigma_i(\boldsymbol{X})$ is the $i$-th singular value of $\boldsymbol{X}$. For vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_2$ denotes its $\ell_2$ norm. For a positive integer $K$, $[K]$ denotes the set $\{1, \dots, K\}$. For two positive quantities $a, b \in \mathbb{R}$, $b = O(a)$ means $b \leq ca$ for some universal constant $c$; likewise, $b = \Omega(a)$ represents $b \geq ca$ for some universal constant $c$. To simplify notations in the following sections, for an order-$N$ TT format tensor with ranks $(r_1, \dots, r_{N-1})$, we define $\overline{r} = \max_{i=1}^{N-1} r_i$ and $\overline{d} = \max_{i=1}^{N} d_i$.

## II. $\ell_1/\ell_2$ -Restricted Isometry Property and Sharpness for Robust TT Recovery

### A. Tensor Train Decomposition

Recall the TT format in (2). Considering that $\boldsymbol{X}_i(:, s_i, :)$ will be extensively used, we denote it by $\boldsymbol{X}_i(s_i) \in \mathbb{R}^{r_{i-1} \times r_i}$ as one "slice" of $\boldsymbol{X}_i$ with the second index being fixed at $s_i$. Thus, for any $\mathcal{X} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_N] \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ in the TT format, we can express its $(s_1, \dots, s_N)$-th element as the following matrix product form

$$\mathcal{X}(s_1, \dots, s_N) = \prod_{i=1}^{N} \boldsymbol{X}_i(:, s_i, :) = \prod_{i=1}^{N} \boldsymbol{X}_i(s_i). \tag{7}$$

We may also arrange the slices $\{\boldsymbol{X}_i(s_i)\}_{s_i=1}^{d_i}$ into the following form:

$$L(\boldsymbol{X}_i) = \begin{bmatrix} \boldsymbol{X}_i(1) \\ \vdots \\ \boldsymbol{X}_i(d_i) \end{bmatrix} \in \mathbb{R}^{d_i r_{i-1} \times r_i}, \quad \forall i \in [N], \tag{8}$$

where $L(\boldsymbol{X}_i)$ is often referred to as the left unfolding of $\boldsymbol{X}_i$ when viewing $\boldsymbol{X}_i$ as a tensor.

The decomposition of the tensor $\mathcal{X}$ into the form of (7) is generally not unique: not only the factors $\boldsymbol{X}_i(s_i)$ are not unique, but also the dimension of these factors can vary. According to [61], there exists a unique set of ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$ for which $\mathcal{X}$ admits a minimal TT decomposition. We say the decomposition (7) is minimal if the rank of the left unfolding matrix $L(\boldsymbol{X}_i)$ in (8) is $r_i$. In addition, the factors can be chosen such that $L(\boldsymbol{X}_i)$ is orthonormal for all $i \in [N-1]$; that is

$$L^\top(\boldsymbol{X}_i) L(\boldsymbol{X}_i) = \mathbf{I}_{r_i}, \quad i \in [N-1]. \tag{9}$$

The resulting TT decomposition is called the left-orthogonal format of $\mathcal{X}$. Moreover, in this case, $r_i$ equals to the rank of the $i$-th unfolding matrix $\mathcal{X}^{\langle i \rangle} \in \mathbb{R}^{(d_1 \cdots d_k) \times (d_{k+1} \cdots d_N)}$ of the tensor $\mathcal{X}$, where the $(s_1 \cdots s_i, s_{i+1} \cdots s_N)$-th element[6] of $\mathcal{X}^{\langle i \rangle}$ is given by $\mathcal{X}^{\langle i \rangle}(s_1 \cdots s_i, s_{i+1} \cdots s_N) = \mathcal{X}(s_1, \dots, s_N)$. This can also serve as an alternative way to define the TT rank. With the $i$-th unfolding matrix $\mathcal{X}^{\langle i \rangle}$[7] and *TT ranks*, we can define 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 $\overline{\sigma}(\mathcal{X}) = \max_{i=1}^{N-1} \sigma_1(\mathcal{X}^{\langle i \rangle})$ and condition number $\kappa(\mathcal{X}) = \frac{\overline{\sigma}(\mathcal{X})}{\underline{\sigma}(\mathcal{X})}$.

### B. $\ell_1/\ell_2$-Restricted Isometry Property

We first prove the $\ell_1/\ell_2$-RIP property for the robust TT recovery problem with Gaussian measurement operators, a "gold standard" for studying random linear measurements in the compressive sensing literature [62], [63], [64], [65]. As previously studied in the contexts of low-rank matrix and Tucker tensor recovery problems [16], [36], [41], [42] and covariance estimation [44], $\ell_1/\ell_2$-RIP establishes a connection between $\|\mathcal{A}(\mathcal{X})\|_1$ and $\|\mathcal{X}\|_F$, differing from previous work on $\ell_2/\ell_2$-RIP [5], [40] on TT recovery problem, which examine the relationship between $\|\mathcal{A}(\mathcal{X})\|_2^2$ and $\|\mathcal{X}\|_F^2$.

*Theorem 1 ($\ell_1/\ell_2$-RIP of Gaussian measurement operators):* Suppose the linear map $\mathcal{A}: \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ is a Gaussian measurement operator where $\{\mathcal{A}_k\}_{k=1}^m$ have i.i.d. standard Gaussian entries. Let $\delta_{\overline{r}} \in (0, \sqrt{2/\pi})$ be a positive constant. If the number of measurements satisfies $m \geq \Omega(N \overline{d} \overline{r}^2 \log N / \delta_{\overline{r}}^2)$, then with probability exceeding $1 - e^{-\Omega(N \overline{d} \overline{r}^2 \log N)}$, $\mathcal{A}$ satisfies the $\ell_1/\ell_2$-restricted isometry property in the sense that

$$(\sqrt{2/\pi} - \delta_{\overline{r}}) \|\mathcal{X}\|_F \leq \frac{1}{m} \|\mathcal{A}(\mathcal{X})\|_1 \leq (\sqrt{2/\pi} + \delta_{\overline{r}}) \|\mathcal{X}\|_F \tag{10}$$

hold for all low-TT-rank tensors $\mathcal{X}$ with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$.

The proof is provided in Appendix B. Theorem 1 guarantees the RIP for Gaussian measurements where the number of measurements $m$ scales linearly, rather than exponentially, with respect to the tensor order $N$. When RIP holds, then for any two distinct TT format tensors $\mathcal{X}_1, \mathcal{X}_2$ with TT ranks smaller than $\overline{r}$, we have distinct measurements since

$$\begin{aligned} \frac{1}{m} \|\mathcal{A}(\mathcal{X}_1) - \mathcal{A}(\mathcal{X}_2)\|_1 &= \frac{1}{m} \|\mathcal{A}(\mathcal{X}_1 - \mathcal{X}_2)\|_1 \\ &\geq (\sqrt{2/\pi} - \delta_{2\overline{r}}) \|\mathcal{X}_1 - \mathcal{X}_2\|_F, \end{aligned} \tag{11}$$

[6] 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.

[7] 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 $\boldsymbol{X}^{\leq i}(s_1 \cdots s_i, :) = \boldsymbol{X}_1(s_1) \cdots \boldsymbol{X}_i(s_i)$ and $\boldsymbol{X}^{\geq i+1}(:, s_{i+1} \cdots s_N) = \boldsymbol{X}_{i+1}(s_{i+1}) \cdots \boldsymbol{X}_N(s_N)$. When factors are in left-orthogonal form, we have $\boldsymbol{X}^{\leq i^\top} \boldsymbol{X}^{\leq i} = \mathbf{I}_{r_i}$ and $\sigma_j(\mathcal{X}^{\langle i \rangle}) = \sigma_j(\boldsymbol{X}^{\geq i+1}), j \in [N-1]$.

which guarantees the possibility of exact recovery in the absence of outliers. In addition, we note that Theorem 1 can also be applicable to other measurement operators, such as subgaussian measurements [66], using a similar analysis.

### C. Sharpness

We now study the $\ell_1$ loss function $f(\mathcal{X}) = \frac{1}{m}\|\mathcal{A}(\mathcal{X}) - \boldsymbol{y}\|_1$ and establish the sharpness property [45], [46] that can ensure exact recovery with corrupted measurements in (1). Let $\mathcal{S} \subseteq \{1, \dots, m\}$ denote the support of the outlier vector $\boldsymbol{s}$, and $\mathcal{S}^c = \{1, \dots, m\} \backslash \mathcal{S}$. We define $p_s = \frac{|\mathcal{S}|}{m}$ as the fraction of outliers in $\boldsymbol{y}$. The following result establishes the sharpness property for $\mathcal{A}$.

*Lemma 1 (Sharpness of Gaussian measurement operators):* Given an unknown target tensor $\mathcal{X}^\star$ with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$, suppose the linear map $\mathcal{A} : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^m$ is a Gaussian measurement operator. Let $\delta_{2\overline{r}} \in (0, (1 - 2p_s)\sqrt{2/\pi})$ be a positive constant. If the number of measurements satisfies $m \geq \Omega(N\overline{d}\overline{r}^2 \log N/\delta_{2\overline{r}}^2)$, then with probability exceeding $1 - 2e^{-\Omega(N\overline{d}\overline{r}^2 \log N)}$, $\mathcal{A}$ satisfies the following sharpness property:

$$\frac{1}{m}\|\mathcal{A}(\mathcal{X} - \mathcal{X}^\star) - \boldsymbol{s}\|_1 - \frac{1}{m}\|\boldsymbol{s}\|_1 \geq ((1 - 2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}})\|\mathcal{X} - \mathcal{X}^\star\|_F \tag{12}$$

holds for all low-TT-rank tensors $\mathcal{X}$ with ranks $\boldsymbol{r}$.

The proof is given in Appendix C. To simplify the notation, we use $\delta_{2\overline{r}}$, as in Theorem 1, to represent the constant. Lemma 1 establishes an exact recovery condition for measurements with outliers (1), showing that when the outlier ratio $p_s \leq \frac{1}{2} - \frac{\delta_{2\overline{r}}}{2\sqrt{2/\pi}} \leq \frac{1}{2}$, the sharpness property (12) implies exact recovery as the left-hand side is equal to $f(\mathcal{X}) - f(\mathcal{X}^\star)$. Additionally, this property indicates that we can tolerate nearly $m/2$ outliers in the measurements when $\delta_{2\overline{r}}$ is sufficiently small. Denote by $\mathcal{A}_\mathcal{S}$ and $\mathcal{A}_{\mathcal{S}^c}$ as the linear operators in $\{\mathcal{A}_k : k \in \mathcal{S}\}$ and $\{\mathcal{A}_k : k \in \mathcal{S}^c\}$, respectively. One can also use the same analysis of [36, Proposition 2] to obtain a similar sharpness by directly using the $\ell_1/\ell_2$-RIP property: assuming that the measurement operators $\mathcal{A}$ and $\mathcal{A}_{\mathcal{S}^c}$ obey $\ell_1/\ell_2$-RIP as in Theorem 1, then we have i.e., $\frac{1}{m}\|\mathcal{A}(\mathcal{X} - \mathcal{X}^\star) - \boldsymbol{s}\|_1 - \frac{1}{m}\|\boldsymbol{s}\|_1 \geq (2(1 - p_s)(\sqrt{2/\pi} - \delta_{2\overline{r}}) - (\sqrt{2/\pi} + \delta_{2\overline{r}}))\|\mathcal{X} - \mathcal{X}^\star\|_F$ with $p_s \leq \frac{1}{2} - \frac{\delta_{2\overline{r}}}{\sqrt{2/\pi} - \delta_{2\overline{r}}}$. Compared to this result, our result provides a more relaxed condition for $\delta_{2\overline{r}}$ when $p_s$ is fixed, or for $p_s$ when $\delta_{2\overline{r}}$ is fixed.

## III. Provably Correct Algorithms for Robust TT Recovery

In this section, we develop gradient-based algorithms to recover $\mathcal{X}^\star$ from corrupted measurements $\boldsymbol{y} = \mathcal{A}(\mathcal{X}^\star) + \boldsymbol{s}$ as described in (1) by solving (3). Specifically, we introduce two iterative algorithms. The first algorithm, the projected subgradient method (PSubGM), optimizes the entire tensor in each iteration and employs the TT-SVD to project the iterates back to the TT format. To address the challenge of high-order tensors, which can be exponentially large, we then propose the factorized Riemannian subgradient method (FRSubGM). This method, based on the factorization approach, directly optimizes over the factors, reducing storage memory requirements at the expense of a slightly slower convergence rate compared to PSubGM. Finally, we show that the commonly used truncated spectral initialization provides a valid starting point for both PSubGM and FRSubGM.

### A. Projected Subgradient Method

We commence by reiterating the loss function in (3), which seeks to minimize the disparity between the measurements $\boldsymbol{y}$ and the linear map of the estimated low-TT-rank tensor $\mathcal{X}$ as:

$$\min_{\substack{\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_N} \\ \text{rank}(\mathcal{X}) = (r_1, \dots, r_{N-1})}} f(\mathcal{X}) = \frac{1}{m}\|\mathcal{A}(\mathcal{X}) - \boldsymbol{y}\|_1. \tag{13}$$

We solve (13) by a Projected SubGradient Method (PSubGM) with the following iterative updates:

$$\mathcal{X}^{(t+1)} = \text{SVD}_{\boldsymbol{r}}^{tt}(\mathcal{X}^{(t)} - \mu_t \partial f(\mathcal{X}^{(t)})), \tag{14}$$

where $\mu_t$ is the step size, $\partial f(\mathcal{X}^{(t)}) = \frac{1}{m}\sum_{k=1}^m \text{sign}(\langle \mathcal{A}_k, \mathcal{X}^{(t)} \rangle - y_k)\mathcal{A}_k$ is a subgradient[8] of $f$, and $\text{SVD}_{\boldsymbol{r}}^{tt}(\cdot)$ denotes the TT-SVD operation [19] that projects a given tensor to a TT format. Computing the optimal low-TT-rank approximation, in general, is NP-hard [67]. While the TT-SVD is not a nonexpansive projection, when two tensors are sufficiently close, it can have an improved guarantee that is independent of $N$, distinguishing it from the result in [19, Corollary 2.4].

*Lemma 2 ([68, Lemma 26]):* Let $\mathcal{X}^\star$ be in TT format with the ranks $(r_1, \dots, r_{N-1})$. For any $\mathcal{E} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ with $C_N\|\mathcal{E}\|_F \leq \underline{\sigma}(\mathcal{X}^\star)$ for some constant $C_N \geq 500N$, we have

$$\|\text{SVD}_{\boldsymbol{r}}^{tt}(\mathcal{X}^\star + \mathcal{E}) - \mathcal{X}^\star\|_F^2 \leq \|\mathcal{E}\|_F^2 + \frac{600N\|\mathcal{E}\|_F^3}{\underline{\sigma}(\mathcal{X}^\star)}. \tag{16}$$

Lemma 2 implies that when the initialization of PSubGM is close to $\mathcal{X}^\star$, the perturbation bound of the TT-SVD is independent of the order $N$ due to $\|\mathcal{E}\|_F \leq \frac{\underline{\sigma}(\mathcal{X}^\star)}{500N}$. To facilitate analyzing the local convergence of the PSubGM, we first establish the robust regularity condition which has been widely built in contexts such as low-rank matrix recovery [69], phase retrieval [70] and robust subspace learning [71]. The result is as follows:

*Lemma 3 (Robust regularity condition of f with respect to the full tensor):* Let the ground truth tensor $\mathcal{X}^\star$ be in TT format with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$. Assume the linear map $\mathcal{A}$ is a Gaussian measurement operator where $\{\mathcal{A}_k\}_{k=1}^m$ have i.i.d. standard Gaussian entries. Then, based on the $\ell_1/\ell_2$-RIP and

[8] The definition of (Fréchet) subdifferential [36] of $f$ at $\mathcal{X}$ is

$$\partial f(\mathcal{X}) = \left\{\mathcal{D} \in \mathbb{R}^{d_1 \times \cdots \times d_N} : \liminf_{\mathcal{X}' \to \mathcal{X}} \frac{f(\mathcal{X}') - f(\mathcal{X}) - \langle \mathcal{D}, \mathcal{X}' - \mathcal{X} \rangle}{\|\mathcal{X}' - \mathcal{X}\|_F} \geq 0\right\}, \tag{15}$$

where each $\mathcal{D} \in \partial f(\mathcal{X})$ is called a subgradient of $f$ at $\mathcal{X}$. In general, a nonsmooth function may have multiple subgraidents at certain points. Here, if there exist multiple subgradients, we pick the one with sign function defined as $\text{sign}(x) = \begin{cases} -1, & x \leq 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$, and with abuse of notation, we use $\partial f(\mathcal{X})$ to denote this subgradient.

sharpness property, $f$ satisfies the robust regularity condition in the sense that

$$\langle \mathcal{X} - \mathcal{X}^\star, \partial f(\mathcal{X}) \rangle \geq ((1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}}) \|\mathcal{X} - \mathcal{X}^\star\|_F, \quad (17)$$

for any low-TT-rank tensors $\mathcal{X}$ with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$.

The proof is given in Appendix D. This result essentially ensures that at any feasible point $\mathcal{X}$, the associated negative search direction $\partial f(\mathcal{X})$ maintains a positive correlation with the error $-(\mathcal{X} - \mathcal{X}^\star)$. This enables subgradient method with an appropriate step size will consistently move the current point closer to the global solution in each update.

In contrast to gradient descent, subgradient method with a constant step size may fail to converge to a critical point of a nonsmooth function, such as the $\ell_1$ loss, even if the function is convex [50], [51], [52]. Therefore, to ensure convergence of PSubGM, it is generally necessary to use a diminishing step size [52], [72]. Based on Lemma 3, we analyze the local convergence of the PSubGM with a diminishing step size.

*Theorem 2 (Local linear convergence of PSubGM):* Let $\mathcal{X}^\star$ be in TT format with ranks $\boldsymbol{r} = (r_1, \dots, r_{N-1})$. Assume that $\mathcal{A}$ obeys the $\ell_1/\ell_2$-RIP and sharpness with a constant $\delta_{2\overline{r}} \leq \frac{1-2p_s-\sqrt{4c/(3+3c)}}{1+\sqrt{4c/(3+3c)}}\sqrt{2/\pi}$ for a positive constant $c \leq \frac{3(1-2p_s)^2}{1+12p_s-12p_s^2}$. Suppose that the PSubGM in (14) is initialized with $\mathcal{X}^{(0)}$ satisfying

$$\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \leq \frac{c\underline{\sigma}(\mathcal{X}^\star)}{600N}, \quad (18)$$

and uses the step size $\mu_t = \lambda q^t$ in (14), where $q = \sqrt{(1+c)(1-\frac{3((1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}})^2}{4(\sqrt{2/\pi}+\delta_{2\overline{r}})^2})}$ and $\lambda = \|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \cdot \frac{(1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}}}{2(\sqrt{2/\pi}+\delta_{2\overline{r}})^2}$. Then, the iterates $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ generated by the PSubGM will converge linearly to $\mathcal{X}^\star$:

$$\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 \leq q^{2t}\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F^2. \quad (19)$$

This proof is provided in Appendix E. Note that the required initialization (18) and the term $1+c$ in (19) are introduced because of the sub-optimality of the TT-SVD operation. While we present the result using a single choice of $\lambda$ and $q$ for simplicity, a wider range of values can be slightly modified to the arguments, without compromising linear convergence. In practice, these parameters should be carefully tuned to ensure convergence. In order to ensure that the upper bound of recovery error in (19) is monotonically decreasing, we can choose $\delta_{2\overline{r}} \leq \frac{1-2p_s-\sqrt{4c/(3+3c)}}{1+\sqrt{4c/(3+3c)}}\sqrt{2/\pi} \leq (1-2p_s)\sqrt{2/\pi}$. This can be guaranteed by choosing sufficiently large $m$ according to Theorem 1. Additionally, it should be noted that the linear convergence rate of the PSubGM improves when either the outlier ratio $p_s$ decreases or the number of measurements $m$ increases.

## B. Factorized Riemannian Subgradient Method

One drawback of PSubGM is that it requires storing the entire tensor ($O(\overline{d}^N)$ size) in each iteration. To reduce the space complexity, an alternative approach is to directly estimate tensor factors $\{\boldsymbol{X}_i\}$, which have a complexity of $O(N\overline{d}\overline{r}^2)$, by solving

$$\min_{\substack{\boldsymbol{X}_i \in \mathbb{R}^{r_{i-1}\times d_i \times r_i},\\ i\in[N]}} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) = \frac{1}{m}\|\mathcal{A}([\boldsymbol{X}_1, \dots, \boldsymbol{X}_N]) - \boldsymbol{y}\|_1,$$
$$\text{s.t.} \sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}, \quad i \in [N-1].$$

The additional constraints $\sum_{s_i=1}^{d_i} \boldsymbol{X}_i^\top(:, s_i, :)\boldsymbol{X}_i(:, s_i, :) = \mathbf{I}_{r_i}$ are introduced to reduce the scaling ambiguity of the factors [49], i.e., recovering the left-orthogonal form. Noticing that constraints define a Stiefel manifold structure, we apply a Riemannian Subgradient method on the Stiefel manifold [73] to optimize it. We call the resulting algorithm FRSubGM, short for factorized Riemannian Subgradient Method, to emphasize the factorization approach. Specifically, recalling the left unfolding $L(\boldsymbol{X}_i)$ of factors in (8), FRSubGM involves iterative updates

$$L(\boldsymbol{X}_i^{(t+1)}) = \text{Retr}_{L(\boldsymbol{X}_i)}\Big(L(\boldsymbol{X}_i^{(t)}) - \frac{\mu_t}{\overline{\sigma}^2(\mathcal{X}^\star)} \mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}}\big(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})\big)\Big), i \in [N-1], \quad (20)$$

$$L(\boldsymbol{X}_N^{(t+1)}) = L(\boldsymbol{X}_N^{(t)}) - \mu_t \partial_{L(\boldsymbol{X}_N)}F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}), \quad (21)$$

where $\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}}(\boldsymbol{U}) = \boldsymbol{U} - \frac{1}{2}L(\boldsymbol{X}_i)(\boldsymbol{U}^\top L(\boldsymbol{X}_i) + (L(\boldsymbol{X}_i))^\top \boldsymbol{U})$ denotes the projection onto the tangent space of the Stifel manifold at the point $L(\boldsymbol{X}_i)$ and the polar decomposition-based retraction is $\text{Retr}_{L(\boldsymbol{X}_i)}(\boldsymbol{G}) = \boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-\frac{1}{2}}$. Moreover, $\mu_t$ is a diminishing step size. Note that we use discrepant step sizes between $\{L(\boldsymbol{X}_i)\}$ and $L(\boldsymbol{X}_N)$, i.e., $\mu_t/\overline{\sigma}^2(\mathcal{X}^\star)$ for $\{L(\boldsymbol{X}_i)\}$ and $\mu_t$ for $L(\boldsymbol{X}_N)$. This is because $\|L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|^2 = 1, i \in [N-1]$ and $\|R(\boldsymbol{X}_N^\star)\|^2 = \sigma_1^2(\mathcal{X}^{\star\langle N-1\rangle}) \leq \overline{\sigma}^2(\mathcal{X}^\star)$, where $R(\boldsymbol{X}_N^\star) = \begin{bmatrix}\boldsymbol{X}_N^\star(1) & \cdots & \boldsymbol{X}_N^\star(d_N)\end{bmatrix} \in \mathbb{R}^{r_{N-1}\times d_N}$, are satisfied in each iteration. For simplicity, we use $\overline{\sigma}^2(\mathcal{X}^\star)$ to unify the step size. However, in practical implementation, we have the flexibility to fine-tune the two step sizes.

Before analyzing the FRSubGM algorithm, we will establish an error metric to quantify the distinctions between factors in two left-orthogonal form tensors, namely $\mathcal{X} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_N]$ and $\mathcal{X}^\star = [\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_N^\star]$. Note that the left-orthogonal form still has rotation ambiguity among the factors in the sense that $\Pi_{i=1}^N \boldsymbol{X}_i^\star(s_i) = \Pi_{i=1}^N \boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i^\star(s_i)\boldsymbol{R}_i$ for any orthonormal matrix $\boldsymbol{R}_i \in \mathbb{O}^{r_i\times r_i}$ (with $\boldsymbol{R}_0 = \boldsymbol{R}_N = 1$). To capture this rotation ambiguity, by defining the rotated factors $L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)$ as $L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star) = \begin{bmatrix}\boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i^\star(1)\boldsymbol{R}_i \\ \vdots \\ \boldsymbol{R}_{i-1}^\top \boldsymbol{X}_i^\star(d_i)\boldsymbol{R}_i\end{bmatrix}$, $\forall \boldsymbol{R}_{i-1}$ and $\boldsymbol{R}_i \in \mathbb{O}^{r_i\times r_i}$, we then apply the distance between the two sets of factors as [49]

$$\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\}) = \min_{\substack{\boldsymbol{R}_i\in\mathbb{O}^{r_i\times r_i},\\ i\in[N-1]}} \sum_{i=1}^{N-1}\overline{\sigma}^2(\mathcal{X}^\star) \\ \|L(\boldsymbol{X}_i)-L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2+\|L(\boldsymbol{X}_N)-L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2, \quad (22)$$

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. Subsequently, we establish a connection between $\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\})$ and $\|\mathcal{X}-\mathcal{X}^\star\|_F^2$, implying the convergence behavior of $\|\mathcal{X}-\mathcal{X}^\star\|_F^2$ as $\{\boldsymbol{X}_i\}$ approaches global minima.

*Lemma 4 ([49, Lemma 1]):* For any two TT format tensors $\mathcal{X}$ and $\mathcal{X}^\star$ with ranks $\boldsymbol{r}=(r_1,\dots,r_{N-1})$ and $\overline{\sigma}^2(\mathcal{X})\le\frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$, let $\{\boldsymbol{X}_i\}$ and $\{\boldsymbol{X}_i^\star\}$ be the corresponding left-orthogonal form factors. Then $\|\mathcal{X}-\mathcal{X}^\star\|_F^2$ and $\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\})$ defined in (22) satisfy

$$\|\mathcal{X}-\mathcal{X}^\star\|_F^2 \ge \frac{\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\})}{8(N+1+\sum_{i=2}^{N-1}r_i)\kappa^2(\mathcal{X}^\star)}, \quad (23)$$

$$\|\mathcal{X}-\mathcal{X}^\star\|_F^2 \le \frac{9N}{4}\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\}). \quad (24)$$

Lemma 4 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. Next, we first provide the robust regularity condition of $F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N)$.

*Lemma 5 (Robust regularity condition of F with respect to tensor factors):* Let the ground truth tensor $\mathcal{X}^\star$ be in TT format with ranks $\boldsymbol{r}=(r_1,\dots,r_{N-1})$. Assume the linear map $\mathcal{A}$ obeys the $\ell_1/\ell_2$-RIP and sharpness with a constant $\delta_{(N+1)\overline{r}}\le(1-2p_s)\sqrt{2/\pi}$. Define the set $\mathcal{C}(b)$ as

$$\mathcal{C}(b) := \Big\{\mathcal{X} : \|\mathcal{X}-\mathcal{X}^\star\|_F^2\le b\Big\}, \quad (25)$$

where $b=\frac{\underline{\sigma}^4(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}})^2}{144(2N^2-2N+1)(N+1+\sum_{i=2}^{N-1}r_i)^2(\sqrt{2/\pi}+\delta_{(N+1)\overline{r}})^2\overline{\sigma}^2(\mathcal{X}^\star)}$. Then for any TT format $\mathcal{X}\in\mathcal{C}(b)$, $F$ satisfies the robust regularity condition:

$$\sum_{i=1}^N\Big\langle L(\boldsymbol{X}_i)-L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\Big(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N)\Big)\Big\rangle \\ \ge\frac{(1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}}}{4\sqrt{2(N+1+\sum_{i=2}^{N-1}r_i)}\kappa(\mathcal{X}^\star)}\mathrm{dist}^2(\{\boldsymbol{X}_i\},\{\boldsymbol{X}_i^\star\}). \quad (26)$$

To simplify the expression, we define the identity operator $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_N)}\mathrm{St}}=\mathcal{I}$ such that $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_N)}\mathrm{St}}(\nabla_{L(\boldsymbol{X}_N)}F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N))=\nabla_{L(\boldsymbol{X}_N)}F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N)$.

The proof is shown in Appendix F. This result guarantees a positive correlation between the errors $-\{L(\boldsymbol{X}_i)-L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\}_{i=1}^N$ and negative Riemannian search directions $\{\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N))\}_{i=1}^N$ for $N$ factors in the Riemannian space, i.e., negative Riemannian direction points

TABLE II
COMPLEXITY COMPARISON BETWEEN PSUBGM AND FRSUBGM

| Algorithm | Space Complexity | Computational Complexity |
|---|---|---|
| IHT | $O((m+1)\overline{d}^N)$ | $O(2m\overline{d}^N+\overline{d}^N\overline{r}^2)$ |
| FRSubGM | $O(m\overline{d}^N+N\overline{d}\overline{r}^2)$ | $O(2m\overline{d}^N+N\overline{d}^N\overline{r}^2+N\overline{d}\overline{r}^3)$ |

towards the true factors. When initialed properly, we then obtain a linear convergence of the FRSubGM with a diminishing step size as following:

*Theorem 3 (Local linear convergence of the FRSubGM):* Let the ground truth tensor $\mathcal{X}^\star$ be in TT format with ranks $\boldsymbol{r}=(r_1,\dots,r_{N-1})$. Assume that the linear map $\mathcal{A}$ obeys the $\ell_1/\ell_2$-RIP and sharpness with a constant $\delta_{(N+1)\overline{r}}\le(1-2p_s)\sqrt{2/\pi}$. Suppose that the FRSubGM in (20) and (21) is initialized with $\mathcal{X}^{(0)}$ satisfying $\mathcal{X}^{(0)}\in\mathcal{C}(b)$. In addition, we set the step size $\mu_t=\lambda q^t$ with $\lambda=\frac{(1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}}}{\sqrt{2(N+1+\sum_{i=2}^{N-1}r_i)}(9N-5)(\sqrt{2/\pi}+\delta_{(N+1)\overline{r}})^2\kappa(\mathcal{X}^\star)}\mathrm{dist}(\{\boldsymbol{X}_i^{(0)}\},\{\boldsymbol{X}_i^\star\})$ and $q=\sqrt{1-\frac{((1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}})^2}{8(N+1+\sum_{i=2}^{N-1}r_i)(9N-5)(\sqrt{2/\pi}+\delta_{(N+1)\overline{r}})^2\kappa^2(\mathcal{X}^\star)}}$. Then, the iterates $\{\boldsymbol{X}_i^{(t)}\}_{t\ge0}$ generated by the FRSubGM will converge linearly to $\{\boldsymbol{X}_i^\star\}$ (up to rotation):

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})\le\mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\},\{\boldsymbol{X}_i^\star\})q^{2t}. \quad (27)$$

The proof is provided in Appendix G. It is important to note that the convergence rate of FRSubGM is still linear, but the convergence rate of FRSubGM depends not only on the values of $\delta_{(N+1)\overline{r}}$ and $p_s$, but also on the ratio $\frac{1}{\kappa^2(\mathcal{X}^\star)}$ and the parameter $N$. Consequently, the convergence rate of FRSubGM could be slower than that of PSubGM. In addition, according to Lemma 4, we can also derive the linear convergence in terms of the entire tensor, i.e., $\|\mathcal{X}^{(t)}-\mathcal{X}^\star\|_F\le O(\frac{\underline{\sigma}^2(\mathcal{X}^\star)}{N^2\overline{r}})q^{2t}$. Ultimately, even though it may not be straightforward to choose exact deterministic values for $\lambda$ and $q$ in practice, we can still select values that are close to these desired values.

## C. Comparison Between PSubGM and FRSubGM

First, the space and computational complexities of PSubGM and FRSubGM are summarized in Table II. As shown, PSubGM has a higher space complexity than FRSubGM, whereas FRSubGM incurs greater computational complexity as a tradeoff for reduced space requirements. Therefore, the choice between these methods necessitates a balance between space and computational efficiency, depending on practical constraints. Additionally, we note that the terms $m\overline{d}^N$ and $2m\overline{d}^N$ in the space and computational complexities originate from the measurement operators $\{\mathcal{A}_k\}_{k=1}^m$. In most cases, such complexity is unavoidable. However, when the measurement operators exhibit local structure—such as those formed by tensor products in quantum state tomography [5]—the exponential complexity can be reduced to linear, specifically $O(mN\overline{r}^2)$.

Next, we discuss the scalability of PSubGM and FRSubGM. $(i)$ The initial requirement $\|\mathcal{X}^{(0)}-\mathcal{X}^\star\|_F$ for these methods is proportional to $O(\frac{\underline{\sigma}(\mathcal{X}^\star)}{N})$ and $O(\frac{\underline{\sigma}(\mathcal{X}^\star)}{N^2\overline{r}\kappa(\mathcal{X}^\star)})$, respectively. This indicates that the initialization requirement depends on $N$ polynomially rather than exponentially. $(ii)$ The convergence rates $q$ for the two methods are given by $(1+c)(1-\frac{3((1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}})^2}{4(\sqrt{2/\pi}+\delta_{2\overline{r}})^2})\le 1$ and $1-O(\frac{((1-2p_s)\sqrt{2/\pi}-\delta_{(N+1)\overline{r}})^2}{N^2\overline{r}(\sqrt{2/\pi}+\delta_{(N+1)\overline{r}})^2\kappa^2(\mathcal{X}^\star)})$. This implies that: (1) The convergence rate of PSubGM primarily depends on $p_s$ and $\delta_{2\overline{r}}$. (2) Due to the highly nonconvex property, the convergence rate of FRSubGM is also influenced by $N$, $\overline{r}$ and $\kappa(\mathcal{X}^\star)$. Nonetheless, both methods exhibit polynomial dependence on these parameters rather than exponential dependence, suggesting that their scalability is not significantly hindered by $\overline{d}$, $\overline{r}$ and $N$, as supported by numerical experiments. $(iii)$ Finally, we analyze the required number of measurements for both methods. Under Gaussian measurements, the conditions $m\ge\Omega(N\overline{d}\overline{r}^2\log N/\delta_{2\overline{r}}^2)$ and $m\ge\Omega(N^3\overline{d}\overline{r}^2\log N/\delta_{(N+1)\overline{r}}^2)$ should be satisfied, further confirming a polynomial rather than exponential dependence on $N$. In summary, PSubGM and FRSubGM maintain polynomial scalability with respect to key parameters, ensuring their feasibility for high-dimensional tensor recovery problems.

### *D. Truncated Spectral Initialization*

The above local linear convergence for both PSubGM and FRSubGM requires an appropriate initialization. To achieve such an initialization, the spectral initialization method is commonly employed in the literature [49], [60], [70], [74], [75]. In the presence of outliers, we employ the truncated spectral initialization method [16], [32], [76]:

$$\mathcal{X}^{(0)}=\text{SVD}_{\boldsymbol{r}}^{tt}\left(\frac{1}{(1-p_s)m}\sum_{k=1}^m y_k\mathcal{A}_k\mathbb{I}_{\{|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\}}\right),\tag{28}$$

where $|\boldsymbol{y}|_{(k)}$ denotes the $k$-th largest amplitude of $\boldsymbol{y}$ and $\mathbb{I}_{\{|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\}}$ indicates that this term is 1 if $|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)}$ and 0 otherwise. Recall that $\text{SVD}_{\boldsymbol{r}}^{tt}(\cdot)$ is the TT-SVD algorithm for finding a TT approximation.

The following result ensures that such an initialization $\mathcal{X}^{(0)}$ provides a good approximation of $\mathcal{X}^\star$.

*Theorem 4:* Let the ground truth tensor $\mathcal{X}^\star$ be in TT format with ranks $\boldsymbol{r}=(r_1,\dots,r_{N-1})$. Suppose the linear map $\mathcal{A}:\mathbb{R}^{d_1\times\cdots\times d_N}\to\mathbb{R}^m$ is a Gaussian measurement operator where $\{\mathcal{A}_k\}_{k=1}^m$ have i.i.d. standard Gaussian entries. Then with probability at least $1-e^{-\Omega(N\overline{d}\overline{r}^2\log N)}-e^{-\Omega(\log((1-p_s)m))}$, the spectral initialization $\mathcal{X}^{(0)}$ generated by (28) satisfies

$$\|\mathcal{X}^{(0)}-\mathcal{X}^\star\|_F\le O\left(\frac{N\overline{r}\log((1-p_s)m)\|\mathcal{X}^\star\|_F\sqrt{\overline{d}}\log N}{\sqrt{(1-p_s)m}}\right).$$

The proof is provided in Appendix H. In summary, Theorem 4 indicates that a sufficiently large $m$ allows for the identification of a suitable initialization that is appropriately close to the ground truth. Additionally, although $p_s$ is specified in (28), it is generally unknown in practical scenarios. Therefore, any constant $\alpha\in[0,1]$ can substitute for $p_s$ in $\lceil p_s m\rceil$. It should be noted that while a smaller $\alpha$ may eliminate more measurements containing outliers, this might necessitate a larger number of measurements $m$ to achieve a satisfactory initialization.

## IV. Numerical Experiments

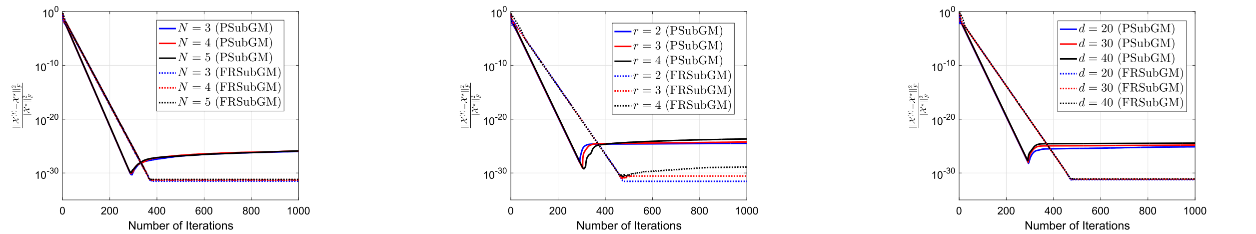
In this section, we conduct numerical experiments to evaluate the performance of PSubGM and FRSubGM algorithms in robust TT recovery. We generate an order-$N$ ground truth tensor $\mathcal{X}^\star\in\mathbb{R}^{d_1\times\cdots\times d_N}$ with ranks $\boldsymbol{r}=(r_1,\dots,r_{N-1})$ by truncating a random Gaussian tensor using a sequential SVD, followed by normalizing it to unit Frobenius norm. To simplify the selection of parameters, we let $d=d_1=\cdots=d_N$ and $r=r_1=\cdots=r_{N-1}$. We then obtain measurements $\{y_k\}_{k=1}^m$ in (1) from measurement operator $\mathcal{A}_k$ which is a random tensor with independent entries generated from the normal distribution, ensuring that $\langle\mathcal{A}_k,\mathcal{X}^\star\rangle\sim\mathcal{N}(0,1)$, and outlier $s_k\sim\mathcal{N}(0,10), k\in\mathcal{S}$ where $|\mathcal{S}|=p_s m$. The elements in $\mathcal{S}$ are randomly selected from the set $\{1,\dots,m\}$. We conduct 20 Monte Carlo trials and take the average over the 20 trials.

In the first experiment, we evaluate the performance of PSubGM and FRSubGM in terms of $N$, $d$, and $r$. Fig. 1(a)–1(c) clearly demonstrate that PSubGM exhibits a faster convergence speed compared to FRSubGM. However, the final recovery error of PSubGM is slightly higher than that of FRSubGM, which can possibly be attributed to the sub-optimality of the TT-SVD. Notably, unlike the significantly slower convergence rates of IHT [40] and FRGD [49] as $N$, $r$, and $d$ increase observed from [48, Fig. 1], PSubGM and FRSubGM do not exhibit such degradation, attributable to the use of a diminishing step size.

In the second experiment, we test the performance of PSubGM and FRSubGM in terms of $\lambda$ and $q$. In Fig. 2(a) and 2(b), we can observe that both larger and smaller values of $\lambda$ or $q$ can potentially result in slower convergence or worse recovery error. Therefore, it is crucial to carefully fine-tune the parameters $\lambda$ and $q$ to ensure optimal performance.

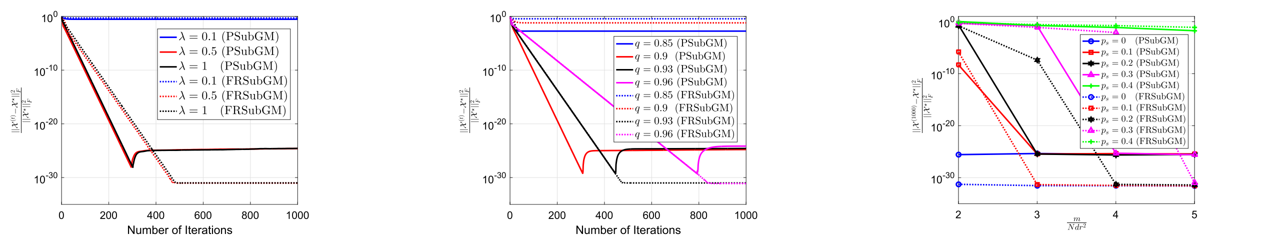
In the third experiment, we test the performance of PSubGM and FRSubGM in terms of $p_s$ and $m$. Fig. 2(c) illustrates the relationship between recovery error and $p_s$ and $m$. It is evident that a larger value of $\frac{m}{Ndr^2}$ ensures better performance, as the $\ell_1/\ell_2$-RIP constant $\delta$ is inversely proportional to $m$. Moreover, as $p_s$ increases, a larger number of measurement operators $m$ is required.

In the fourth experiment, we investigate the necessary value of $m$ for different $N$ utilizing PSubGM and FRSubGM. This investigation is illustrated in Fig. 3(a) and 3(b), where we evaluate the success rate of achieving $\frac{\|\mathcal{X}^{(1000)}-\mathcal{X}^\star\|_F^2}{\|\mathcal{X}^\star\|_F^2}\le 10^{-5}$ over 100 independent trials. Our findings demonstrate a trend: as the value of $m$ increases and $N$ decreases, the success rate of recovery also improves. In addition, we establish a linear correlation between the number of measurements, $m$, and the tensor oders $N$, aligning with the conditions stipulated in Theorem 1. It is important to recognize that FRSubGM necessitates a larger value of $m$ compared to PSubGM. This difference arises from our analysis, where FRSubGM must adhere to the $(N+1)\overline{r}$-$\ell_1/\ell_2$-RIP, whereas PSubGM only requires the $3\overline{r}$-$\ell_1/\ell_2$ -RIP.



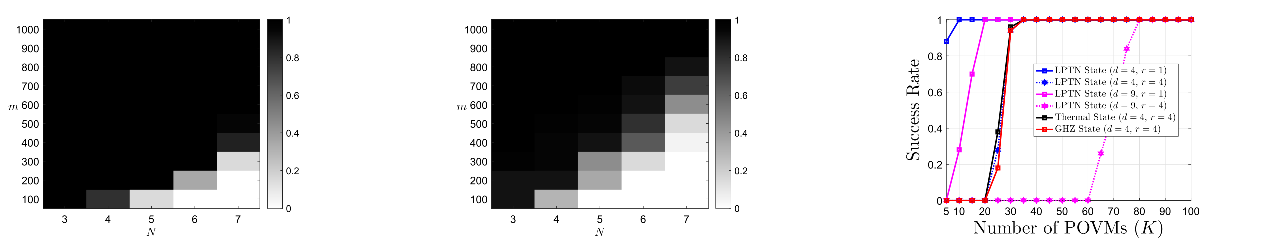

(a) (b) (c)

Fig. 1. Performance comparison of the PSubGM and FRSubGM in the robust tensor recovery, (a) for different $N$ with $d=6$, $r=2$, $m=1500$, $p_s=0.3$, $\lambda=0.5$, $q=0.9$ (PSubGM) and $q=0.91$ (FRSubGM) (b) for different $r$ with $d=40$, $N=3$, $m=12000$, $p_s=0.3$, $\lambda=0.5$, $q=0.9$ (PSubGM) and $q=0.93$ (FRSubGM) (c) for different $d$ with $N=3$, $r=2$, $m=3000$, $p_s=0.3$, $\lambda=0.5$, $q=0.9$ (PSubGM) and $q=0.93$ (FRSubGM).




(a) (b) (c)

Fig. 2. Performance comparison of the PSubGM and FRSubGM in the robust tensor recovery, (a) for different $\lambda$ with $N=3$, $d=30$, $r=2$, $m=2000$, $p_s=0.3$, $q=0.9$ (PSubGM) and $q=0.93$ (FRSubGM), (b) for different $q$ with $N=3$, $d=30$, $r=2$, $m=2000$, $p_s=0.3$ and $\lambda=0.5$, (c) for different $p_s$ and $m$ with $N=3$, $d=10$, $r=2$, $\lambda=0.5$, $q=0.9$ (PSubGM) and $q=0.91$ (FRSubGM).




(a) (b) (c)

Fig. 3. Performance comparison of (a) PSubGM and (b) FRSubGM in the robust tensor recovery, for different $N$ and $m$ with $d=4$, $r=2$, $p_s=0.05$, $\lambda=0.07$ and $q=0.99$, (c) PSubGM in the quantum state tomography using $K$ POVMs with $N=3$, $p_s=0.1$, $\lambda=1$, $q=0.97$ for states with $d=4$, as well as $\lambda=2$, $q=0.965$ or $q=0.97$ for LPTN states with $d=9$ and $r=1$ or $r=4$.

In the final experiment, we evaluate the proposed method for quantum state tomography in the presence of outliers. As demonstrated in previous results, PSubGM requires fewer measurements than FRSubGM. Therefore, we primarily employ PSubGM to determine the necessary number of measurements to ensure that $\frac{\|\mathcal{X}^{(1000)}-\mathcal{X}^\star\|_F^2}{\|\mathcal{X}^\star\|_F^2} \leq 10^{-10}$ over 100 independent trials. In this setting, we do not introduce measurement errors; instead, outliers are incorporated as $s_k \sim \mathcal{N}(0,10), k \in \mathcal{S}$. To conduct the experiments, we generate $m = d^{\frac{3}{2}}K$ measurements using $K$ positive operator-valued measures (POVMs), which are chosen as Haar-random projective measurements. The target states include the random locally purified tensor network (LPTN) state [4], the thermal state with temperature $T=0.2$ [6], and the Greenberger-Horne-Zeilinger (GHZ) state [6]. The detailed model can be found in [5]. From Fig. 3(c), we observe that PSubGM successfully recovers the ground truth tensor even when the magnitude of outliers is significantly larger than that of the pure measurements. Moreover, as $d$ and $r$ increase, a greater number of POVMs are required to ensure a higher success rate.

## V. Conclusion

In this paper, we develop efficient algorithms with guaranteed performance for robust tensor train (TT) recovery in the presence of outliers. We first prove the $\ell_1/\ell_2$-RIP of the Gaussian measurement operator and the sharpness property of the robust $\ell_1$ loss formulation, implying the possibility of exact recovery even when the measurements are corrupted by outliers. We then propose two iterative algorithms, namely the projected subgradient method (PSubGM) and the factorized Riemannian subgradient method (FRSubGM), to solve the corresponding recovery problems. With suitable initialization and diminishing step sizes, we show that both PSubGM and FRSubGM converge to the ground truth tensor at a linear rate and can tolerate $(0.5-0.5\delta/\sqrt{2/\pi}) \times 100\%$ of outliers, where $\delta \in (0, \sqrt{2/\pi})$ represents the $\ell_1/\ell_2$-RIP constant. We also demonstrate that a truncated spectral method can provide an appropriate initialization to ensure the local convergence of both algorithms.

As mentioned earlier, the convergence rate of the FRSubGM is influenced by the condition number $\kappa(\mathcal{X}^\star)$, with larger values of $\kappa(\mathcal{X}^\star)$ leading to slower convergence. A potential direction for future research is to incorporate the scaled technique [47], [59], [77] into our algorithms to mitigate the impact of $\kappa(\mathcal{X}^\star)$ on convergence. Another promising avenue for future exploration is the investigation of robust overparameterized TT recovery, building upon the advancements made in the matrix case [78], [79], [80].

## Appendix A
## Technical Tools Used in Proofs

We present some useful results for the proofs in the next sections.

*Lemma 6 ([49, Lemma 3]):* For any $\boldsymbol{A}_i, \boldsymbol{A}_i^\star \in \mathbb{R}^{r_{i-1}\times r_i}, i\in[N]$, we have $\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star = \sum_{i=1}^N \boldsymbol{A}_1^\star\cdots\boldsymbol{A}_{i-1}^\star(\boldsymbol{A}_i-\boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}\cdots\boldsymbol{A}_N$.

*Lemma 7:* Consider the loss function $f(\mathcal{X}) = \frac{1}{m}\|\mathcal{A}(\mathcal{X}) - \boldsymbol{y}\|_1$, where the measurement operator $\mathcal{A}$ satisfies the $\ell_1/\ell_2$-RIP with a constant $\delta_{2\overline{r}}$. Then for any $\mathcal{X}$, it holds that $\|\mathcal{D}\|_F \le \sqrt{2/\pi}+\delta_{2\overline{r}}, \forall \mathcal{D}\in\partial f(\mathcal{X})$.

*Proof:* Recall the definition of (Fréchet) subdifferential of $f$ at $\mathcal{X}$ in (15). Now for any $\mathcal{X}'\in\mathbb{R}^{d_1\times\cdots\times d_N}$, we have $|f(\mathcal{X}')-f(\mathcal{X})| = \frac{1}{m}\big|\|\mathcal{A}(\mathcal{X}')-\boldsymbol{y}\|_1 - \|\mathcal{A}(\mathcal{X})-\boldsymbol{y}\|_1\big| \le \frac{1}{m}\|\mathcal{A}(\mathcal{X}'-\mathcal{X})\|_1 \le (\sqrt{2/\pi}+\delta_{2\overline{r}})\|\mathcal{X}'-\mathcal{X}\|_F$, where the second inequality follows from the $\ell_1/\ell_2$-RIP of $\mathcal{A}$. This further implies that $\liminf_{\mathcal{X}'\to\mathcal{X}}\frac{|f(\mathcal{X}')-f(\mathcal{X})|}{\|\mathcal{X}'-\mathcal{X}\|_F} \le \lim_{\mathcal{X}'\to\mathcal{X}}\frac{(\sqrt{2/\pi}+\delta_{2\overline{r}})\|\mathcal{X}'-\mathcal{X}\|_F}{\|\mathcal{X}'-\mathcal{X}\|_F} = \sqrt{2/\pi}+\delta_{2\overline{r}}$.

Upon taking $\mathcal{X}' = \mathcal{X}+t\mathcal{D}, t\to 0$ and invoking (15), we have $\|\mathcal{D}\|_F\le\sqrt{2/\pi}+\delta_{2\overline{r}}, \quad \forall\mathcal{D}\in\partial f(\mathcal{X})$. □

*Lemma 8 ([81, Lemma 1]):* Let $\boldsymbol{X}^\top\boldsymbol{X} = \mathbf{I}$ and $\boldsymbol{\xi}$ on the tangent space of Stiefel manifold be given. Consider the point $\boldsymbol{X}^+ = \boldsymbol{X}+\boldsymbol{\xi}$. Then, the polar decomposition-based retraction satisfies $\text{Retr}_{\boldsymbol{X}}(\boldsymbol{X}^+) = \boldsymbol{X}^+({\boldsymbol{X}^+}^\top\boldsymbol{X}^+)^{-\frac{1}{2}}$ and

$$\|\text{Retr}_{\boldsymbol{X}}(\boldsymbol{X}^+) - \overline{\boldsymbol{X}}\|_F \le \|\boldsymbol{X}^+ - \overline{\boldsymbol{X}}\|_F = \|\boldsymbol{X}+\boldsymbol{\xi}-\overline{\boldsymbol{X}}\|_F \tag{29}$$

for any $\overline{\boldsymbol{X}}^\top\overline{\boldsymbol{X}} = \mathbf{I}$.

Finally, we introduce a new operation related to the multiplication of submatrices within the left unfolding matrices $L(\boldsymbol{X}_i) = \big[\boldsymbol{X}_i^\top(1)\cdots\boldsymbol{X}_i^\top(d_i)\big]^\top \in\mathbb{R}^{(r_{i-1}d_i)\times r_i}, i\in[N]$. For simplicity, we will only consider the case $d_i = 2$, but extending to the general case is straightforward. In particular, 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 $\boldsymbol{A}\overline{\otimes}\boldsymbol{B} = \begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}\overline{\otimes}\begin{bmatrix}\boldsymbol{B}_1\\ \boldsymbol{B}_2\end{bmatrix} = \big[(\boldsymbol{A}_1\boldsymbol{B}_1)^\top\ (\boldsymbol{A}_2\boldsymbol{B}_1)^\top\ (\boldsymbol{A}_1\boldsymbol{B}_2)^\top\ (\boldsymbol{A}_2\boldsymbol{B}_2)^\top\big]^\top$.

According to [49, Lemma 2], we can conclude that for any left-orthogonal TT format tensor $\mathcal{X}^\star = [\boldsymbol{X}_1^\star,\dots,\boldsymbol{X}_N^\star]$, we have

$$\|\mathcal{X}^\star\|_F = \|L(\boldsymbol{X}_1^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_N^\star)\|_2 = \|L(\boldsymbol{X}_N^\star)\|_2, \tag{30}$$

$$\begin{aligned}&\|L(\boldsymbol{X}_i^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_{N-1}^\star)\overline{\otimes}L(\boldsymbol{X}_N^\star)\|_2\\ &\quad\le\Pi_{l=i}^{N-1}\|L(\boldsymbol{X}_l^\star)\|\|L(\boldsymbol{X}_N^\star)\|_2,\quad\forall i\in[N-1],\end{aligned} \tag{31}$$

$$\begin{aligned}&\|L(\boldsymbol{X}_i^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_j^\star)\|\\ &\quad\le\Pi_{l=i}^{j}\|L(\boldsymbol{X}_l^\star)\| = 1,\quad i\le j,\quad\forall i,j\in[N-1],\end{aligned} \tag{32}$$

$$\begin{aligned}&\|L(\boldsymbol{X}_i^\star)\overline{\otimes}\cdots\overline{\otimes}L(\boldsymbol{X}_j^\star)\|_F\\ &\quad\le\Pi_{l=i}^{j-1}\|L(\boldsymbol{X}_l^\star)\|\|L(\boldsymbol{X}_j^\star)\|_F,\quad i\le j,\quad\forall i,j\in[N-1].\end{aligned} \tag{33}$$

## Appendix B
## Proof of Theorem 1

*Proof:* We first compute the covering number for any low-TT-rank tensor $\mathcal{X} = [\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]\in\mathbb{R}^{d_1\times\cdots\times d_N}$ with ranks $(r_1,\dots,r_{N-1})$. Given that any TT format can be converted to its left-orthogonal form, we denote $\mathcal{X} = [\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]$ as the left-orthogonal format. According to [82], we can construct $\epsilon$-net $\{L(\boldsymbol{X}_i^{(1)}),\dots,L(\boldsymbol{X}_i^{(n_i)})\}, i\in[N-1]$ for each set of matrices $\{L(\boldsymbol{X}_i)\in\mathbb{R}^{d_ir_{i-1}\times r_i}, i\in[N-1]: \|L(\boldsymbol{X}_i)\|\le 1\}$ $(r_0 = 1)$ such that $\sup_{L(\boldsymbol{X}_i):\|L(\boldsymbol{X}_i)\|\le 1}\min_{p_i\le n_i}\|L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^{(p_i)})\|\le\epsilon$ with the covering number $n_i\le(\frac{4+\epsilon}{\epsilon})^{d_ir_{i-1}r_i}$. Also, we can construct $\epsilon$-net $\{L(\boldsymbol{X}_N^{(1)}),\dots,L(\boldsymbol{X}_N^{(n_N)})\}$ for $\{L(\boldsymbol{X}_N)\in\mathbb{R}^{d_Nr_{N-1}\times 1}: \|L(\boldsymbol{X}_N)\|_2\le 1\}$ such that $\sup_{L(\boldsymbol{X}_N):\|L(\boldsymbol{X}_N)\|_2\le 1}\min_{p_N\le n_N}\|L(\boldsymbol{X}_N) - L(\boldsymbol{X}_N^{(p_N)})\|_2\le\epsilon$ with the covering number $n_N\le(\frac{2+\epsilon}{\epsilon})^{d_Nr_{N-1}r_N}$. Hence, for any low-rank TT format $\mathcal{X}$ with $\|\mathcal{X}\|_F\le\|\boldsymbol{X}_N\|_F\le 1$ derived from (31), its covering argument is $\Pi_{i=1}^N n_i\le(\frac{4+\epsilon}{\epsilon})^{d_1r_1+\sum_{i=2}^{N-1}d_ir_{i-1}r_i+d_Nr_{N-1}}\le(\frac{4+\epsilon}{\epsilon})^{N\overline{d}\overline{r}^2}$ where $\overline{r} = \max_i r_i$ and $\overline{d} = \max_i d_i$.

Without loss of the generality, we assume that $\mathcal{X}$ is in TT format with $\|\mathcal{X}\|_F = 1$. For simplicity, we use $\mathcal{I}$ to denote the index set $[n_1]\times\cdots\times[n_N]$. According to the construction of the $\epsilon$-net, there exists $p = (p_1,\dots,p_N)\in\mathcal{I}$ such that $\|L(\boldsymbol{X}_i) - L(\boldsymbol{X}_i^{(p_i)})\|\le\epsilon,\quad i\in[N-1]$ and $\|L(\boldsymbol{X}_N) - L(\boldsymbol{X}_N^{(p_N)})\|_2\le\epsilon$.

Taking $\epsilon = \frac{c\delta_{\overline{r}}}{N}$ with a positive constant $c$ gives

$$\begin{aligned}&\sup_{\mathcal{X}}\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^{(p)})\|_1\\ &\quad=\sup_{[\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]}1/m\sum_{k=1}^m|\langle\mathcal{A}_k,[\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]\\ &\qquad-[\boldsymbol{X}_1^{(p_1)},\dots,\boldsymbol{X}_N^{(p_N)}]\rangle|\\ &\quad=\sup_{[\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]}1/m\sum_{k=1}^m|\langle\mathcal{A}_k,\sum_{a_1=1}^N[\boldsymbol{X}_1^{(p_1)},\dots,\boldsymbol{X}_{a_1}^{(p_{a_1})}\\ &\qquad-\boldsymbol{X}_{a_1},\dots,\boldsymbol{X}_N]\rangle|\\ &\quad\le\sup_{\mathcal{X}}\frac{N\epsilon}{m}\|\mathcal{A}(\mathcal{X})\|_1 = \sup_{\mathcal{X}}\frac{c\delta_{\overline{r}}}{m}\|\mathcal{A}(\mathcal{X})\|_1,\end{aligned} \tag{34}$$

where we write $[\boldsymbol{X}_1,\dots,\boldsymbol{X}_N] - [\boldsymbol{X}_1^{(p_1)},\dots,\boldsymbol{X}_N^{(p_N)}]$ in the second line as the sum of $N$ terms via Lemma 6.

According to (34), we have

$$\begin{aligned}
&\sup_{\mathcal{X}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right| \\
&\leq \sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(p)})\|_1-\sqrt{2/\pi}\right|+\sup_{\mathcal{X}}\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^{(p)})\|_1 \\
&\leq \sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(p)})\|_1-\sqrt{2/\pi}\right|+\sup_{\mathcal{X}}\frac{c\delta_{\overline{r}}}{m}\|\mathcal{A}(\mathcal{X})\|_1 \\
&\leq \sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(p)})\|_1-\sqrt{2/\pi}\right| \\
&\quad+\sup_{\mathcal{X}} c\delta_{\overline{r}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right|+c\delta_{\overline{r}}\sqrt{2/\pi}, \qquad (35)
\end{aligned}$$

and then it follows

$$\begin{aligned}
&\sup_{\mathcal{X}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right| \\
&\quad\leq \frac{\sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(p)})\|_1-\sqrt{2/\pi}\right|+c\delta_{\overline{r}}\sqrt{2/\pi}}{1-c\delta_{\overline{r}}}. \qquad (36)
\end{aligned}$$

To finish our derivation, we need to obtain the concentration inequality with respect to $\frac{1}{m}\sum_{k=1}^m|\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle|$. For any fixed TT format tensor $\mathcal{X}^{(p)}$ with $\|\mathcal{X}^{(p)}\|_F=1$, $|\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle|$ obeys standard Gaussian with mean $\sqrt{2/\pi}$ and unit variance since $\{\mathcal{A}_k\}_{k=1}^m$ have i.i.d. standard Gaussian entries. Hence, based on the tail function of Gaussian random variable, we have $\mathbb{P}\left(\sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\sum_{k=1}^m|\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle|-\sqrt{2/\pi}\right|\geq\frac{\delta_{\overline{r}}}{2}\right)\leq 2e^{-c_1 m\delta_{\overline{r}}^2}$, where $c_1$ is a constant.

Based on (36), we can derive

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathcal{X}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right|\geq\frac{\frac{\delta_{\overline{r}}}{2}+c\delta_{\overline{r}}\sqrt{2/\pi}}{1-c\delta_{\overline{r}}}\right) \\
&\leq \mathbb{P}\Bigg((\sup_{\mathcal{X}^{(p)}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X}^{(p)})\|_1-\sqrt{2/\pi}\right|+c\delta_{\overline{r}}\sqrt{2/\pi})/(1-c\delta_{\overline{r}}) \\
&\quad\geq\frac{\frac{\delta_{\overline{r}}}{2}+c\delta_{\overline{r}}\sqrt{2/\pi}}{1-c\delta_{\overline{r}}}\Bigg)\leq\Pi_{i=1}^N n_i e^{1-c_1 m\delta_{\overline{r}}^2} \\
&\leq\left(\frac{4+\epsilon}{\epsilon}\right)^{N\overline{d}\overline{r}^2}e^{1-c_1 m\delta_{\overline{r}}^2}\leq e^{1-c_1 m\delta_{\overline{r}}^2+c_2 N\overline{d}\overline{r}^2\log N}, \qquad (37)
\end{aligned}$$

where in the last line, we choose $\epsilon=\frac{c\delta_{\overline{r}}}{N}$, and $c_2$ is a positive constant. Based on (37), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathcal{X}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right|\leq\frac{\frac{\delta_{\overline{r}}}{2}+c\delta_{\overline{r}}\sqrt{2/\pi}}{1-c\delta_{\overline{r}}}\right) \\
&\geq 1-e^{1-c_1 m\delta_{\overline{r}}^2+c_2 N\overline{d}\overline{r}^2\log N}. \qquad (38)
\end{aligned}$$

To guarantee that $\frac{\frac{\delta_{\overline{r}}}{2}+c\delta_{\overline{r}}\sqrt{2/\pi}}{1-c\delta_{\overline{r}}}\leq\delta_{\overline{r}}\leq\sqrt{2/\pi}$, we can select $c=\frac{\sqrt{2\pi}}{8}$ in $\epsilon=\frac{c\delta_{\overline{r}}}{N}$. Furthermore, if $m\geq\Omega(N\overline{d}\overline{r}^2\log N/\delta_{\overline{r}}^2)$, we obtain the following result: $\mathbb{P}\left(\sup_{\mathcal{X}}\left|\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1-\sqrt{2/\pi}\right|\leq\delta_{\overline{r}}\right)\geq 1-e^{-c_3 N\overline{d}\overline{r}^2\log N}$, where $c_3$ is a constant. In other words, with probability at least $1-e^{-c_3 N\overline{d}\overline{r}^2\log N}$, it holds that

$$(\sqrt{2/\pi}-\delta_{\overline{r}})\|\mathcal{X}\|_F\leq\frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_1\leq(\sqrt{2/\pi}+\delta_{\overline{r}})\|\mathcal{X}\|_F \qquad (39)$$

for any low-TT-rank tensor $\mathcal{X}\in\mathbb{R}^{d_1\times\cdots\times d_N}$ with ranks $(r_1,\dots,r_{N-1})$. □

## APPENDIX C
## PROOF OF LEMMA 1

*Proof:* We first expand $\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^\star)-\boldsymbol{s}\|_1-\frac{1}{m}\|\boldsymbol{s}\|_1$ as

$$\begin{aligned}
&\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^\star)-\boldsymbol{s}\|_1-\frac{1}{m}\|\boldsymbol{s}\|_1 \\
&\quad=\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}-\mathcal{X}^\star)\|_1+\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}-\mathcal{X}^\star) \\
&\qquad-\boldsymbol{s}\|_1-\frac{1}{m}\|\boldsymbol{s}\|_1 \\
&\quad\geq\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}-\mathcal{X}^\star)\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}-\mathcal{X}^\star)\|_1. \qquad (40)
\end{aligned}$$

In the subsequent part, we focus on analyzing the lower bound of $\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1$ for any tensor $\mathcal{X}$ in TT format with TT ranks smaller than $2\overline{r}$. We can construct an $\epsilon$-net $\{\mathcal{X}^{(p)}\}$ with the covering number $(\frac{4+\epsilon}{\epsilon})^{4N\overline{d}\overline{r}^2}$ for any low-TT-rank tensor $\mathcal{X}$ with ranks $(2r_1,\dots,2r_{N-1})$ such that (34) holds. Without loss of the generality, we assume that $\mathcal{X}=[\boldsymbol{X}_1,\dots,\boldsymbol{X}_N]$ is in left-orthogonal TT format with $\|\mathcal{X}\|_F=1$. Then we define $Y_k=\begin{cases}-|\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle|+\sqrt{2/\pi}, & k\in\mathcal{S},\\ |\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle|-\sqrt{2/\pi}, & k\in\mathcal{S}^c.\end{cases}$

Hoeffding inequality for Gaussian random variables tells that

$$\begin{aligned}
&\mathbb{P}\Bigg(\frac{1}{m}\sum_{k=1}^m Y_k=\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1 \\
&\quad-(1-2p_s)\sqrt{2/\pi}\geq-\frac{\delta_{2\overline{r}}}{2}\Bigg)\geq 1-e^{-c_1 m\delta_{2\overline{r}}^2}, \qquad (41)
\end{aligned}$$

where $c_1$ is a constant.

On the other hand, $\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1\geq\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}-\mathcal{X}^{(p)})\|_1$ and $\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\leq\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1+\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}-\mathcal{X}^{(p)})\|_1$ hold for any tensor $\mathcal{X}$. Hence, we have $\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\geq\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^{(p)})\|_1$.

Combing (34), we can get

$$\begin{aligned}
&\inf_{\mathcal{X}}\left(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\right) \\
&\geq\inf_{\mathcal{X}^{(p)}}\left(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1\right) \\
&\quad-\sup_{\mathcal{X}}\frac{1}{m}\|\mathcal{A}(\mathcal{X}-\mathcal{X}^{(p)})\|_1 \\
&\geq\inf_{\mathcal{X}^{(p)}}\left(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1\right) \\
&\quad-\sup_{\mathcal{X}}\frac{c\delta_{2\overline{r}}}{m}\|\mathcal{A}(\mathcal{X})\|_1 \\
&\geq\inf_{\mathcal{X}^{(p)}}\left(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X}^{(p)})\|_1-\frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X}^{(p)})\|_1\right) \\
&\quad-c\delta_{2\overline{r}}\left(\frac{\delta_{2\overline{r}}/2+c\delta_{2\overline{r}}\sqrt{2/\pi}}{1-c\delta_{2\overline{r}}}+\sqrt{2/\pi}\right),
\end{aligned}$$

where the last line follows the (38) with probability $1 - e^{-c_2 N\overline{d}\overline{r}^2 \log N}$ in which $c_2$ is a positive constant. Denote the event $F := \{(38) \text{ is satisfied}\}$ which holds with probability $1 - e^{-c_2 N\overline{d}\overline{r}^2 \log N}$. We can take the union bound with (41) to conclude

$$\begin{aligned}&\mathbb{P}\bigg(\inf_{\mathcal{X}}\bigg(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\bigg) \geq (1-2p_s)\sqrt{2/\pi}\\&\quad - \frac{\delta_{2\overline{r}}}{2} - c\delta_{2\overline{r}}\bigg(\frac{\delta_{2\overline{r}}/2 + c\delta_{2\overline{r}}\sqrt{2/\pi}}{1-c\delta_{2\overline{r}}} + \sqrt{2/\pi}\bigg)\bigg|F\bigg)\\&\geq 1 - \Pi_{i=1}^N n_i 2e^{c_1 m\delta_{2\overline{r}}^2} \geq 1 - \bigg(\frac{4+\epsilon}{\epsilon}\bigg)^{4N\overline{d}\overline{r}^2} e^{1-c_1 m\delta_{2\overline{r}}^2}\\&\geq 1 - e^{1-c_1 m\delta_{2\overline{r}}^2 + c_3 N\overline{d}\overline{r}^2\log N},\end{aligned} \tag{43}$$

where $c_3$ is a positive constant.

To guarantee that $-\frac{\delta_{2\overline{r}}}{2} - c\delta_{2\overline{r}}(\frac{\delta_{2\overline{r}}/2+c\delta_{2\overline{r}}\sqrt{2/\pi}}{1-c\delta_{2\overline{r}}} + \sqrt{2/\pi}) \geq -\delta_{2\overline{r}}$ and $\delta_{2\overline{r}} \leq \sqrt{2/\pi}$, we set $c = \frac{\sqrt{2\pi}}{8}$. While $m \geq \Omega(N\overline{d}\overline{r}^2\log N/\delta_{2\overline{r}}^2)$, we can obtain $\mathbb{P}\big(\inf_{\mathcal{X}}\big(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\big) \geq (1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}}\big|F\big) \geq 1 - e^{-c_4 N\overline{d}\overline{r}^2\log N}$, where $c_4$ is a constant. In the end, we can derive (44), shown at the bottom of the page, where $c_5$ is a positive constant. □

## Appendix D
## Proof of Lemma 3

*Proof:* We can apply the norm inequalities to derive

$$\begin{aligned}&\langle\partial f(\mathcal{X}), \mathcal{X} - \mathcal{X}^\star\rangle\\&= \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle\mathcal{A}_k, \mathcal{X}\rangle - y_k)(\langle\mathcal{A}_k, \mathcal{X} - \mathcal{X}^\star\rangle - s_k)\\&\quad + \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle\mathcal{A}_k, \mathcal{X}\rangle - y_k)s_k\\&\geq \frac{1}{m}\|\mathcal{A}(\mathcal{X} - \mathcal{X}^\star) - \boldsymbol{s}\|_1 - \frac{1}{m}\|\boldsymbol{s}\|_1\\&\geq ((1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}})\|\mathcal{X} - \mathcal{X}^\star\|_F,\end{aligned} \tag{45}$$

where the last line follows Lemma 1 for $\delta_{2\overline{r}} \leq (1-2p_s)\sqrt{2/\pi}$. □

## Appendix E
## Proof of Theorem 2

*Proof:* We expand $\|\mathcal{X}^{(t+1)} - \mathcal{X}^\star\|_F^2$ as following:

$$\begin{aligned}&\|\mathcal{X}^{(t+1)} - \mathcal{X}^\star\|_F^2 = \|\text{SVD}_{\boldsymbol{r}}^{tt}(\mathcal{X}^{(t)} - \mu_t\partial f(\mathcal{X}^{(t)})) - \mathcal{X}^\star\|_F^2\\&\leq \bigg(1 + \frac{600N}{\underline{\sigma}(\mathcal{X}^\star)}\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F\bigg)\|\mathcal{X}^{(t)} - \mu_t\partial f(\mathcal{X}^{(t)}) - \mathcal{X}^\star\|_F^2\\&= \bigg(1 + \frac{600N}{\underline{\sigma}(\mathcal{X}^\star)}\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F\bigg)\big(\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2\\&\quad - 2\mu_t\langle\partial f(\mathcal{X}^{(t)}), \mathcal{X}^{(t)} - \mathcal{X}^\star\rangle + \mu_t^2\|\partial f(\mathcal{X}^{(t)})\|_F^2\big)\\&\leq \bigg(1 + \frac{600N}{\underline{\sigma}(\mathcal{X}^\star)}\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F\bigg)\big(\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2\\&\quad - 2\mu_t((1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}})\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F\\&\quad + \mu_t^2(\sqrt{2/\pi} + \delta_{2\overline{r}})^2\big),\end{aligned} \tag{46}$$

where we use Lemma 2 in the first inequality and subsequently employ Lemma 3 and Lemma 7 in the last line.

Under the initial condition $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \leq \frac{c\underline{\sigma}(\mathcal{X}^\star)}{600N}$ with a constant $c$, (46) can be rewritten as

$$\begin{aligned}&\|\mathcal{X}^{(t+1)} - \mathcal{X}^\star\|_F^2 \leq (1+c)\big(\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - 2\mu_t((1-2p_s)\\&\quad \cdot\sqrt{2/\pi} - \delta_{2\overline{r}})\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F + \mu_t^2(\sqrt{2/\pi} + \delta_{2\overline{r}})^2\big).\end{aligned} \tag{47}$$

Based on the discussion in Appendix E-A, we can select $\mu_t = \lambda q^t$ where $\lambda = \frac{((1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}})}{2(\sqrt{2/\pi}+\delta_{2\overline{r}})^2}\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F$ and $q = \sqrt{(1+c)(1 - \frac{3((1-2p_s)\sqrt{2/\pi}-\delta_{2\overline{r}})^2}{4(\sqrt{2/\pi}+\delta_{2\overline{r}})^2}})$, and then obtain

$$\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 \leq \|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F^2 q^{2t}. \tag{48}$$

□

### A. Proof of (48)

*Proof:* Define $c_1 = 2((1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}})$ and $c_2 = (\sqrt{2/\pi} + \delta_{2\overline{r}})^2$.

Now we can simplify (47) as $\|\mathcal{X}^{(t+1)} - \mathcal{X}^\star\|_F^2 \leq (1+c)(\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F^2 - \mu_t c_1\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F + \mu_t^2 c_2)$. Next, we aim to show

$$\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F \leq c_0 q^{\frac{t}{2}} = c_0(1+c)^{\frac{t}{2}}p^{\frac{t}{2}} \tag{49}$$

where $c_0 = \|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F$ and $p$ is a parameter that needs to be determined. Let us therefore fix a value $x \in [0,1]$ satisfying $\|\mathcal{X}^{(t)} - \mathcal{X}^\star\|_F = xc_0(1+c)^{\frac{t}{2}}p^{\frac{t}{2}}$. Assume the above induction hypothesis (49) holds at the $t$-iteration. We need to further prove

$$\begin{aligned}&\|\mathcal{X}^{(t+1)} - \mathcal{X}^\star\|_F^2 \leq (1+c)\big(c_0^2(1+c)^t p^t x^2 - \mu_t c_0 c_1(1+c)^{\frac{t}{2}}\\&\quad \cdot p^{\frac{t}{2}}x + \mu_t^2 c_2\big) \leq c_0^2(1+c)^{t+1}p^{t+1}.\end{aligned} \tag{50}$$

$$\begin{aligned}&\mathbb{P}\left(\inf_{\mathcal{X}}\big(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\big) \geq (1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}}\right)\\&\geq \mathbb{P}\left(\inf_{\mathcal{X}}\big(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\big) \geq (1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}} \cap F\right)\\&= \mathbb{P}(F)\,\mathbb{P}\left(\inf_{\mathcal{X}}\big(\frac{1}{m}\|\mathcal{A}_{\mathcal{S}^c}(\mathcal{X})\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{S}}(\mathcal{X})\|_1\big) \geq (1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}}\middle|F\right)\\&\geq (1 - e^{-c_2 N\overline{d}\overline{r}^2\log N})(1 - e^{-c_4 N\overline{d}\overline{r}^2\log N}) \geq 1 - 2e^{-c_5 N\overline{d}\overline{r}^2\log N},\end{aligned} \tag{44}$$

When we select $\mu_t = \lambda q^t = c_0 a(1+c)^{\frac{t}{2}} p^{\frac{t}{2}}$ where $a$ is a parameter which needs to be determined, (50) can be simplified as

$$x^2 - c_1 a x + c_2 a^2 \leq p. \tag{51}$$

Note that the left hand side of (51) is a convex quadratic in $x$ and therefore the maximum between [0, 1] must occur either at $x = 0$ or $x = 1$.

- When selecting $x = 0$, we can derive the condition $c_2 a^2 \leq p$.
- When choosing $x = 1$, we have $c_2 a^2 - c_1 a + 1 - p \leq 0$. Since $c_1, c_2 > 0$, we have $\frac{c_1 - \sqrt{c_1^2 - 4c_2(1-p)}}{2c_2} \leq a \leq \frac{c_1 + \sqrt{c_1^2 - 4c_2(1-p)}}{2c_2}$.

By selecting $p = 1 - \frac{3c_1^2}{16c_2}$ and $a = \frac{c_1}{4c_2}$, we can ensure that conditions $c_2 a^2 \leq p$ and $c_2 a^2 - c_1 a + 1 - p \leq 0$ holds for $\delta_{2\overline{r}} \geq 0$. Hence we can respectively choose $\lambda = \frac{c_0 c_1}{4c_2}$ and $q = \sqrt{(1+c)(1 - \frac{3c_1^2}{16c_2})}$, which further guarantees (50). □

## Appendix F
## Proof of Lemma 5

*Proof:* First, we provide one useful property. Based on $\operatorname{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})^2}{18(2N^2-2N+1)(N+1+\sum_{i=2}^{N-1} r_i)(\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2}$ which can be obtained by $\mathcal{X} \in \mathcal{C}(b)$ and Lemma 4, for $i \in [N-1]$, we can obtain

$$\begin{aligned}
&\sigma_1^2(\mathcal{X}^{\langle i \rangle}) = \|\mathcal{X}^{\geq i+1}\|^2 \\
&\leq \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 2\|\boldsymbol{R}_i^\top \mathcal{X}^{\star \geq i+1}\|^2 + 2\|\mathcal{X}^{\geq i+1} - \boldsymbol{R}_i^\top \mathcal{X}^{\star \geq i+1}\|^2 \\
&\leq 2\overline{\sigma}^2(\mathcal{X}^\star) + \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 2\|\mathcal{X}^{\langle i \rangle} - \mathcal{X}^{\star \langle i \rangle} + \mathcal{X}^{\star \langle i \rangle} \\
&\quad - \mathcal{X}^{\leq i} \boldsymbol{R}_i^\top \mathcal{X}^{\star \geq i+1}\|^2 \\
&\leq 2\overline{\sigma}^2(\mathcal{X}^\star) + 4\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \\
&\quad + \min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}} 4\|\boldsymbol{R}_i^\top \mathcal{X}^{\star \geq i+1}\|^2 \|\mathcal{X}^{\leq i} - \mathcal{X}^{\star \leq i} \boldsymbol{R}_i\|_F^2 \\
&\leq 2\overline{\sigma}^2(\mathcal{X}^\star) + \left(4 + \frac{16\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)}\right)\|\mathcal{X} - \mathcal{X}^\star\|_F^2 \leq 2\overline{\sigma}^2(\mathcal{X}^\star) \\
&\quad + \frac{45N\overline{\sigma}^2(\mathcal{X}^\star)}{\underline{\sigma}^2(\mathcal{X}^\star)} \operatorname{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4},
\end{aligned} \tag{52}$$

where the fourth and last lines respectively follow [49, eq. (60)] and Lemma 4. Note that $\overline{\sigma}^2(\mathcal{X}) = \max_{i=1}^{N-1} \sigma_1^2(\mathcal{X}^{\langle i \rangle}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}$.

Then we need to define the subgradient of $F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)$ as following:

$$\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) = \begin{bmatrix} \partial_{\boldsymbol{X}_i(1)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \\ \vdots \\ \partial_{\boldsymbol{X}_i(d_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \end{bmatrix}. \tag{53}$$

Here the subgradient with respect to each factor $\boldsymbol{X}_i(s_i)$ can be computed as $\partial_{\boldsymbol{X}_i(s_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) = \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle - y_k) \sum_{\substack{s_1,\dots,s_{i-1},\\ s_{i+1},\dots,s_N}} (\mathcal{A}_k(s_1, \dots, s_N) \boldsymbol{X}_{i-1}^\top(s_{i-1}) \cdots \boldsymbol{X}_1^\top(s_1) \cdot \boldsymbol{X}_N^\top(s_N) \cdots \boldsymbol{X}_{i+1}^\top(s_{i+1}))$.

Before analyzing the robust regularity condition, we need to define three matrices for $i \in [N]$ as follows:

$$\begin{aligned}
\boldsymbol{D}_1(i) &= \begin{bmatrix} \boldsymbol{X}_{i-1}^\top(1) \cdots \boldsymbol{X}_1^\top(1) & \cdots & \boldsymbol{X}_{i-1}^\top(d_{i-1}) \cdots \boldsymbol{X}_1^\top(d_1) \end{bmatrix} \\
&= L^\top(\boldsymbol{X}_{i-1}) \overline{\otimes} \cdots \overline{\otimes} L^\top(\boldsymbol{X}_1) \in \mathbb{R}^{r_i \times (d_1 \cdots d_{i-1})},
\end{aligned} \tag{54}$$

$$\boldsymbol{D}_2(i) = \begin{bmatrix} \boldsymbol{X}_N^\top(1) \cdots \boldsymbol{X}_{i+1}^\top(1) \\ \vdots \\ \boldsymbol{X}_N^\top(d_N) \cdots \boldsymbol{X}_{i+1}^\top(d_{i+1}) \end{bmatrix} \in \mathbb{R}^{(d_{i+1} \cdots d_N) \times r_i}, \tag{55}$$

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 $\boldsymbol{E}(s_i)(s_1 \cdots s_{i-1}, s_{i+1} \cdots s_N) = \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle - y_k)\mathcal{A}_k(s_1, \dots, s_N)$.

Based on the aforementioned notations, we can derive

$$\begin{aligned}
&\|\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)\|_F^2 \\
&= \sum_{s_i=1}^{d_i} \|\partial_{\boldsymbol{X}_i(s_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)\|_F^2 \\
&= \sum_{s_i=1}^{d_i} \|\boldsymbol{D}_1(i)\boldsymbol{E}(s_i)\boldsymbol{D}_2(i)\|_F^2 \\
&\leq \sum_{s_i=1}^{d_i} \|L^\top(\boldsymbol{X}_{i-1}) \overline{\otimes} \cdots \overline{\otimes} L^\top(\boldsymbol{X}_1)\|^2 \|\boldsymbol{D}_2(i)\|^2 \|\boldsymbol{E}(s_i)\|_F^2 \\
&\leq \|L(\boldsymbol{X}_1)\|^2 \cdots \|L(\boldsymbol{X}_{i-1})\|^2 \|\mathcal{X}^{\geq i+1}\|^2 \|\frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle \\
&\quad - y_k)\mathcal{A}_k\|_F^2 \leq \begin{cases} \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}(\sqrt{2/\pi} + \delta_{2\overline{r}})^2, & i \in [N-1], \\ (\sqrt{2/\pi} + \delta_{2\overline{r}})^2, & i = N, \end{cases}
\end{aligned} \tag{56}$$

where we use (A), $\|\boldsymbol{D}_2(i)\| = \|\mathcal{X}^{\geq i+1}\|$ and $\sum_{s_i=1}^{d_i} \|\boldsymbol{E}(s_i)\|_F^2 = \|\frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle - y_k)\mathcal{A}_k\|_F^2$ in the second inequality. In addition, the third inequality follows $\|\mathcal{X}^{\geq i+1}\| = \sigma_1(\mathcal{X}^{\langle i \rangle}) \leq \frac{3\overline{\sigma}(\mathcal{X}^\star)}{2}$ and Lemma 7.

Now, we rewrite the cross term in the robust regularity condition as following:

$$\begin{aligned}
&\sum_{i=1}^N \left\langle L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}\big(\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)\big) \right\rangle \\
&= \sum_{i=1}^N \left\langle L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \right\rangle - T \\
&= \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle - y_k)\langle \mathcal{A}_k, \mathcal{X} - \mathcal{X}^\star\rangle \\
&\quad + \frac{1}{m}\sum_{k=1}^m \operatorname{sign}(\langle \mathcal{A}_k, \mathcal{X}\rangle - y_k)\langle \operatorname{vec}(\mathcal{A}_k), \boldsymbol{h}\rangle - T,
\end{aligned} \tag{57}$$

where $\boldsymbol{h} = L_{\boldsymbol{R}}(\boldsymbol{X}_1^\star) \overline{\otimes} \cdots \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star) - L(\boldsymbol{X}_1) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_{N-1}) \overline{\otimes} L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star) + \sum_{i=1}^{N-1} L(\boldsymbol{X}_1) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_{i-1}) \overline{\otimes} (L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)) \overline{\otimes} L(\boldsymbol{X}_{i+1}) \overline{\otimes} \cdots \overline{\otimes} L(\boldsymbol{X}_N)$ and $T =$

$\sum_{i=1}^{N-1} \langle L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)) \rangle$.

To get the lower bound of (57), we need to obtain upper bounds of $\boldsymbol{h}$ and $T$. According to [49, eq. (82)], we directly obtain

$$\|\boldsymbol{h}\|_2^2 \leq \frac{9N(N-1)}{8\overline{\sigma}^2(\mathcal{X}^\star)} \mathrm{dist}^4(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}). \tag{58}$$

Then, we can derive

$$\begin{aligned} T &= \sum_{i=1}^{N-1} \left\langle \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)), \partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N) \right\rangle \\ &\leq \frac{1}{2} \sum_{i=1}^{N-1} \|L(\boldsymbol{X}_i)\| \|L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 \|\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)\|_F \\ &\leq \frac{3\overline{\sigma}(\mathcal{X}^\star)}{4} (\sqrt{2/\pi} + \delta_{2\overline{r}}) \sum_{i=1}^{N-1} \|L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 \\ &\leq \frac{3}{4\overline{\sigma}(\mathcal{X}^\star)} (\sqrt{2/\pi} + \delta_{2\overline{r}}) \mathrm{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}), \end{aligned} \tag{59}$$

where the first inequality follows (56) and $\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(\cdot)$ is defined as

$$\begin{aligned} &\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}^{\perp}(L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)) \\ &= L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star) - \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)) \\ &= \frac{1}{2} L(\boldsymbol{X}_i)((L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star))^\top L(\boldsymbol{X}_i) \\ &\quad + L^\top(\boldsymbol{X}_i)(L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star))) \\ &= \frac{1}{2} L(\boldsymbol{X}_i)(2\mathbf{I}_{r_i} - L_{\boldsymbol{R}}^\top(\boldsymbol{X}_i^\star) L(\boldsymbol{X}_i) - L^\top(\boldsymbol{X}_i) L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)) \\ &= \frac{1}{2} L(\boldsymbol{X}_i)((L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star))^\top (L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star))). \end{aligned} \tag{60}$$

Ultimately, we arrive at

$$\begin{aligned} &\sum_{i=1}^{N} \left\langle L(\boldsymbol{X}_i) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star), \mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)) \right\rangle \\ &\geq \frac{1}{m} \sum_{k=1}^{m} |\langle \mathcal{A}_k, \mathcal{X} - \mathcal{X}^\star \rangle - s_k| - \frac{1}{m} \sum_{k=1}^{m} |s_k| - \frac{1}{m} \sum_{k=1}^{m} |\langle \mathrm{vec}(\mathcal{A}_k), \\ &\boldsymbol{h} \rangle| - \frac{3}{4\overline{\sigma}(\mathcal{X}^\star)} (\sqrt{2/\pi} + \delta_{2\overline{r}}) \mathrm{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \\ &\geq ((1-2p_s)\sqrt{2/\pi} - \delta_{2\overline{r}}) \|\mathcal{X} - \mathcal{X}^\star\|_F - (\sqrt{2/\pi} + \delta_{(N+1)\overline{r}}) \|\boldsymbol{h}\|_2 \\ &\quad - \frac{3}{4\overline{\sigma}(\mathcal{X}^\star)} (\sqrt{2/\pi} + \delta_{2\overline{r}}) \mathrm{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \\ &\geq \frac{\underline{\sigma}(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})}{4\sqrt{2(N+1+\sum_{i=2}^{N-1} r_i)}\overline{\sigma}(\mathcal{X}^\star)} \mathrm{dist}(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \end{aligned} \tag{61}$$

where (59) is used in the first inequality. Note that $\boldsymbol{h}$ can be viewed as a TT format where the rank is at most $((N-1)r_1, \dots, (N-1)r_{N-1})$. Hence, we apply $\ell_1/\ell_2$-RIP and Lemma 1 in the second inequality. The last line follows $\delta_{2\overline{r}} \leq \delta_{(N+1)\overline{r}} \leq (1-2p_s)\sqrt{2/\pi}$, (58), Lemma 4 and $\mathrm{dist}^2(\{\boldsymbol{X}_i\}, \{\boldsymbol{X}_i^\star\}) \leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})^2}{18(2N^2-2N+1)(N+1+\sum_{i=2}^{N-1} r_i)(\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2}$. □

## Appendix G
## Proof of Theorem 3

*Proof:* To utilize the robust regularity condition of Lemma 5 in the derivation of Theorem 3, we need to prove conditions in Lemma 5. Due to the retraction operation, we can guarantee that $L(\boldsymbol{X}_i^{(t)})$ are orthonormal. In addition, we assume that

$$\begin{aligned} &\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\ &\leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})^2}{18(2N^2-2N+1)(N+1+\sum_{i=2}^{N-1} r_i)(\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2}, \end{aligned} \tag{62}$$

which can be proven later, and following (52), then obtain $\sigma_1^2(\mathcal{X}^{(t)\langle i \rangle}) \leq \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i \in [N-1]$.

Next, we define the best rotation matrices as following: $(\boldsymbol{R}_1^{(t)}, \dots, \boldsymbol{R}_{N-1}^{(t)}) = \arg\min_{\boldsymbol{R}_i \in \mathbb{O}^{r_i \times r_i}, i \in [N-1]} \sum_{i=1}^{N-1} \overline{\sigma}^2(\mathcal{X}^\star) \cdot \|L(\boldsymbol{X}_i^{(t)}) - L_{\boldsymbol{R}}(\boldsymbol{X}_i^\star)\|_F^2 + \|L(\boldsymbol{X}_N^{(t)}) - L_{\boldsymbol{R}}(\boldsymbol{X}_N^\star)\|_2^2$.

Now we can prove the assumption $\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq \mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\})$, expand $\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\})$ and subsequently derive (63), shown at the bottom of the next page, where the first inequality follows the nonexpansiveness property of Lemma 8.

Based on (56), we can easily obtain

$$\begin{aligned} &\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \|\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)}))\|_F^2 \\ &\quad + \|\partial_{L(\boldsymbol{X}_N)} F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})\|_2^2 \\ &\leq \frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)} \sum_{i=1}^{N-1} \|\partial_{L(\boldsymbol{X}_i)} F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})\|_F^2 \\ &\quad + \|\partial_{L(\boldsymbol{X}_N)} F(\boldsymbol{X}_1^{(t)}, \dots, \boldsymbol{X}_N^{(t)})\|_2^2 \leq \frac{9N-5}{4} (\sqrt{2/\pi} + \delta_{2\overline{r}})^2, \end{aligned} \tag{64}$$

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 $\|\mathcal{P}_{\mathrm{T}_{L(\boldsymbol{X}_i)}\mathrm{St}}(\boldsymbol{B})\|_F^2 \leq \|\boldsymbol{B}\|_F^2$.

Hence, combing the robust regularity condition in Lemma 5 and (64), we have

$$\begin{aligned} &\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\}, \{\boldsymbol{X}_i^\star\}) \leq \mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\ &\quad - \frac{\underline{\sigma}(\mathcal{X}^\star)((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})}{2\sqrt{2(N+1+\sum_{i=2}^{N-1} r_i)}\overline{\sigma}(\mathcal{X}^\star)} \mu_t \mathrm{dist}(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \\ &\quad + \frac{9N-5}{4} (\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2 \mu_t^2. \end{aligned} \tag{65}$$

Following the same analysis of (48) in Appendix E-A, we can set $\mu_t = \lambda q^t$ where we respectively select $\lambda = \frac{(1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}}}{\sqrt{2(N+1+\sum_{i=2}^{N-1} r_i)(9N-5)(\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2 \kappa(\mathcal{X}^\star)}} \mathrm{dist}(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\})$ and $q = \sqrt{1 - \frac{((1-2p_s)\sqrt{2/\pi} - \delta_{(N+1)\overline{r}})^2}{8(N+1+\sum_{i=2}^{N-1} r_i)(9N-5)(\sqrt{2/\pi} + \delta_{(N+1)\overline{r}})^2 \kappa^2(\mathcal{X}^\star)}}$, and then guarantee that

$$\mathrm{dist}^2(\{\boldsymbol{X}_i^{(t)}\}, \{\boldsymbol{X}_i^\star\}) \leq \mathrm{dist}^2(\{\boldsymbol{X}_i^{(0)}\}, \{\boldsymbol{X}_i^\star\}) q^{2t}. \tag{66}$$

*Proof:* of (62): We now prove (62) by induction. First note that (62) holds for $t = 0$ which can be proved by combing $\mathcal{X}^{(0)} \in \mathcal{C}(b)$ and Lemma 4. We now assume it holds at $t=t'$, which implies that $\sigma_1^2(\mathcal{X}^{(t')^{\langle i\rangle}}) \le \frac{9\overline{\sigma}^2(\mathcal{X}^\star)}{4}, i\in[N-1]$. By invoking (66), we have $\text{dist}^2(\{\boldsymbol{X}_i^{(t'+1)}\},\{\boldsymbol{X}_i^\star\}) \le \text{dist}^2(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})$. Consequently, (62) also holds at $t=t'+1$. By induction, we can conclude that (62) holds for all $t\ge 0$. This completes the proof. □

## Appendix H
## Proof of Theorem 4

*Proof:* Before analyzing the truncated spectral initialization, we first define the following restricted Frobenius norm for any tensor $\mathcal{H}\in\mathbb{R}^{d_1\times\cdots\times d_N}$: $\|\mathcal{H}\|_{F,\overline{r}} = \max_{\substack{\mathcal{X}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{X}\|_F\le 1,\\ \text{rank}(\mathcal{X})=(r_1,\dots,r_{N-1})}} \langle \mathcal{H},\mathcal{X}\rangle$, where $\text{rank}(\mathcal{X})$ denotes the TT ranks of $\mathcal{X}$.

Following the same analysis of [49, eq. (89)], we have

$$\begin{aligned}
&\|\mathcal{X}^{(0)}-\mathcal{X}^\star\|_F \le (1+\sqrt{N-1})\\
&\times\left\|\frac{1}{(1-p_s)m}\sum_{k=1}^m y_k\mathcal{A}_k\mathbb{I}_{\{|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\}}-\mathcal{X}^\star\right\|_{F,2\overline{r}}\\
&=\max_{\substack{\mathcal{X}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{X}\|_F\le 1,\\ \text{rank}(\mathcal{X})=(2r_1,\dots,2r_{N-1})}} \frac{1+\sqrt{N-1}}{(1-p_s)m}\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}\rangle,
\end{aligned} \tag{67}$$

where $\mathcal{S}'=\{k:|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)},k\in[m]\}$ and $|\mathcal{S}'|=\lceil(1-p_s)m\rceil$. Since $|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\le\max_k|\langle\mathcal{A}_k,\mathcal{X}^\star\rangle|$ where $\langle\mathcal{A}_k,\mathcal{X}^\star\rangle\sim\mathcal{N}(0,\|\mathcal{X}^\star\|_F^2)$, we first have $\mathbb{P}(|\langle\mathcal{A}_k,\mathcal{X}^\star\rangle|\ge t)\le 2e^{-\frac{t^2}{2\|\mathcal{X}^\star\|_F^2}}$, and it follows that $\mathbb{P}(\max_k|\langle\mathcal{A}_k,\mathcal{X}^\star\rangle|\le t)\ge 1-2\lceil(1-p_s)m\rceil e^{-\frac{t^2}{2\|\mathcal{X}^\star\|_F^2}}\ge 1-e^{1-\frac{t^2}{2\|\mathcal{X}^\star\|_F^2}+\log((1-p_s)m)}$.

Then taking $t=c_1\|\mathcal{X}^\star\|_F\log((1-p_s)m)$ with a positive constant $c_1$, with probability $1-e^{-\Omega(\log((1-p_s)m))}$ we can obtain $|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\le\max_k|\langle\mathcal{A}_k,\mathcal{X}^\star\rangle|\le O(\|\mathcal{X}^\star\|_F\log((1-p_s)m))$.

Next, according to [49, Appendix E], we can construct an $\epsilon$-net $\{\mathcal{X}^{(1)},\dots,\mathcal{X}^{(p)}\}$ with covering number $p\le(\frac{4+\epsilon}{\epsilon})^{4N\overline{d}\overline{r}^2}$ for any TT format tensors $\mathcal{X}$ with TT ranks $(2r_1,\dots,2r_{N-1})$ such that

$$\begin{aligned}
&\max_{\substack{\mathcal{X}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{X}\|_F\le 1,\\ \text{rank}(\mathcal{X})=(2r_1,\dots,2r_{N-1})}}\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}\rangle\\
&\le 2\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle
\end{aligned}\tag{68}$$

using $\epsilon=\frac{1}{2N}$.

Note that $\mathbb{E}_{\mathcal{A}_k}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle=\mathbb{E}_{\mathcal{A}_k}(\langle\mathcal{A}_k,\mathcal{X}^\star\rangle+s_k)\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle-\langle\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle=\langle\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle-\langle\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle=0$ since each element in $\mathcal{A}_k$ follows the normal distribution. In addition, $\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle$ is a subgaussian random variable with subgaussian norm $\|\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle\|_{\psi_2}\le|y_k|\|\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle\|_{\psi_2}+\|\mathcal{X}^\star\|_F\|\mathcal{X}^{(p)}\|_F\le O(\log((1-p_s)m)\|\mathcal{X}^\star\|_F)$ where we use $|y_k|\le|\boldsymbol{y}|_{(\lceil p_s m\rceil)}\le O(\|\log((1-p_s)m)\|\mathcal{X}^\star\|_F)$, $\langle\mathcal{A}_k,\mathcal{X}^{(p)}\rangle\sim\mathcal{N}(0,\|\mathcal{X}^{(p)}\|_F^2)$ and $\|\mathcal{X}^{(p)}\|_F\le 1$. According to the General Hoeffding's inequality [66, Theorem 2.6.2], we have

$$\begin{aligned}
&\mathbb{P}\Big(|\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle|\ge t\Big)\\
&\le 2e^{-\frac{c_2t^2}{(1-p_s)m(\log((1-p_s)m))^2\|\mathcal{X}^\star\|_F^2}},
\end{aligned}\tag{69}$$

where $c_2$ is a positive constant. Combing (68) with (69), we further derive

$$\begin{aligned}
&\mathbb{P}\Big(\max_{\substack{\mathcal{X}\in\mathbb{R}^{d_1\times\cdots\times d_N},\|\mathcal{X}\|_F\le 1,\\ \text{rank}(\mathcal{X})=(2r_1,\dots,2r_{N-1})}}\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}\rangle\ge t\Big)\\
&\le\mathbb{P}\left(\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle\ge\frac{t}{2}\right)\\
&\le\mathbb{P}\left(|\sum_{k\in\mathcal{S}'}\langle y_k\mathcal{A}_k-\mathcal{X}^\star,\mathcal{X}^{(p)}\rangle|\ge\frac{t}{2}\right)\\
&\le\left(\frac{4+\epsilon}{\epsilon}\right)^{4N\overline{d}\overline{r}^2}e^{1-\frac{c_2t^2}{4(1-p_s)m(\log((1-p_s)m))^2\|\mathcal{X}^\star\|_F^2}}\\
&\le e^{1-\frac{c_2t^2}{4(1-p_s)m(\log((1-p_s)m))^2\|\mathcal{X}^\star\|_F^2}+c_3N\overline{d}\overline{r}^2\log N},
\end{aligned}\tag{70}$$

$$\begin{aligned}
&\text{dist}^2(\{\boldsymbol{X}_i^{(t+1)}\},\{\boldsymbol{X}_i^\star\})\\
&=\sum_{i=1}^{N-1}\overline{\sigma}^2(\mathcal{X}^\star)\|L(\boldsymbol{X}_i^{(t+1)})-L_{\boldsymbol{R}^{(t+1)}}(\boldsymbol{X}_i^\star)\|_F^2+\|L(\boldsymbol{X}_N^{(t+1)})-L_{\boldsymbol{R}^{(t+1)}}(\boldsymbol{X}_N^\star)\|_2^2\\
&\le\sum_{i=1}^{N-1}\overline{\sigma}^2(\mathcal{X}^\star)\left\|L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star)-\frac{\mu_t}{\overline{\sigma}^2(\mathcal{X}^\star)}\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}}\big(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1^{(t)},\dots,\boldsymbol{X}_N^{(t)})\big)\right\|_F^2\\
&\quad+\|L(\boldsymbol{X}_N^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_N^\star)-\mu_t\partial_{L(\boldsymbol{X}_N)}F(\boldsymbol{X}_1^{(t)},\dots,\boldsymbol{X}_N^{(t)})\|_2^2\\
&=\text{dist}^2(\{\boldsymbol{X}_i^{(t)}\},\{\boldsymbol{X}_i^\star\})-2\mu_t\sum_{i=1}^N\left\langle L(\boldsymbol{X}_i^{(t)})-L_{\boldsymbol{R}^{(t)}}(\boldsymbol{X}_i^\star),\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}}\big(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1^{(t)},\dots,\boldsymbol{X}_N^{(t)})\big)\right\rangle\\
&\quad+\mu_t^2\left(\frac{1}{\overline{\sigma}^2(\mathcal{X}^\star)}\sum_{i=1}^{N-1}\|\mathcal{P}_{\text{T}_{L(\boldsymbol{X}_i)}\text{St}}\big(\partial_{L(\boldsymbol{X}_i)}F(\boldsymbol{X}_1^{(t)},\dots,\boldsymbol{X}_N^{(t)})\big)\|_F^2+\|\partial_{L(\boldsymbol{X}_N)}F(\boldsymbol{X}_1^{(t)},\dots,\boldsymbol{X}_N^{(t)})\|_2^2\right),
\end{aligned}\tag{63}$$

where $c_3$ is a constant and based on the assumption in (68), $\frac{4+\epsilon}{\epsilon} = \frac{4+\frac{1}{2N}}{\frac{1}{2N}} = 8N + 1$.

Taking $t = \log((1-p_s)m)\|\mathcal{X}^\star\|_F\sqrt{(1-p_s)mN\overline{d}\overline{r}^2\log N}$, with probability $1 - e^{-\Omega(N\overline{d}\overline{r}^2\log N)} - e^{-\Omega(\log((1-p_s)m))}$, we have $\|\mathcal{X}^{(0)} - \mathcal{X}^\star\|_F \leq O(\frac{N\overline{r}\log((1-p_s)m)\|\mathcal{X}^\star\|_F\sqrt{\overline{d}}\log N}{\sqrt{(1-p_s)m}})$. □

## References


[1] A. Cichocki et al., "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar. 2015.

[2] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, Jul. 2017.

[3] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, Mar. 2000.

[4] A. Lidiak et al., "Quantum state tomography with tensor train cross approximation," 2022, *arXiv:2207.06397*.

[5] Z. Qin, C. Jameson, Z. Gong, M. B. Wakin, and Z. Zhu, "Quantum state tomography for matrix product density operators," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 5030–5056, Jul. 2024.

[6] Z. Qin, J. M. Lukens, B. T. Kirby, and Z. Zhu, "Enhancing quantum state reconstruction with structured classical shadows," 2025, *arXiv:2501.03144*.

[7] A. K. Smilde, P. Geladi, and R. Bro, *Multi-Way Analysis: applications in the Chemical Sciences.* Hoboken, NJ, USA: Wiley, 2005.

[8] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 1, pp. 6–20, Jan. 2009.

[9] V. Hore et al., "Tensor decomposition for multiple-tissue gene expression experiments," *Nat. Genet.*, vol. 48, no. 9, pp. 1094–1100, 2016.

[10] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vision*, vol. 54, pp. 117–142, 2003.

[11] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[12] W. Guo, I. Kotsia, and I. Patras, "Tensor learning for regression," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 816–827, Feb. 2012.

[13] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Am. Stat. Assoc.*, vol. 108, no. 502, pp. 540–552, 2013.

[14] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 397–408, Jan. 2016.

[15] L. Li and X. Zhang, "Parsimonious tensor response regression," *J. Am. Stat. Assoc.*, vol. 112, no. 519, pp. 1131–1146, 2017.

[16] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned robust low-rank tensor regression," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 9072–9076.

[17] K. Li and S. Cong, "A robust compressive quantum state tomography algorithm using ADMM," *IFAC Proc. Volumes*, vol. 47, no. 3, pp. 6878–6883, 2014.

[18] S. Hara, T. Ono, R. Okamoto, T. Washio, and S. Takeuchi, "Anomaly detection in reconstructed quantum states using a machine-learning technique," *Phys. Rev. A*, vol. 89, no. 2, 2014, Art. no. 022104.

[19] I. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.

[20] R. Bro, "Parafac. Tutorial and applications," *Chemometrics Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, 1997.

[21] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[22] H. Johan, "Tensor rank is np-complete," *J. Algorithms*, vol. 4, no. 11, pp. 644–654, 1990.

[23] V. De Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1084–1127, 2008.

[24] A. Cichocki, N. Lee, I. Oseledets, A. Phan, Q. Zhao, and D. Mandic, "Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions," *Found. Trends® Mach. Learn.*, vol. 9, nos. 4–5, pp. 249–429, 2016.

[25] H. Cai, Z. Chao, L. Huang, and D. Needell, "Robust tensor cur decompositions: Rapid low-tucker-rank tensor recovery with sparse corruptions," *SIAM J. Imag. Sci.*, vol. 17, no. 1, pp. 225–247, 2024.

[26] C. Llosa-Vite and R. Maitra, "Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2282–2296, Feb. 2023.

[27] P. D. Hoff, "Multilinear tensor regression for longitudinal relational data," *Ann. Appl. Statist.*, vol. 9, no. 3, 2015, Art. no. 1169.

[28] Y. Liu, J. Liu, and C. Zhu, "Low-rank tensor train coefficient array estimation for tensor-on-tensor regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5402–5411, Dec. 2020.

[29] F. D. Mandanas and C. L. Kotropoulos, "Robust multidimensional scaling using a maximum correntropy criterion," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 919–932, Feb. 2017.

[30] Z.-Y. Wang, H. C. So, and A. M. Zoubir, "Robust low-rank matrix recovery via hybrid ordinary-welsch function," *IEEE Trans. Signal Process.*, vol. 71, pp. 2548–2563, 2023.

[31] Z.-Y. Wang, H. C. So, and A. M. Zoubir, "Low-rank tensor completion via novel sparsity-inducing regularizers," *IEEE Trans. Signal Process.*, vol. 72, pp. 3519–3534, 2024.

[32] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *Inf. Inference: A J. IMA*, vol. 9, no. 2, pp. 289–325, 2020.

[33] Z.-Y. Wang, X. P. Li, and H. C. So, "Robust matrix completion based on factorization and truncated-quadratic loss function," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1521–1534, Apr. 2023.

[34] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[35] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Inf. Inference: A J. IMA*, vol. 8, no. 3, pp. 471–529, 2019.

[36] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal, "Nonconvex robust low-rank matrix recovery," *SIAM J. Optim.*, vol. 30, no. 1, pp. 660–686, 2020.

[37] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, "Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence," *Found. Comput. Math.*, vol. 21, no. 6, pp. 1505–1593, 2021.

[38] J. Ma and S. Fattahi, "Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers," 2021, *arXiv:2102.02969*.

[39] L. Ding, L. Jiang, Y. Chen, Q. Qu, and Z. Zhu, "Rank overspecified robust matrix recovery: Subgradient method and exact recovery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26–767.

[40] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra Appl.*, vol. 523, pp. 220–262, 2017.

[41] M. Zhang, Z.-H. Huang, and Y. Zhang, "Restricted *p*-isometry properties of nonconvex matrix recovery," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4316–4323, Jul. 2013.

[42] M.-C. Yue and A. M.-C. So, "A perturbation inequality for concave functions of singular values and its applications in low-rank matrix recovery," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 396–416, 2016.

[43] T. Tong, "Scaled gradient methods for ill-conditioned low-rank matrix and tensor estimation," Ph.D. dissertation, Carnegie Mellon Univ., 2022,

[44] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 4034–4059, Jul. 2015.

[45] J. V. Burke and M. C. Ferris, "Weak sharp minima in mathematical programming," *SIAM J. Control Optim.*, vol. 31, no. 5, pp. 1340–1359, 1993.

[46] P. Marcotte and D. Zhu, "Weak sharp solutions of variational inequalities," *SIAM J. Optim.*, vol. 9, no. 1, pp. 179–189, 1998.

[47] T. Tong, C. Ma, and Y. Chi, "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," *IEEE Trans. Signal Process.*, vol. 69, pp. 2396–2409, 2021.

[48] Z. Qin and Z. Zhu, "Computational and statistical guarantees for tensor-on-tensor regression with tensor train decomposition," 2024, *arXiv:2406.06002*.

[49] Z. Qin, M. B. Wakin, and Z. Zhu, "Guaranteed nonconvex factorization approach for tensor train recovery," *J. Mach. Learn. Res.*, vol. 25, no. 383, pp. 1–48, 2024.

[50] A. Nedić and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," *Stochastic Optim.: Algorithms Appl.*, 2001, pp. 223–264.

[51] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization," *Optim. Mach. Learn., Neural Inf. Process. Ser.*, 2012, pp. 85–119.

[52] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3. New York, NY, USA: Springer, 2012.

[53] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow," in *Proc. Int. Conf. Mach. Learn.,* PMLR, 2016, pp. 1022–1031.

[54] J. Ma and S. Fattahi, "Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization," *J. Mach. Learn. Res.*, vol. 24, no. 96, pp. 1–84, 2023.

[55] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 172–208, 2019.

[56] R. Grotheer, S. Li, A. Ma, D. Needell, and J. Qin, "Iterative hard thresholding for low cp-rank tensor models," *Linear Multilinear Algebra*, vol. 70, no. 22, pp. 7452–7468, 2022.

[57] S. Budzinskiy and N. Zamarashkin, "Tensor train completion: Local recovery guarantees via Riemannian optimization," 2021, *arXiv:2110.03975*.

[58] Y. Luo and A. R. Zhang, "Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap, and their interplay," 2022, *arXiv:2206.08756*.

[59] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi, "Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements," 2021, *arXiv:2104. 14526*.

[60] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," *Ann. Statist.*, vol. 50, no. 1, pp. 1–29, 2022.

[61] S. Holtz, T. Rohwedder, and R. Schneider, "On manifolds of tensors of fixed TT-rank," *Numerische Mathematik*, vol. 120, no. 4, pp. 701–731, 2012.

[62] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[63] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[64] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[65] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[66] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, MA, USA: Cambridge Univ. Press, 2018.

[67] C. J. Hillar and L.-H. Lim, "Most tensor problems are np-hard," *J. ACM*, vol. 60, no. 6, pp. 1–39, 2013.

[68] J.-F. Cai, J. Li, and D. Xia, "Provable tensor-train format tensor completion by Riemannian optimization," *J. Mach. Learn. Res.*, vol. 23, no. 123, pp. 1–77, 2022.

[69] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via Procrustes flow," in *Proc. Int. Conf. Mach. Learn.,* PMLR, 2016, pp. 964–973.

[70] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[71] Z. Zhu, T. Ding, D. Robinson, M. Tsakiris, and R. Vidal, "A linearly convergent method for non-smooth non-convex optimization on the Grassmannian with applications to robust subspace and dictionary learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[72] J.-L. Goffin, "On convergence rates of subgradient optimization methods," *Math. Program.*, vol. 13, no. 1, pp. 329–347, 1977.

[73] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds.* Princeton, NJ, USA: Princeton Univ. Press, 2009.

[74] W. Luo, W. Alghamdi, and Y. M. Lu, "Optimal spectral initialization for signal recovery with applications to phase retrieval," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2347–2356, May 2019.

[75] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[76] H. Zhang, Y. Chi, and Y. Liang, "Median-truncated nonconvex approach for phase retrieval with outliers," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7287–7310, Nov. 2018.

[77] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *J. Mach. Learn. Res*, vol. 22, no. 150, pp. 1–63, 2021.

[78] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23–831.

[79] L. Jiang, Y. Chen, and L. Ding, "Algorithmic regularization in model-free overparametrized asymmetric matrix factorization," 2022, *arXiv: 2203.02839*.

[80] L. Ding, Z. Qin, L. Jiang, J. Zhou, and Z. Zhu, "A validation approach to over-parameterized matrix and image recovery," 2022, *arXiv: 2209.10675*.

[81] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So, "Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods," *SIAM J. Optim.*, vol. 31, no. 3, pp. 1605–1634, 2021.

[82] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338, Nov. 2018.